

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahlab Blida

N° D'ordre :



Faculté des sciences

Département d'informatique

Mémoire Présenté par :

IRKI Noussaiba

BOUNEDJAR Mehdi

En vue d'obtenir le diplôme de master

Domaine : Mathématique et informatique

Filière : Informatique

Spécialité : Informatique

Option : Ingénierie de logiciel

Thème : Système de Question Réponse pour l'interrogation d'une ontologie en langue Arabe du domaine juridique.

Soutenu le :

Mme. FARAH M

Président

Mme. ARKAM.

Examineur

Mme. Aliane Hassina Maitre de recherche CERIST, Promotrice

Promotion 2015 / 2016

MA-004-302-1

Remerciement

*Nous remercions tout d'abord, ALLAH qui nous a donné la force et le courage
pour terminer nos études et élaborer ce modeste travail.*

*Nous tenons à présenter nos sincères remerciements à Mme ALLANE Hassina,
maitre de recherche CERIST, de nous avoir encadré, ainsi que pour ses conseils
et encouragements.*

*Nous tenons à remercier aussi tous nos enseignants qui ont participé à notre
formation tout au long de notre cursus.*

*Nous tenons à remercier nos parents, nos familles et amis pour leur soutien tout
au long de ce mémoire.*

Merci.

المستخلص

المجال القانوني لديه الآن كمية كبيرة جدا من المعلومات النصية على شبكة الإنترنت، بما في ذلك الجريدة الرسمية التي تراكمت لديها قاعدة من الوثائق الإلكترونية منذ عام 1962، ويسمح لك بالبحث عن أي معلومات.

ونظرا لمحدودية الجريدة الرسمية الحالية على الموقع، سوف تستخدم الأنطولوجيا لتمثيل نموذج من الجريدة الرسمية، يسمح بالمرونة والفهرسة أفضل ولبناء نظام أفضل بالتالي سوف نقوم بتطوير نظام سؤال جواب.

تم تصميم هذه الأنظمة للرد على إجابة محددة، النص قد يحتوي على إجابة محددة على السؤال من قبل المستخدم. أعطيت لنا كجزء من هذا المشروع، تطوير نظام الاستجابة لتحسين الوصول إلى قوانين الجريدة الرسمية. مهمتنا هي وضع نظام للرد على أسئلة باللغة العربية للجريدة الرسمية.

كلمات البحث: نظام السؤال / الجواب، الجريدة الرسمية العربية، الأنطولوجيا، أوزان.

Résumé

Le domaine juridique dispose aujourd'hui d'un très grand volume d'information textuel sur le web, notamment le journal officiel qui dispose d'une base de documents électroniques accumulée depuis 1962 à nos jours, permettant la recherche d'une information quelconque.

Vu les limites du journal officiel actuel sur le site qui est figé, nous utilisons les ontologies comme modèle de représentation du journal officiel, elle permet une souplesse et une meilleure indexation et qui permet de construire un système de recherche plus performant donc en l'occurrence un système de question réponse.

Ces systèmes ont pour objectif de répondre par une réponse précise ou bien un passage de texte susceptible de contenir la réponse précise à la question de l'utilisateur.

Il nous a été confié, dans le cadre de ce projet de fin d'étude, le développement d'un système de question réponse pour améliorer l'accès au journal officiel. Notre travail consiste à la mise en place d'un système de question réponse en langage naturel pour le journal officiel arabe.

Mots-clés : Système question /réponse, journal officiel arabe, ontologie, patron.

Abstract

The legal field now has a very large volume of textual information on the web, including the official newspaper who have a base of electronic documents that has accumulated since 1962, it allows the search of any information

Given the limitations of the current official newspaper on the site which is frozen, we use the ontologies as a model of representation of the official newspaper, it allows flexibility and better indexing and for building a better search system like a system question answer.

These systems are designed to answer a specific question and then display the text might contain the best answer to the question of the user.

It was given to us as part of this project, development a system question answering to improve access to the official site. Our job is to set up a system to answer questions in natural language for the Arab official newspaper.

Keywords: question answering systems, Arab official newspaper, ontology, pattern.

Remerciement	
Résumé	
Sommaire	
Liste des abréviations.....	
Liste des figures.....	
Liste des tableaux.....	

Introduction générale

1. Contexte et objectifs :	1
2. Problématique :	1
3. Contributions :	2
4. Organisation du mémoire :	2

Chapitre 01 : Les systèmes de question réponse

1. Introduction :	5
2. Rappel sur les systèmes de recherche d'informations :	5
2.1 Définition d'un Système de Recherche d'Information :	5
2.2 Domaine d'application :	6
2.3 Architecture général d'un SRI :	6
2.4 Méthodes d'évaluation :	8
3. La relation de la RI avec les systèmes de question\réponse QR :	10
4. Du système de recherche d'information vers système question/ réponse QR :	11
5. Les systèmes de question/réponses Q /R :	12
5.1 Définition d'un système question/réponse :	12
6. La différence entre le moteur de recherche classique et un système de question réponse :	13
7. Architecture des systèmes de questions réponses :	13
7.1 Module d'analyse de la question :	14
7.2 Module de recherche des documents candidats :	15
7.3 Module d'extraction de la réponse :	15
8. Classification des systèmes question réponse :	15
8.1 Classification par type de question :	15
8.2 Classification par domaine :	17

8.2.1 Les systèmes de question réponse à domaine ouvert :	17
8.2.2 Les systèmes de question réponse à domaine fermé :	21
9. Les systèmes de questions réponses et la langue Arabe :	24
10. L'évaluation des SQR :	30
11. Conclusion :	31

Chapitre02 : les ontologies

1. Introduction :	33
2. Définition :	33
3. Les différents types d'ontologies :	34
3.1 Selon le degré de formalisme :	34
3.1.1 Les ontologies hautement informelles	35
3.1.2 Les ontologies semi informelles	35
3.1.3 Les ontologies semi formelles	35
3.1.4 Les ontologies rigoureusement formelles	35
3.2 Selon les objets modélisés :	35
3.2.1 Les ontologies supérieures:	36
3.2.2 Les ontologies de domaine (Domain ontologies) et les ontologies de tâche (task ontologies) :	36
3.2.3 Les ontologies d'application:	37
3.2.4 Les ontologies de représentation:	37
3.2.5 Les ontologies de raisonnement:	37
3.3 Selon la granularité	38
3.3.1 Granularité fine:	38
3.3.2 Granularité large:	38
4. Méthode de construction d'une ontologie :	39
4.1 La méthode développée par l'université de Stanford :	39
5. Les Langages de spécification d'ontologies :	41
6. Conclusion :	43

Chapitre03 : Conception et implémentation

1. Introduction :	45
2. Construction et implémentation de l'ontologie du journal officiel :	45
2.1 Structure du Journal Officiel « JO » :	45
2.2 Construction de notre ontologie (méthode de Stanford) :	46
2.3 L'implémentation de l'ontologie :	48
3. Conception et implémentation du système de question/réponse :	49
3.1 L'approche proposée :	49
3.2 Pourquoi l'approche à base de patrons :	49
4. Typologie des questions proposées :	50
5. Patrons lexico-syntaxiques :	57
5.1 L'expansion de la requête :	57
6. Fonctionnement du système :	58
7. L'architecture générale du système :	58
7.1 Analyse des questions :	59
7.1.1 Normalisation :	60
7.1.2 Analyse morphologique :	60
7.1.3 Extraction des termes par les patrons :	60
7.2 Extraction de la réponse :	62
7.2.1 Le langage SPARQL :	62
7.2.2 Traduction de la question analysée en SPARQL :	62
7.2.3 Filtrage selon les propriétés :	63
7.2.4 Le moteur d'inférence :	64
7.2.5 Extraction du texte (ou/et passage) :	64
8. Conclusion :	65

Chapitre04 : Expérimentation et évaluation

1. Introduction :	67
2. Les outils utilisés :	67
3. Expérimentation et évaluation :	68
3.1. Les questions utilisées :	68
4. Comparaison avec le JORADP :	69
5. Lecture des résultats :	74
6. Conclusion :	74

Conclusion générale

Conclusion générale.....76

Liste des abréviations :

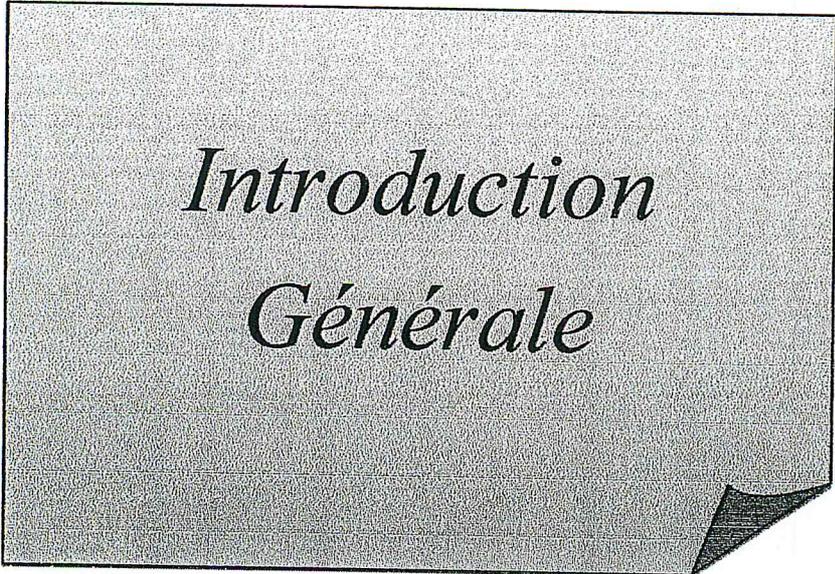
- RI Recherche d'information.
- SQR Système Question Réponse.
- TREC Text REtrieval Conference
- OWL Ontology Web Language.
- OML Ontology Markup Language.
- URI Universal Resource Identifier.
- XML extensible Markup Language.
- XOL Ontology Exchange Language.
- W3C World Wide Web Consortium.
- IBM International Business Machines.
- RDF Resource Description Framework.
- CLEF Cross-language Evaluation Form.
- SRI Système de recherche d'information.
- SHOE Simple HTML Ontology Extension
- TAL Traitement Automatique des Langues.
- DAML+OIL DARPA Agent Markup Language
- RDFS Resource Description Framework Schema.
- SPARQL SPARQL Protocol and RDF Query Language.
- OIL Ontology Interchange Language and Ontology Inference Layer.

Liste des figures :

Figure 1 : Processus en U de recherche d'information	7
Figure 2 : Evaluation, classification des documents	9
Figure 3: Evaluation, bruit et silence.....	9
Figure 4: Schéma type d'un système de Q/R.....	13
Figure 5: Architecture générale du système QRISTAL	19
Figure 6 : Architecture générale du système AQAS.....	26
Figure 7 : L'architecture du système El Bayan	29
Figure 8: Classification des ontologies	36
Figure 9 : Schéma représentatif des relations entre les ontologies de domaine, de raisonnement, de représentation et de haut niveau.	38
Figure 10: Les langages de spécification de l'ontologie.....	41
Figure 11 : Taxonomie de l'ontologie du journal officiel " JORADP" 'Onto JORADP-	48
Figure 12 : Processus d'expansion de la requête	57
Figure 13: l'architecture générale du système	59
Figure 14 : Détail de la phase d'analyse.....	61
Figure 15 : le processus d'interrogation d'ontologie.	62
Figure 16 : la recherche avec JORADP	69
Figure 17 : Résultat du mot « العقوبات » avec JORADP.....	70
Figure 18 : Recherche avec JO-Arabe.....	70
Figure 19 : Résultat du mot « الإستيرادات » avec JORADP.....	72
Figure 20 : Résultat du mot « مبلغ الإستيرادات » avec JORADP.....	72
Figure 21 : Résultat du mot « الثورة الزراعية » avec JORADP.....	73

Liste des tableaux :

Tableau 1: les classes de notre ontologie.	62
Tableau 2: les classes et les propriétés de l'ontologie.....	63



*Introduction
Générale*

Introduction générale

1. Contexte et objectif :

Aujourd'hui, le Web est devenu la source principale de l'information où beaucoup de téraoctets de données sont ajoutées tous les jours dans les différents domaines. Nous nous intéressons dans ce travail à l'amélioration de l'accès au journal officiel arabe, il possède comme ressource une base d'information accumulée depuis 1962 à nos jours, et qui sert à diffuser des textes juridiques (lois, décrets, décisions ...), ainsi que d'autres informations juridiques officielles.

Leur accès est souvent facile et gratuit, notamment grâce au réseau Internet.

Celui-ci touche différents secteurs de l'état. Ces textes sont promulgués par différentes instances conformément aux lois constitutionnelles de chaque pays.

Cependant, cette masse documentaire est devenue de plus en plus difficile à exploiter. L'exploitation de cette quantité de données a rendu la recherche complexe et coûteuse en termes de temps. Désormais, l'utilisation du journal officiel est très contraignante.

2. Problématique :

Actuellement, le secrétariat général du gouvernement algérien gère le site web www.joradp.dz, où nous pouvons consulter tous les journaux officiels, et d'accéder à n'importe quel texte juridique par rapport à sa référence, ou sa date de publication... etc., il permet aussi l'accès aux codes et à la constitution de l'état algérien.

Mais, cette approche est parfois difficile, car elle consiste à chercher des documents pertinents par rapport aux mots-clés de l'utilisateur, et c'est à l'utilisateur que revient la tâche d'explorer les documents pour en prendre l'information dont il a besoin.

Les systèmes administratifs en Algérie sont arabisés, une grande communauté travaille dans ce domaine (domaine juridique), le journal officiel est bilingue, nous nous intéressons à la langue arabe donc afin de résoudre ce problème, et répondre aux attentes des intéressés, nous allons développer un Système de Question Réponse pour l'interrogation d'une ontologie du journal officiel en langue arabe.

Introduction générale

3. Contribution :

Les systèmes de Question/Réponse (Q/R) sont une extension des systèmes de recherche documentaire. Ce type de système permet à un utilisateur de poser une question en langage naturel et de retourner une réponse à cette question au lieu d'un ensemble de documents jugés pertinents, comme c'est le cas des moteurs de recherche. En effet, face à une question donnée, les moteurs de recherche renvoient tous les documents jugés pertinents par rapport à la question, et c'est à l'utilisateur que revient la tâche d'explorer ces documents afin de trouver la réponse à sa question. Répondre à des questions précises requiert une analyse des documents sélectionnés afin d'en extraire l'information recherchée.

L'objectif de ce travail est de proposer un système de question-réponse qui interroge le journal officiel arabe représenté sous forme d'une ontologie pour pallier aux limites du joradp actuel.

4. Organisation du mémoire :

Afin d'atteindre le but de notre travail, le mémoire sera présenté en deux parties :

La première partie :

Contient deux chapitres, elle présente le contexte du travail, elle a pour but de présenter les systèmes de recherche d'information en générale, et les systèmes question réponse ainsi que les ontologies.

Chapitre 1 : Un état de l'art des systèmes de recherche d'information en montrant l'évolution des moteurs de recherche vers des systèmes plus performants tels que les systèmes de question-réponse. Nous présentons ensuite l'architecture typique d'un système de question réponse et ses différents modules.

Nous abordons par la suite quelques approches des systèmes Q/R multilingue et arabe.

Chapitre 2 : Ce chapitre propose une généralité sur les ontologies. Nous présentons les ontologies avec leurs caractéristiques, leurs constructions et leurs méthodes d'application, ainsi que les domaines d'application de ces ontologies. Nous allons décrire quelques langages de représentation des ontologies utilisées.

Introduction générale

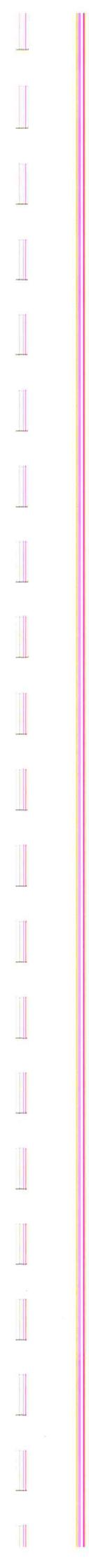
La deuxième partie :

Après les généralités présentées dans la première partie. Cette partie est consacrée pour la conception et le développement du système.

Chapitre3 : C'est le cœur de notre mémoire, nous présentons la conception et l'implémentation de l'ontologie et de notre système de question réponse, nous abordons aussi l'approche utilisée pour réaliser le système.

Chapitre4 : Ce chapitre est le résultat de la conception pour réaliser notre système, il s'agit de transformer les éléments décrits lors de la conception en éléments du langage cible. A la fin de ce chapitre nous allons effectuer des tests sur notre application, ainsi qu'une évaluation sur les résultats obtenus.

Et dans la conclusion, nous présenterons les principaux résultats obtenus dans ce mémoire ainsi que quelques perspectives.



Chapitre 01 :
Les systèmes de
questions réponses

1. Introduction :

Le grand service rendu par l'Internet a facilité l'accès à l'information. Cependant l'utilisation est confrontée à diverses entraves causant des perturbations aux utilisateurs. Afin de faire face à ce problème, certains outils sont apparus tels que les moteurs de recherche et les systèmes Question/Réponse (Q/R), essayant d'apporter un accès flexible et aisé à l'information.

Les systèmes de questions-réponses (SQR) sont considérés comme étant les plus avancés, et ont pour fonction de répondre à des questions formulées en langue naturelle, contrairement aux moteurs de recherche où l'utilisateur saisit un certain nombre de mots-clés correspondant à sa recherche. Une autre différence avec les moteurs de recherche est le résultat de la requête : alors qu'un moteur de recherche retourne un ensemble de pages correspondant aux mots-clés de la requête, les SQR renvoient la réponse précise à la question, avec souvent un passage de document la justifiant.

2. Rappel sur les systèmes de recherche d'informations :

2.1 Définition d'un Système de Recherche d'Information :

Il existe plusieurs définitions d'un SRI qui sont plus ou moins proches :

Tomek strzalkowski définit un SRI comme suit (Strzalkowski, 1993) :

La tâche typique de la recherche d'information est de sélectionner des documents dans une base de données, en réponse à une requête de l'utilisateur, et leur rangement par ordre de pertinence.

Tandis qu'Alan Smeaton donne la définition suivante (Smeaton, 1989) :

Le but d'un système de recherche d'information est de retrouver des documents en réponse à une requête des usagers, de manière à ce que les contenus des documents soient pertinents au besoin initial d'information de l'utilisateur.

Salton et McGill donnent une définition d'un SRI plus précise et complète (SM, 1983)

Un SRI traite de la représentation, du stockage des informations, de l'organisation de ces informations et de l'accès aux éléments de l'information

La recherche d'information est la science qui étudie la manière de répondre pertinemment à une requête en retrouvant de l'information dans un corpus.

Nous définissons un SRI comme étant un système permettant de retrouver les documents pertinents à une requête d'utilisateur écrite dans un langage libre, à partir d'une base de documents volumineuse. Dans cette définition, il y a trois notions clés les documents, la requête et la pertinence :

- **Documents** : « Le document est l'élément centrale du SRI, c'est un objet complexe sans cesse en évolution car il est lié aux développements des technologies de la communication ». Un document peut être un texte, un morceau de texte, une page WEB, une image, une bande vidéo, etc. On appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur.
- **Requête** : une requête est une façon d'exprimer un besoin en information de l'utilisateur par un ensemble de mots clés, ce besoin est traduit à l'aide d'un langage naturel ou booléen.
- **Pertinence** : La pertinence est une notion complexe et un peu floue qui dépend de l'utilisateur et de la requête mais de façon générale, le but de la RI est de retrouver Seulement les documents pertinents et un document pertinent doit contenir l'information que l'utilisateur recherche. C'est sur cette notion que les SRI sont jugés. (BOUCHAM,2009).

2.2 Domaine d'application :

La RI est un domaine vaste qui se situe à la frontière de plusieurs disciplines tel que :

- Recherche adhoc,
- Classification /catégorisation (clustering).
- Filtrage d'information (filtering/recommendation)
- Méta-moteurs (data-fusion, Meta-search)
- Résumé automatique
- Croisement de langues)
- Fouille de textes (Text mining).

2.3 Architecture général d'un SRI :

Le processus de recherche d'information pertinentes que le SRI est sensé restituer à un utilisateur, consiste en la mise en correspondance des représentations des informations

contenues dans un fond documentaire des besoins de cet utilisateur exprimé par une requête.

Cette notion de pertinence peut être appréhendée à deux niveaux :

- **Niveau utilisateur** : la pertinence correspond à la satisfaction de l'utilisateur de l'ensemble des documents restitués par le SRI.
- **Niveau système** : le système mesure un degré de pertinence, une valeur de similitude entre un document et une requête.

Le but de tout SRI est de rapprocher la pertinence système de la pertinence utilisateur. Pour effectuer de façon efficace cette fonction, le SRI doit réaliser l'indexation des documents, la formulation de la question, la comparaison requête-documents et enfin la reformulation de la requête. D'après (BOUCHAM,2009) Nous pouvons représenter schématiquement un SRI, comme illustré par la figure 1, Ce qui est appelé communément le processus en U de recherche d'information. (BOUCHAM,2009)

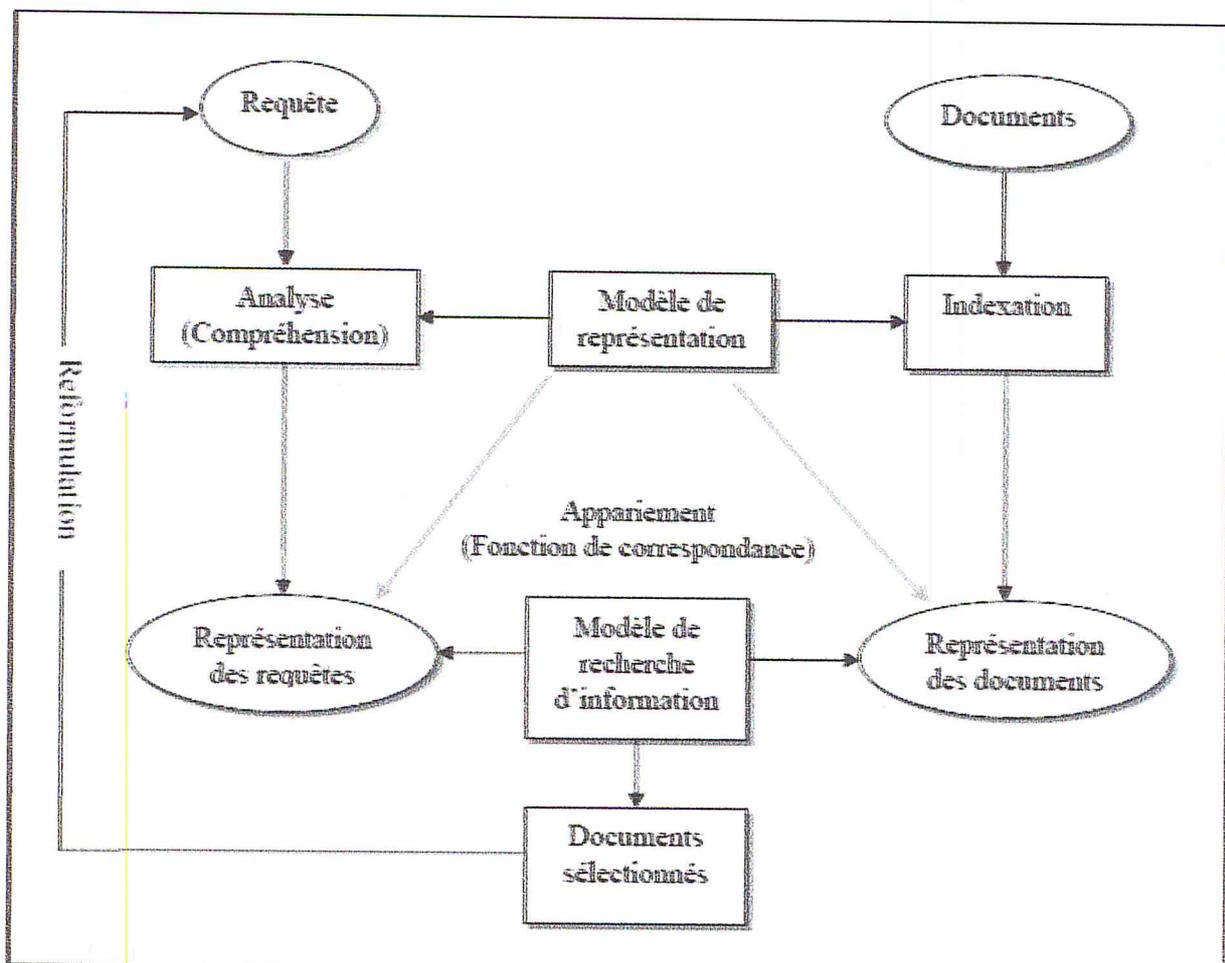


Figure 1 : Processus en U de recherche d'information

Le bruit représente les documents extraits mais non pertinents.

$$\text{Bruit} = \frac{\text{Nombre de documents retournés et non pertinents}}{\text{Nombre de documents extraits}}$$

Le silence (aussi appelé facteur d'omission) représente les documents pertinents non extraits

$$\text{Silence} = \frac{\text{Nombre de documents non retrouvés et pertinents}}{\text{Nombre de documents pertinents}}$$

Un SRI sera d'autant meilleur que le bruit et le silence seront faibles. On pourra représenter le bruit et le silence en proportion du nombre de documents extraits.

Deux autres mesures sont souvent utilisées, qui sont en fait les compléments des deux précédentes :

La précision qui représente le nombre de documents pertinents extraits par rapport au nombre de documents extraits.

$$\text{Precision} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents extraits}}$$

Le rappel, qui représente le nombre de documents pertinents extraits par rapport au nombre de documents pertinents.

$$\text{Rappel} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents pertinents}}$$

Un SRI sera d'autant meilleur que la précision et le rappel seront forts.

3. La relation de la RI avec les systèmes de question\réponse QR :

Un système QR permet de répondre aux questions relatives à un petit domaine. Par exemple, on peut poser la question "quelle est la dernière version de Word ?" à un système spécialisé sur le marché de logiciel. Pour cela, il faut qu'on crée une modélisation du domaine d'application dans lequel les concepts ou objets sont reliés par des relations sémantiques. Ce modèle permettra de retrouver le concept ou l'objet et ainsi donner une réponse directe à la question. Pour notre exemple, la réponse peut être "Word 2016 et Word 2013", par exemple.

Nous voyons ici qu'il y a une différence sur la nature de réponse entre les deux types de système. Dans la RI, c'est une réponse indirecte à une question : nous identifions les documents dans

lesquels l'utilisateur peut trouver des réponses directes à sa question. Tandis que dans un système QR, on fournit une réponse directe.

Il y a des tentatives de rapprocher la RI vers des systèmes QR, mais cela s'avère très difficile. La raison principale est que la RI s'applique en général à tous les domaines sans restriction. Il est difficile, dans ce cas, de créer un modèle nécessaire pour déduire la réponse directe à une question dans un système QR. Dans certains contextes très spécialisés, la RI incorpore une base de connaissances. Elle utilise aussi des raisonnements pour déduire si un document peut être pertinent ou pas. Donc, le fonctionnement de ce type de RI ressemble un peu plus à celui d'un système de question réponse (Jian-Yun,2009).

Une tentative plus restreint consiste à raffiner la notion de document dans la réponse : au lieu de fournir un document complet comme une réponse, d'identifier un passage dans le document (passage retrieval). C'est une étape qui diminue un peu la distance entre la RI et la QR. Mais la différence fondamentale reste la même (Jian-Yun,2009).

4. Du système de recherche d'information vers système question réponse QR :

L'accès aux systèmes de recherche d'information sur Internet (bibliothèques numériques, moteurs de recherche ...) reste problématique pour les non-spécialistes du processus de recherche. (Ihadjadene, 1999) et (Spink, 2001) ont montré que les ressources sont sous-utilisées dans les catalogues en ligne, les bibliothèques numériques ou les moteurs de recherche. L'une des premières difficultés pour les usagers est d'identifier les sources pertinentes et d'avoir une vision claire des contenus et de leur structure. Par ailleurs, bien que les moteurs de recherche proposent des fonctionnalités avancées comme les opérateurs booléens, les usagers les emploient rarement. Plusieurs études ont également montré que le choix des termes ne se fait pas selon un plan de recherche méthodique mais résulte au contraire d'une démarche d'essai erreur, aggravée par le fait que les usagers n'ont souvent qu'une vague idée de ce qu'ils cherchent. Qu'ils préfèrent changer le contenu de la requête plutôt que de modifier sa formulation (Stéphane,et.al,2004).

Les systèmes de questions-réponses (QR) peuvent se voir comme des systèmes de RI avancés. Dans un système de RI (par ex. le moteur de recherche GOOGLE), l'utilisateur définit sa recherche par une requête contenant des mots-clés. Le résultat de la recherche est un ensemble

de documents ou de liens vers des documents que l'utilisateur doit ensuite fouiller afin de trouver l'information qu'il recherche. Par exemple, pour répondre à la question « Quel acteur joue le rôle principal dans le film Golden Eye ? » sans rien ne savoir ni du film ni de l'acteur recherché, il est nécessaire à un internaute de suivre le 3^{ème} résultat parmi les réponses proposées par Google à la requête « Golden Eye main actor » et ainsi vérifier que l'acteur Pierce Brosnan mentionné dans la section de ce résultat, est bien l'information qu'il recherche.

Par contre, les systèmes de question réponse fournissent à l'utilisateur des réponses précises à leur question formulée en langue naturelle. Ainsi, pour la même recherche que celle faite précédemment à GOOGLE, on posera par exemple au système question/réponse START la question « Who is the main actor in Golden Eye ? » et il va répondre en retour par : Pierce Brosnan (Foucault,2013).

5. Les systèmes de question/réponses Q /R :

5.1 Définition d'un système question/réponse :

Les SQR cherchant une réponse dans une collection de documents combinent plusieurs domaines dont notamment la recherche d'information et le traitement automatique des langues à travers l'extraction d'information. En effet, là où des moteurs de recherche renvoient des références de documents (avec éventuellement un extrait de ces documents) suite à une requête saisie sous forme de mots-clés, les SQR travaillent à partir d'une question en langue naturelle dont tous les mots ne sont pas forcément pertinents pour la recherche d'information. Après une analyse de la question propre à chaque système, ils sélectionnent un ensemble de documents de la collection puis extraient des candidats pour la réponse recherchée depuis ces documents. Les SQR extraient plusieurs candidats à la réponse, que nous désignerons désormais sous le terme de candidats-réponses, et les ordonnent selon des critères qui leur sont propres. Enfin, une étape de validation de réponse peut également intervenir, aussi bien durant l'ordonnancement qu'après. L'ultime étape est bien évidemment de présenter la ou les réponses à l'utilisateur (Mathieu-Henri,2014).

Les systèmes varient en fonction de plusieurs paramètres :

- 1) le domaine d'application : n'importe quel sujet peut-il être abordé ? Ou y-a-t-il un ou plusieurs domaines circonscrits ?
- 2) le type d'application : commerce, support, accueil, etc.

3) les bases dans lesquelles ils vont puiser les réponses ;

4) leur fonctionnement (analyse de la question ; recherche de la réponse ...).

Pour résumer, nous distinguons donc plusieurs parties communes à la quasi-totalité des SQR :

- a. Une analyse de la question afin de :
 - Transformer la question en langue naturelle en une requête formalisée.
 - Déterminer le type de la question et le type de la réponse attendue.
- b. La recherche de documents à l'aide de cette requête formalisée.
- c. L'extraction des candidats-réponses.
- d. L'ordonnancement, l'agrégation et éventuellement la validation des candidats-réponses.
- e. La présentation de la ou des réponses

6. La différence entre le moteur de recherche classique et un système de question réponse :

Contrairement aux moteurs de recherche, les systèmes de Question/Réponses ne se contentent pas de retrouver les documents contenant une certaine combinaison de chaînes de caractères mais essaient plutôt d'obtenir une réponse exacte à une question spécifique (la question et la réponse sont formulées toutes les deux en langage naturel).

7. Architecture des systèmes de questions réponses :

L'architecture des SQR dépend des stratégies utilisées : chaque système a ses stratégies propres et de ce fait sa propre architecture.

En général un SQR est composé de trois modules principaux :

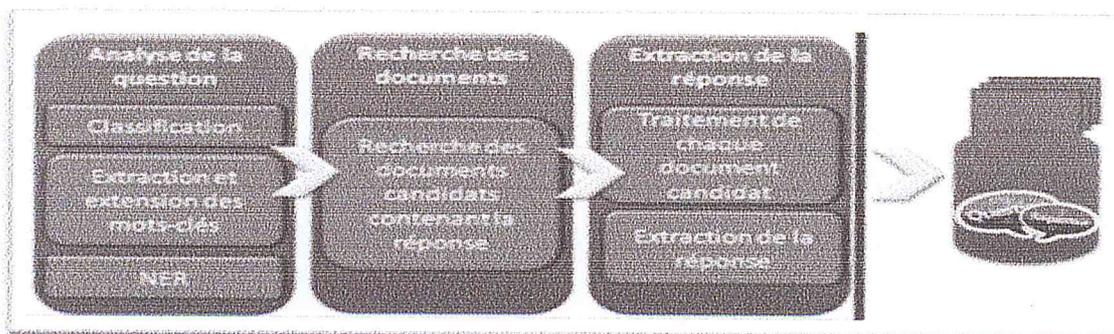


Figure 3: Schéma type d'un système de Q/R (BOUZOUBAA et al., 2010).

7.1 Module d'analyse de la question :

L'analyse de la question est un élément primordial d'un système de question réponse. Si cette analyse est erronée, les chances de trouver une bonne réponse à la question sont sérieusement compromises ; l'analyse de la question doit déterminer :

- **Le typage de la question** : il permet d'attribuer à la question une catégorie selon une classification prédéfinie (Définition, Factuelle, Booléenne).
- **Les entités nommées de la question** : il s'agit de repérer toutes les entités nommées présentes dans la question telles que personne, lieu, argent, pourcentage etc.
- **Le type de la réponse attendue** : Le type de la réponse attendue correspond à l'identification de l'objet de la question, ce type est généralement formalisé sous la forme d'un type d'entité nommée (personne, date, lieu, ...) ou d'un type d'entité plus élargi (un évènement)
- **Le focus de la question** : il s'agit d'extraire l'objet sur lequel porte la question, c'est-à-dire un élément susceptible d'être présent dans le passage réponse.
- **NER (Named-entity recognition)** : Un module de reconnaissance des noms d'entités qui permet de distinguer personnes, pays, organismes, etc.
- Un module d'extension des mots clés originaux de la requête pour augmenter les chances de trouver des documents contenant une ou plusieurs parties de la réponse, l'extension peut être morphologique mais aussi sémantique, par synonymes, mots de la définition, les super types et les sous types du concept.

Exemple :

Les informations extraites lors de l'analyse de la question (la catégorie, le type général, les entités nommées, le focus, les termes et leurs variations sémantiques) sont ensuite utilisées par le moteur de recherche pour créer une requête qui fournira des documents contenant les mots de la question, ou bien leurs variations (synonymes).

Les informations extraites lors de l'analyse de la question « qui est le président de l'Algérie en ce moment ? », « من هو الرئيس الحالي للجزائر ؟ » Seraient :

– catégorie : qui / هو

– type général : président / الرئيس

- entité nommée : /
- type d'entité attendue : président / الرئيس
- focus : président/الرئيس
- verbe principal : être (est)

Ces informations seront nécessaires pour la sélection des documents. Enfin, lors de l'extraction de la réponse précise, le focus, le verbe principal et l'entité nommée attendue jouent un rôle déterminant grâce à des patrons d'extraction de la réponse définis sur ces éléments.

7.2 Module de recherche des documents candidats :

- **La sélection des documents candidats** extrait un sous-ensemble de documents contenant possiblement une réponse à la question ou de l'information pertinente devant être analysée pour en déduire une réponse.
- **L'analyse des documents candidats** est effectuée pour extraire de l'information provenant des documents candidats pour préciser la question ou ajouter des faits qui aideront à trouver la réponse.

7.3 Module d'extraction de la réponse :

À partir du corpus de documents : il traite chaque document candidat de la même manière que la question, et pour extraire les phrases qui contiennent la réponse, un filtrage est appliqué sur l'ontologie utilisées (qui est considérée le noyau de système), grâce aux relations implicites ou explicites entre les classes de l'ontologie. Il est fondé sur les processus d'inférences (une inférence est un mode de raisonnement consistant à aller d'une connaissance à une autre qui lui est liée).

8. Classification des systèmes question réponse :

8.1 Classification par type de question :

- a. **Question de définition** : Les questions de définition cherchent à obtenir la définition d'un objet « Qu'est-ce que l'Atlantis ? » ou demandent une information sur une personne ou une organisation comme « Qui est Tim berners-lee ? ». Pour ce type de questions, certaines

réponses peuvent être courtes (« inventeur du World Wide Web (WWW) ») mais d'autres peuvent être beaucoup plus élaborées. La réponse à la dernière question pourrait ainsi parler de son parcours politique, de son enfance ... L'extraction des réponses à ce type de question est souvent fondée sur l'utilisation de patrons d'extraction comme « Tim berners-lee est <Réponse> ». La validation de réponses peut elle aussi être simple en retenant la réponse la plus souvent extraite ou provenant d'un patron fiable.

- b. **Les questions factuelles** : les questions factuelles cherchent à obtenir une information précise concernant un événement ou une entité comme une date (« En quelle année l'Espagne est-elle devenue championne du monde de football ? »), un lieu (« Dans quelle ville se trouve la tour Eiffel ? ») ou une personne (« Qui a tué Henri IV ? »). Certaines questions attendent une entité nommée d'un type particulier en retour et d'autres non (« Citer une théorie de Albert Einstein. ») mais attendent généralement dans ce cas un groupe nominal.
- c. **Les questions de type listes** : attendent un certain nombre de réponses, ce nombre étant spécifié (« Quels sont les pays de l'union européenne ? »). Le mécanisme visant à valider une réponse peut être perçu comme une partie du traitement des questions factuelles de type liste pour lesquelles il est aussi nécessaire de s'assurer que toutes
- d. **Questions booléennes** : attendent généralement comme réponse soit OUI soit NON. C'est par exemple le cas de la question « existe-t-il plusieurs sortes de bicarbonate de soude ? » la réponse sera oui ou non plus un passage justificatif correspond au texte et la forme déclarative de la question à l'hypothèse, ici « Il existe trois types de bicarbonate de soude
- e. **Les questions complexes** : attendent une réponse plus élaborée et dont la taille peut aller jusqu'à un paragraphe. Elles peuvent être séparées en plusieurs catégories. Certaines questions demandent ainsi l'opinion d'une personne concernant un événement (« Que pensez-vous de la dernière version de Windows ?») et d'autres une explication comme les questions commençant par les pronoms interrogatifs « comment » (« Comment utiliser le braille ») et « pourquoi » (« Pourquoi la terre tourne ?»).

8.2 Classification par domaine :

8.2.1 Les systèmes de question réponse à domaine ouvert :

Les systèmes question réponse du domaine ouvert répondent à des questions à un sujet non spécifique. Un système est dit en domaine ouvert quand il peut supporter n'importe quel sujet abordé dans une question de l'utilisateur, quel que soit sa formulation et son sujet. Le corpus de documents pour la recherche des réponses peut éventuellement être le web avec des formats et des styles très divers, ou alors des grands ensembles de documents électroniques.

❖ Système JAVELIN :

Les développeurs du JAVELIN ont proposé un système de question-réponse JAVELIN (Justification based Answer Valuation through Language INterpretation) (Nyberg et al., 2002) fondé sur une interaction avec l'utilisateur dont l'intérêt est d'élucider la question et de déterminer une stratégie de recherche adaptée pour trouver la réponse.

Il est composé de trois principaux modules :

1/ Il utilise un analyseur qui permet de déterminer certaines caractéristiques tel que : le type de la question, les mots clés, le type de la réponse attendue, les synonymes ...)

2/Pour la recherche, il exploite plusieurs bases de données selon le type de la question, après avoir eu les documents nécessaires, il élimine les passages non pertinents selon un indice de confiance (à vrai dire c'est selon le nombre de mots clés trouvés dans le passage).

3/Au final il classe les réponses selon un ordonnancement qui s'appuie sur le résultat obtenu dans l'étape précédente.

Le système JAVELIN offre à l'utilisateur la possibilité de définir la procédure de recherche appliquée sur la question, cela c'est pour lui apporter plus de précision à sa requête et aussi pour guider la stratégie de recherche. A la fin, il lui affiche non seulement la réponse mais aussi une justification de la réponse (le passage où la réponse a été prise).

❖ Piquant :

Le système PIQUANT d'IBM (Chu-Carroll et al. 2002) se base sur l'utilisation de plusieurs systèmes de questions réponses, l'avantage de ce système c'est la redondance et la pluralité des réponses trouvées, ce qui donne une réponse assez pertinente.

PIQUANT se fonde sur différents agents indépendants pour rechercher une réponse. Parmi ces « agents réponses », on note un agent fondé sur des outils statistiques et d'autres, sur des outils de traitement automatique des langues. Le typage de la question repose sur une analyse syntaxique qui permet de déterminer le type de la question, type de la réponse, mots clés ... etc. Il utilise plusieurs sources de connaissances comme WordNet (pour avoir plus de variantes d'un mot clé) et CYC (pour les connaissances de sens commun qui permet au système de réaliser des inférences), (Cyc est un projet d'intelligence artificielle, qui cherche à développer une ontologie globale, et une base de connaissance).

❖ **Système Power Answer :**

Power Answer est un système de question/réponse avec une architecture fondée sur le raisonnement logique. Proposé par (Moldovan et al. 2002).

Il repose sur la représentation sous forme de formules logiques de la question, de la réponse ainsi que des sources de données servant à extraire la réponse.

Il utilise pour l'analyse de la question des données sémantiques du WordNet et ainsi un module de reconnaissances d'entité nommées, cela pour déterminer : le type de la question, le type de la réponse, les mots clés, les entités nommées ... etc.

Pour extraire la réponse, il utilise le même module pour repérer les entités nommées au type de la réponse dans le passage sélectionné.

En cas où le système ne trouve pas de réponses, il utilise un programme de démonstration automatique de réponse en utilisant une base d'axiomes à partir de WordNet pour déduire une réponse.

Il affiche la réponse, plus la chaîne de raisonnement (en cas où il utilise le programme).

❖ **Système d'InsightSoft :**

Le système d'InsightSoft se base sur l'application massive des patrons d'extraction sous forme d'expression régulière et l'exploitation d'une base de connaissances factuelle.

Il analyse la question pour déterminer le type de la question, puis il sélectionne les patrons à appliquer pour extraire la réponse d'un passage. Chaque passage sélectionné sera analysé par ces patrons-là.

Ce système s'est classé premier lors de l'évaluation TREC-10, avec 77% de réponses correctes (Embarque mehdi, 2009).

❖ **QRISTAL :**

QRISTAL (Dominique, et.al,2005) est un système de questions / réponses multilingue (Français, Anglais, Italien, portugais et polonais) conçu pour extraire des réponses de documents placés sur un disque dur ou pour extraire des réponses à partir du WEB sur la base des pages ou passages retournés par le moteur WEB classique (GOOGLE, AOL.Etc.)

Le système reconnaît un grand nombre de format (HTML ; XML ; Txt ; .PDF ; .HLP...etc.) autorisant ainsi l'indexation de l'immense majorité des textes mais également des e-mails ou encore des fichiers d'aide.

• **Architecture**

L'architecture du système est modulaire ; le schéma général est décrit en dessous :

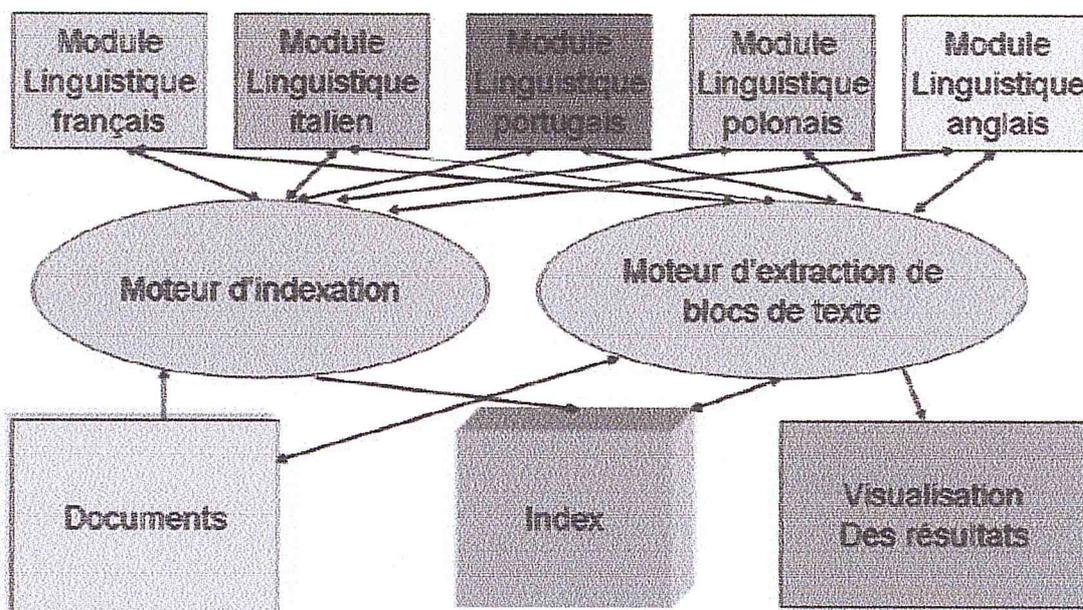


Figure 4: Architecture générale du système (Dominique,2005)

Ce système est donc un moteur complet d'indexation de réponse. Toutefois l'indexation n'est effectuée que pour les documents « statiques », la recherche sur le web se faisant à l'aide d'un méta moteur, par conséquent sans indexation préalable des pages.

❖ QALC :

Le système QALC (Question Answering program of the Language and Cognition group) (Ferret et al. 2000 ; Ferret et al., 2001) a été le premier système de question-réponse développé pour l'anglais au sein du LIMSI dans le cadre de la campagne d'évaluation TREC en 1999. Il a constitué la base des systèmes suivants ayant participé à d'autres campagnes d'évaluation comme EQueR 2004 et CLEF-Q

Depuis 2005, une version cross lingue a également été élaborée pour participer aux campagnes d'évaluation CLEF-QA en 2005 et 2006 (Grau et al. 2006). Ce système, MUSCLEF, prend en entrée des questions en français, et recherche leurs réponses dans des documents en anglais. Cela correspond en réalité à deux sous-systèmes, utilisant deux stratégies parallèles pour passer d'une langue à l'autre. La première stratégie s'appuie sur la traduction de la question par un traducteur automatique qui est ensuite passée en entrée du système QALC. La deuxième stratégie quant à elle consiste à traduire chaque terme de la question. L'ensemble des termes traduits est implémenté dans un système particulier appelé MUSQUAT. Enfin, les résultats des deux sous-systèmes sont ensuite combinés, et les réponses résultant de cette fusion sont celles du système MUSCLEF.

• L'architecture :

L'architecture de QALC comporte six modules principaux, deux pour le traitement des questions, trois agissant sur le corpus et un sixième module d'appariement entre la question et des phrases

Analyse des questions : L'analyse des questions est effectuée par l'application de règles permettant de discriminer les types de questions et de leur assigner une (ou des) catégorie(s). Ces catégories décrivent les types d'entités qui devront constituer la réponse.

Extraction de termes : L'extraction de termes utilise des patrons syntaxiques décrivant des groupes nominaux complexes. Nous cherchons l'appariement maximal de ces modèles sur les phrases afin de reconnaître de tels groupes nominaux dans les questions.

Indexation automatique et reconnaissance de variantes : L'indexation automatique se fonde sur l'utilisation de FASTR, un analyseur transformationnel qui reconnaît les occurrences de termes

et leurs variantes dans le corpus. Les termes et les variantes constituent un index des documents, et sont utilisés dans le processus de classement des documents et des réponses.

Reconnaissance des entités nommées : Les entités nommées sont recherchées dans les documents afin de produire des indices permettant d'évaluer le degré de similarité entre une question et une phrase candidate pour la réponse. Ces entités sont extraites grâce à des patrons lexico-syntaxiques et à l'exploitation de lexiques.

Ordonnement des documents et sélection : Les documents sont ordonnés selon un poids calculé à partir des termes et des types de variantes qu'ils contiennent. Seuls les n premiers sont retenus. Une sélection plus fine est appliquée ensuite si la courbe des valeurs de pertinence des documents correspond à un plateau.

Appariement Question/Réponse : Toutes les données extraites des documents et des questions sont utilisées pour évaluer le degré de similarité entre chaque question et les phrases des documents sélectionnés pour chacune d'elles. La similarité tient compte des mots, des termes des variantes et des catégories d'entités nommées communes (Ferret et Grau, 1999).

8.2.2 Les systèmes de question réponse à domaine fermé :

Les systèmes question-réponse dans un domaine fermé, traitent de questions sur un domaine spécifique (par exemple de la géographie, la médecine, tourisme... etc.) permettent à des experts du domaine ainsi qu'à des non spécialiste d'interroger des documents ou une base de connaissances spécifiques au domaine pour l'obtention des réponses pertinentes.

Les premiers systèmes de question-réponse de ce domaine sont : « Baseball » qui répond à des questions factuelles concernant des matchs de baseball de la ligue américaine et « Lunar » qui répond à des questions concernant des échantillons de sol lunaire en utilisant une base de données. Ces systèmes sont maintenant connus par l'appellation d'interface en langage naturel à une base de données.

❖ **Système WebCoop :**

Le système WEBCOOP (COOPérativité pour le WEB) (Benamara, 2004) a été développé pour le domaine du tourisme, un système de génération de réponses coopératives. L'idée est de proposer à l'utilisateur des informations additionnelles (explications, justifications, etc.).

Ce système de question-réponse permet de retourner une réponse même quand la question posée comporte des fausses présuppositions ou des malentendus. Il a été développé pour répondre à des questions portant sur le domaine du tourisme en s'appuyant essentiellement sur une ontologie du domaine et des bases de connaissances regroupant les aspects hébergement et transport.

Les questions sont classées selon deux catégories où chaque catégorie se compose de plusieurs classes sémantiques permettant de mieux préciser le type de la réponse recherchée.

Une première catégorie concerne les questions qui attendent une entité nommée comme réponse, tandis qu'une deuxième catégorie porte sur les questions dont la réponse est une entité textuelle.

La question est représentée sémantiquement sous forme d'un triplé : la catégorie de la question, le type de la réponse attendue et une représentation de la question en formules logiques du premier ordre.

Pour l'extraction de la réponse, il utilise un moteur d'inférence qui permet de comparer la question avec les documents de la base documentaire grâce à des procédures de Raisonnement.

La réponse se décompose en deux parties : la première partie sera une réponse directe, par contre la deuxième partie sera une réponse coopérative.

❖ **TeLQAS :**

Un système question / réponse pour le domaine des technologies des télécommunications. Dans un processus en ligne, le système accepte les questions des utilisateurs en anglais, et après la récupération des documents de texte liés de sa base de données soit local ou à travers web ; il résume les documents de texte récupéré avec la plus grande pertinence.

Architecture :

TeLQAS comprend trois sous-systèmes principaux :

Un sous-système en ligne, un sous-système hors ligne, et une ontologie.

Le sous-système en ligne répond aux questions soumises par les utilisateurs en temps réel. Tant que le processus reste en ligne, TeLQAS traite les questions à l'aide d'un composant de traitement du langage naturel qui implémente une simple analyse syntaxique et l'étiquetage d'une partie du discours. Le sous-système en ligne utilise également un moteur d'inférence afin de procéder à l'inférence nécessaire sur les petits éléments de connaissance.

Le sous-système hors ligne automatiquement indexe les documents collectés par un robot Web focalisé sur le web. Un serveur d'ontologie ainsi que son API est utilisé pour la représentation des connaissances.

Les principaux concepts et les classes de l'ontologie sont créés par des experts du domaine.

❖ EPO Care :

L'objectif du projet EPO Care (Evidence at Point of Care) est de développer des méthodes pour répondre à des questions dans le domaine médical en utilisant CE ou bien « clinical Evidence » comme une source.

Le système accepte les questions qui contiennent de mots clés au format PICO (Sackett et al., 2000).

Dans ce format, une question clinique est représenté par un ensemble de quatre champs qui correspondent à les éléments de base de la question :

P : une description du patient (ou le problème) ;

I : une intervention ;

C : une intervention de comparaison ou de contrôle (peut être omise) ;

O : le résultat clinique.

Par exemple, la question ci-dessus peuvent être représentés

en format PICO comme suit :

P : infarctus du myocarde

I: thrombolyse

C: -

O: mortalité dans le projet est d'étendre le mot-clé pour que le système de récupération des mots pour répondre aux questions exprimées en langage naturel.

Dans nos travaux antérieurs (Niu et al., 2003), ils ont montré que les technologies actuelles comme les systèmes question repense factuelle ne sont pas suffisantes pour les questions

cliniques, dont les réponses doivent souvent être obtenues par la synthèse d'un contexte pertinent.

Pour adapter à cette nouvelle caractéristique de QR dans le domaine médical, nous exploitons des classes sémantiques et les relations entre eux, dans un texte médical, les classes sémantiques sont importantes pour cette tâche parce que l'information contenue en eux est souvent un bon candidat pour répondre à des questions cliniques.

Dans l'exemple ci-dessus, les éléments PICO correspondent trois classes sémantiques.

MALADIE (problème médicale du patient), **INTERVENTIONS** (médicaments appliquée à la maladie) et la **RÉSULTAT CLINIQUE**. Ils constituent ensemble un **SCENARIO** du traitement. De même, un scénario de diagnostic souvent comprend des symptômes, **PROCÉDURE D'ESSAI**, et **MALADIES** supposée. Pour comprendre la sémantique du texte médical et de trouver des réponses aux questions de la clinique nous avons besoin de savoir comment ces classes se rapportent les uns aux autres dans un scénario spécifique. Par exemple, est ce que ce médicament, un type spécial de d'un autre est-ce ce médicament appliqué à cette maladie ? Voici les types de relations qui nous intéressent. Dans ce travail, une approche « cue-word » ou bien des mots de repères est utilisée pour identifier des classes sémantiques dans le scénario de traitement, et analyser les relations entre eux. Le processus de classification automatique est également appliqué pour déterminer la polarité d'un résultat.

9. Les systèmes de questions réponses et la langue Arabe :

Les systèmes de question réponse pour la langue arabe sont très peu par rapport aux langues latines telles que l'Anglais et le Français. Généralement, cela est dû à l'absence de l'accessibilité aux ressources linguistiques, comme des corpus et des outils de TAL de base (tokenizers, analyseur morphologique, etc.). De plus, la langue arabe a une très complexe morphologie, et les textes souffrent de la rareté des voyelles. Ces spécificités de la langue arabe introduisent de nombreux problèmes de traitement liés à la tokenization, l'identification et la catégorisation des entités nommées, etc.

Pour les SQR en arabe, voici quelques implémentations qui existent :

❖ **QARAB (Hammond, et al., 2002) :**

Il traite des questions exprimées en langue arabe et fournit des réponses courtes. Sa principale source de connaissance est une collection de journaux arabes extraits d'Al Raya, un journal publié au Qatar, mais il ne procède pas à une analyse sémantique de la question.

Le traitement de question réponse de base en QARAB est composé de trois étapes principales :

- Traitement de la question d'entrée
- Recherche des documents candidats (Paragraphes) contenant des réponses de l'IR système
- Traitement de chaque document candidat (de la même manière que la question est traitée) et retourner des phrases qui peuvent contenir la réponse

Le système de QARAB atteint une précision de 97,3% et un rappel de 97,3%. L'évaluation a été effectuée directement par quatre natifs arabes orateurs qui ont présenté 113 questions au système et jugent l'exactitude des réponses.

❖ **AQUASYS (Bekhit, et al, 2010) :**

Est un système Question réponse conçu pour permettre à l'utilisateur de saisir une question formulée en langage naturel arabe (des questions commencent par des pronoms interrogatifs (qui, quoi, comment combien et quand). et obtient une réponse précise à la question posée.

Aquasys peut répondre à des questions liées à :

Une entité nommée qui peut être de tout type : personne, emplacement, organisation, temps, et la quantité, etc.

L'architecture d'Aquasys repose sur quatre modules principaux,

- Analyse de la question,
- Le filtrage des phrases,
- Trouver des réponses candidats
- Modules de classement.

Chaque module a été développé sur la base d'un certain nombre de sous-modules et / ou des procédés.

- ❖ AQAS (Mohammed et al. 1993) : Il est basé sur la connaissance, et extrait les réponses seulement des données structurées et non pas de textes bruts.

L'architecture de système AQAS est représentée dans en dessous :

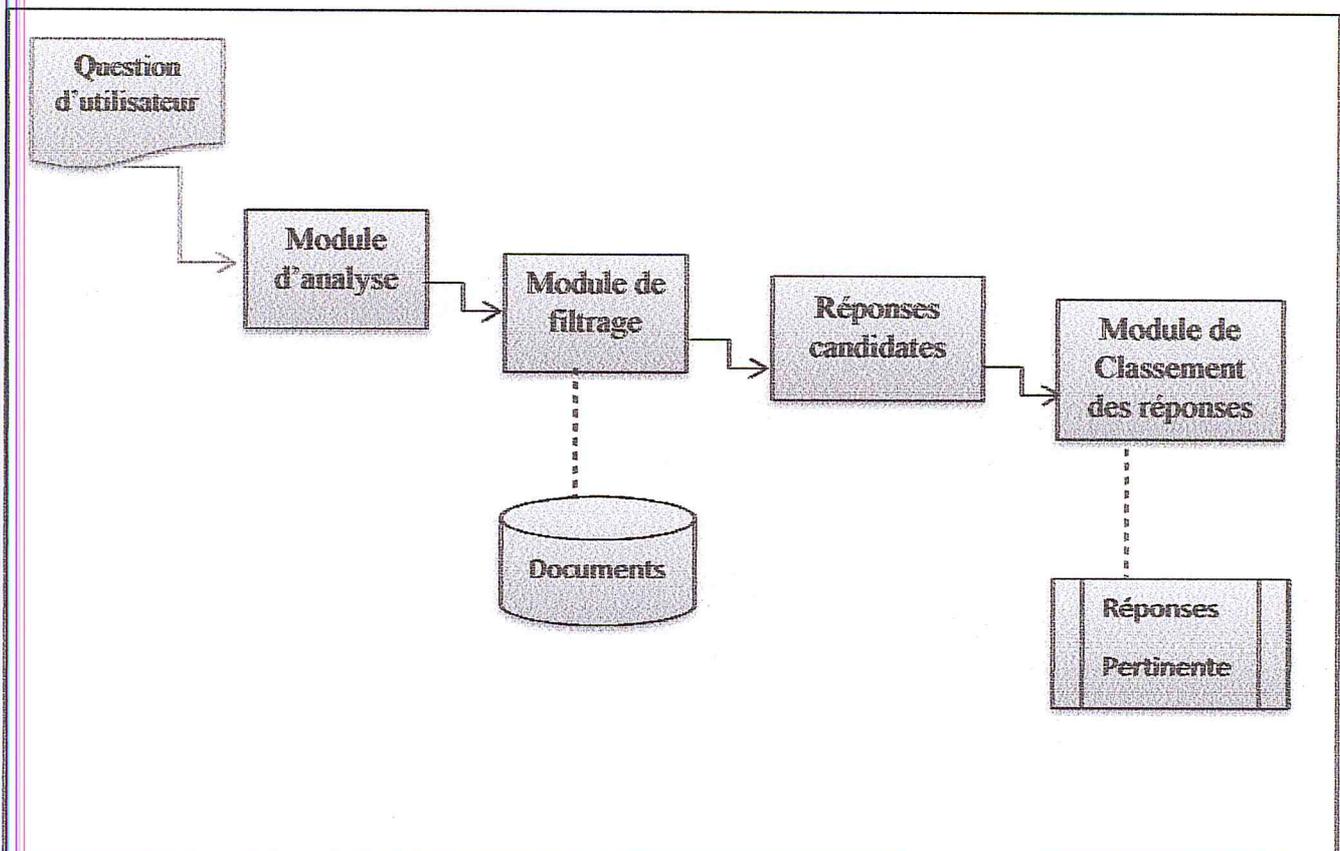


Figure 5: Architecture générale du système AQAS.

❖ **Al-Bayan (Abdelnasse, et al,2008) :**

Al-Bayan est un système de question/réponse, qui traite des questions exprimées en langue arabe concernant le coran. La source de la connaissance de ce système c'est le coran et ses livres d'interprétation(Tafsir).

Ce système récupère les versets du coran les plus pertinents, puis il fait l'extraction du passage qui contient la réponse (du coran et du livre d'interprétation).

Le système de QARAB atteint une précision de 85% en utilisant les trois premiers résultats.

Exemple des questions posées : « من هي ملكة سبأ »

La réponse :

فَكَتَبَ سُلَيْمَانُ كِتَابًا إِلَى بَلْقِيسَ مَلِكَةِ سَبَأَ، وَأَمَرَ الْمَلَائِكَةَ بِحَمَلِهِ إِلَيْهَا، وَبِالْقَائِدِ بَيْنَ يَدَيْهَا، ثُمَّ أَمَرَ
بِالْتَّحِي حَتَّى جَاءُوا بِالْمَلِكَةِ مَا سَعَتْهُ بِالْكِتَابِ، وَمَاذَا يَكُونُ رَدُّهَا عَلَيْهِ، فَحَمَلَتِ الْمَلَائِكَةُ الْكِتَابَ
إِلَيْهَا، وَالْقَائِدَ بَيْنَ يَدَيْهَا

1. Analyse de la question :

Cette étape se compose deux partie classification et prétraitement de la question, le système prend d'abord la question arabe qui est prétraitée pour en extraire la requête qui sera utilisée dans le module récupérer l'information. La question est également classée pour obtenir le type de la question, et par conséquent le type de la réponse attendue, qui sera ensuite utilisé dans le module d'extraction des réponses module.

2. Recherche d'information :

La question prétraitée sera envoyée au module de récupération de l'information qui récupère la plupart des versets sémantiquement liés du Coran et ses livres d'interprétation.

L'approche de ce système est basée sur une analyse sémantique explicite (Gabrilovitch et al, 2007) qui augmente les représentations des mots-clés de texte avec les concepts, qui sont extraites automatiquement à partir des dépôts de connaissances (qui est une ontologie), une ontologie Coranique qui classe les versets du Coran selon leurs sujets.

La technique utilisée d'apprentissage, pour construire un interpréteur sémantique comme dans (Gabrilovitch,et al ,2007) qui retrouve des fragments de texte en langage naturel dans un vecteur pondérée de concepts. Chaque concept de feuille dans l'ontologie a une liste de versets, qui sont liés à ce concept. Pour chaque concept de la feuille C_i , un document est construit D_i où D_i est

un document des versets et leur Tafsir qui appartiennent à C_i , puis on fait un prétraitement sur D_i qui au final va être indexé.

Chaque concept coranique sera représenté par un vecteur des termes qui apparaissent dans le document correspondant.

3. Extraction de la réponse :

Après l'extraction des versets pertinents, leur Tafsir et le type de réponse attendue ils sont tous envoyés à l'étape d'extraction pour extraire la réponse finale et répondre à la question d'entrée. Nous définissons la réponse comme une phrase qui contient une réponse attendue. La phase d'extraction de réponse se compose des étapes suivantes : d'abord, les entités nommées dans l'entrée question sont identifiées. Puis, plusieurs caractéristiques sont extraites qui sont utilisées pour classer chaque réponse candidate.

3.1.Reconnaissance des entités arabes REN :

La REN est une sous-tâche d'extraction d'information, où chaque nom propre entrée - tels que des personnes, des lieux et numéros est attribué une étiquette d'entité nommée.

L'architecture de système El-Bayan est représentée dans en dessous :

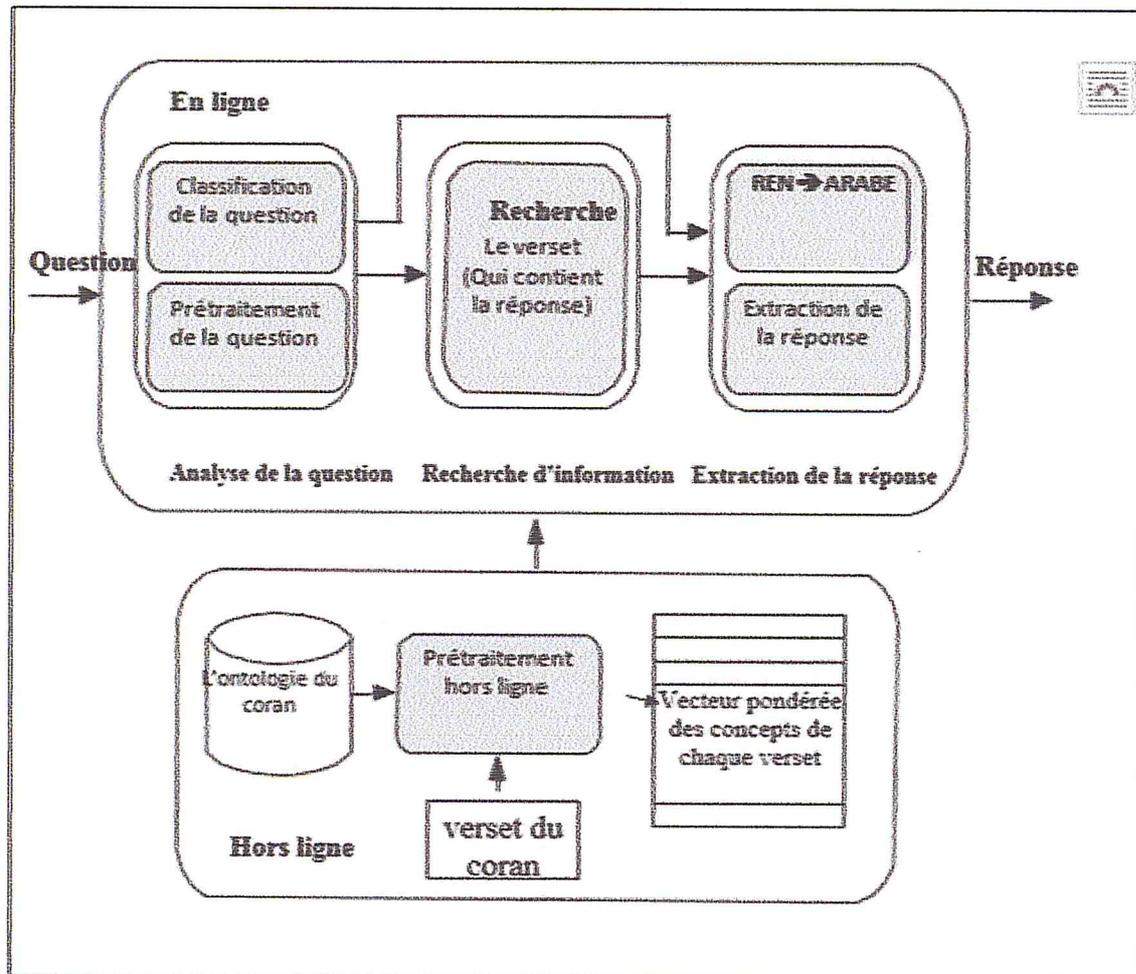


Figure 6 : L'architecture du système El Bayan

❖ **ArabiQA (Benajiba, et al., 2007) :**

Un système question de réponse, qui intègre un système de reconnaissance des entités nommés (REN) appelé ANERsys et il adapte l'arabe JIRS pour extraire des passages de documents arabes.

Il comprend :

- Un module d'extraction de texte,
- Un système de reconnaissance d'entités nommées (NER),
- Un module d'extraction de réponses dédié plus particulièrement aux questions types.

❖ **QASAL (Brini et al,2009) :**

Systeme Question /Réponse qui utilise des techniques de traitement du langage naturel pour le traitement des questions factuelles arabes

QASAL a une architecture pipeline constitué de trois composants :

- Analyse de la Question
- Recherche de Passage
- Modules d'extraction des réponses.

10. L'évaluation des SQR :

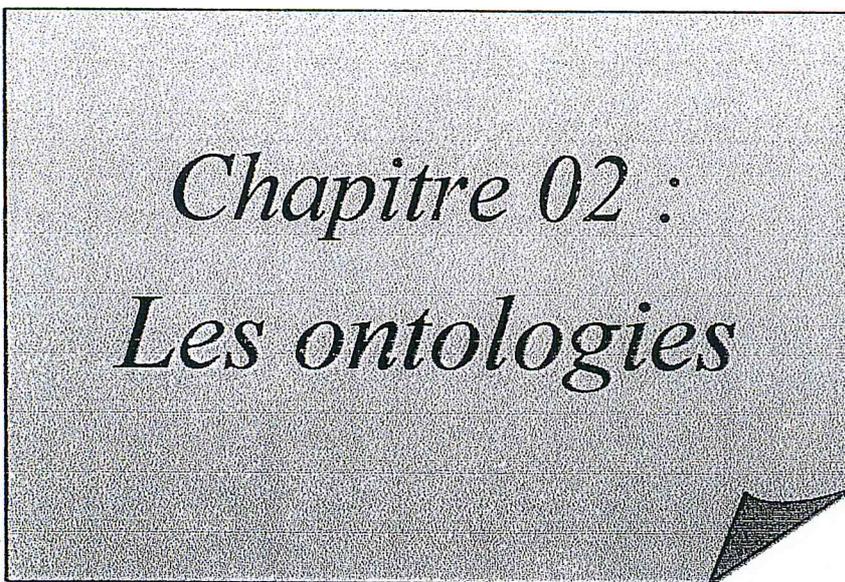
L'évaluation des SQR peut se faire au niveau de la satisfaction utilisateur (point de vue applicatif et qualitatif) ou par l'intermédiaire d'une métrique (point de vue comparatif car quantitatif). Les campagnes d'évaluations de SQR ont pour but de jauger les performances des différentes approches et proposent pour cela un nombre de questions significatif pour les catégories les plus fréquentes : factuelles, booléennes, définition, complexes, liste et Nil (questions n'ayant pas de réponses dans la collection de documents). Les systèmes doivent fournir plusieurs réponses pour chaque question (de trois à cinq généralement) et sont le plus souvent évalués grâce à la métrique du MRR (Mean Reciprocal Rank) qui favorise ainsi les SQR fournissant une réponse correcte dans les premiers rangs. Il est très difficile de garantir qu'une unique réponse correcte puisse être obtenue à partir de la collection de documents disponible pour l'évaluation, ce qui serait peu intéressant d'ailleurs, et une évaluation humaine des réponses doit parfois avoir lieu pour juger la réponse ainsi que le passage justificatif. Dans les campagnes traitant des question-listes (questions de type liste), pour indiquer qu'on n'attend pas une réponse unique, une marque de pluriel est toujours présente mais le nombre de réponses attendues n'est pas toujours mentionné dans la question comme dans les exemples suivants : Quelles sont les 4 localisations possibles des neuroblastomes ? (EQueR, Quaero 2008, 2009, TREC 2001, 2002) ou Quels sont les secteurs qui recrutent ? (Quaero 2010, TREC 2003 à 2007). La métrique utilisée pour évaluer les réponses à ce type de questions est alors dans le premier cas la précision moyenne (nombre de réponses correctes/nombre de réponses attendues) et dans le second la F-mesure (en considérant l'ensemble des réponses jugées correctes par les évaluateurs. Par l'utilisation du MRR, les campagnes analysant un triplet

question/réponse/passage obligent donc les SQR à faire un choix d'au plus N réponses par question. Une réponse issue d'un recoupement d'informations entre plusieurs documents est donc difficile à justifier dans le cadre d'une campagne d'évaluation. De plus, la réponse et le passage doivent obligatoirement être du texte issu d'un document de la collection alors qu'il peut être parfois plus pertinent de renvoyer un élément structural (un tableau par exemple). Ces éléments structuraux sont très présents dans les documents Web mais, de toutes les campagnes évoquées jusqu'à présent, seule Quaero utilise une collection de documents Web et impose un format de réponse assez identique à celui des autres campagnes (Mathieu,2014).

11. Conclusion :

Dans ce chapitre, nous avons présenté les systèmes de question réponse ainsi que l'intérêt de la recherche d'informations précises. Nous avons présenté quelques systèmes de question réponse, ces derniers s'appuient sur le même principe « 3 modules » :

Le premier sur l'analyse de la question, le deuxième sur la recherche et la sélection des documents, tandis que le dernier est l'extraction de la réponse. Cette étude va nous permettre de choisir quelle méthode du système question réponse nous allons utiliser pour construire notre propre système, pour l'interrogation du journal officiel en langue arabe.



Chapitre 02 :
Les ontologies

1. Introduction :

Après avoir donné une vision générale sur les systèmes de question réponses avec leurs architectures et quelques exemples sur ces derniers, maintenant nous allons définir les ontologies en donnant leurs classifications et présentant la méthode utilisée parmi celles qui existent pour construire l'ontologie du domaine juridique, et à la fin nous allons voir les langages de spécification d'ontologies.

2. Définition :

Une ontologie en informatique est un ensemble structuré de concepts permettant de donner un sens aux informations (GRUBER, 1993). Les concepts sont organisés dans un graphe dont les relations peuvent être :

- Des relations sémantiques.
- Des relations de composition et d'héritage (au sens objet).

L'objectif premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné.

À l'origine, l'ontologie est une notion philosophique, et nous considérons que l'étude de l'ontologie était une partie de la métaphysique, qui s'intéresse à l'étude des propriétés de l'être, et par extension de l'existence. Le mot ontologie provient du grec « onto » qui signifie ce qui existe, comme l'être, et l'existant, et « logos » qui veut dire étude. John McCarthy a introduit l'ontologie en intelligence artificielle en 1980, par le principe que les concepteurs de systèmes intelligents fondés sur la logique devraient d'abord établir une base de connaissances complète de ce qui existe, et ensuite utiliser ces données. Les ontologies sont apparues en intelligence artificielle, comme réponses aux problématiques de représentation et de manipulation des connaissances au sein des systèmes informatiques.

Dans le cadre de l'intelligence artificielle Neeches et ses collègues (Neeches, et.al, 1993) furent les premiers à proposer une définition à savoir « une ontologie définit les termes et les relations de base de vocabulaire d'un domaine, ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire ».

En 1993, (Gruber, et.al, 1993) propose la définition suivante : « une ontologie est une spécification explicite d'une conceptualisation ».

En 1995, (GUARINO, 1997) a modifié légèrement la définition de GRUBER, et la définit par : « une ontologie est une spécification partielle et formelle d'une conceptualisation ».

En 1997, ces deux dernières définitions sont regroupées dans celle de (BORST, 1997) comme « une ontologie est définie comme étant une spécification formelle d'une conceptualisation partagée ».

En 1998 (WEL, 1998) l'a défini comme suit : « Une ontologie est la définition de concepts, relations entre concepts, contraintes et règles d'inférences qui seront utilisés par un système de représentation des connaissances. »

- Conceptualisation : le modèle abstrait d'un phénomène du monde réel par identification des concepts clefs de ce phénomène.
- Explicite : le type de concepts utilisés et les contraintes liées à leur usage sont définis explicitement.
- Formelle : une ontologie doit être traduite en langage interprétable par une machine.
- Partagée : capture la connaissance consensuelle, c'est-à-dire non réservée à quelques individus, mais partagée par un groupe ou une communauté.

3. Les différents types d'ontologies :

Uschold. (Uschold, et. al, 1996) détermine trois dimensions pour la classification des ontologies :

3.1 Selon le degré de formalisme :

M. Uschold et M. Grüninger ont identifié quatre types d'ontologies :

Les ontologies hautement informelles, les ontologies informelles, les ontologies semi formelles et les ontologies rigoureusement formelles (Uschold, et. al, 1996).

3.1.1 Les ontologies hautement informelles : exprimées en langage naturel.

3.1.2 Les ontologies semi informelles : Elles sont exprimées sous une forme limitée, restreinte et structurée du langage naturel (en utilisant des modèles), c'est à dire des patrons ont été mis en œuvre.

3.1.3 Les ontologies semi formelles : exprimées dans un langage défini artificiellement et formellement.

3.1.4 Les ontologies rigoureusement formelles : exprimées dans un langage contenant une sémantique formelle, des théorèmes et des preuves de propriétés telles que la robustesse, l'exhaustivité, la complétude et la consistance. La plupart des ontologies sont aujourd'hui implémentées en langage formel (Ontolingua, CycL, Loom, Flogic). Une classification de ce type d'ontologies peut se définir soit en « intension » ou en « extension » :

- Une ontologie formalisée en extension (GRUBER T,1993), a sa définition formelle avec une sémantique déclarative (vocabulaire, thésaurus, etc.). Elle est constituée d'un ensemble de « lexons », expressions élémentaires construites d'un élément de contexte, de terme et de rôle.
- Une ontologie formalisée en intension (GRUBER T,1993) possède une définition à travers des « mondes possibles » (thésaurus de domaine vus comme une synthèse organisée des arrangements de termes possibles linguistiquement).

3.2 Selon les objets modélisés :

Les ontologies ont été aussi regroupées dans (GÓMEZ-PÉREZ, et.al,2004) en se basant sur les objets modélisés par l'ontologie afin de répondre à un but précis.

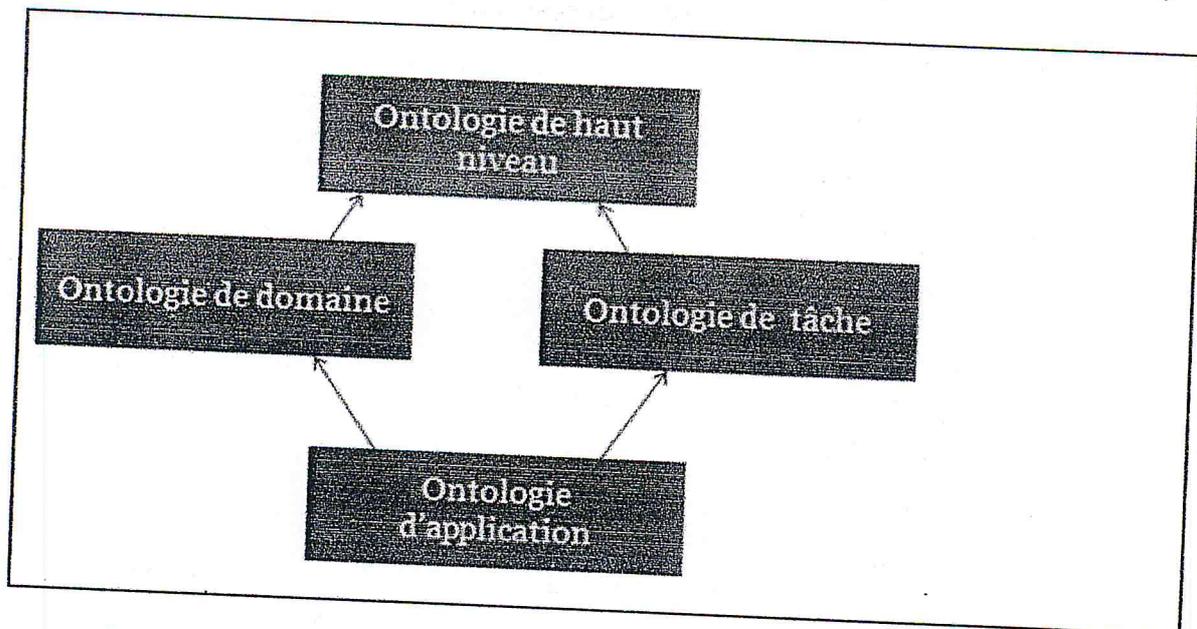


Figure 8 : Classification des ontologies selon (GUARINO, et.al,1994).

Différents types d'ontologie selon leur degré de dépendance vis-à-vis d'une tâche particulière ou d'un point de vue (Les flèches représentent des relations de spécialisation).

3.2.1 Les ontologies supérieures :

Dites aussi de haut niveau « top level ontologies ou Upper Level ontologies », ontologies « génériques ou noyaux d'ontologies », « méta ontologie », « ontologies de sens commun/général » : décrivent les concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc., qui sont indépendants d'un problème ou d'un domaine d'application particulier. (Kaveh, 2004), Parmi les ontologies de haut niveau : Sumo, ontolingua, Il n'existe pas pour le moment d'ontologies de haut niveau unifiées.

3.2.2 Les ontologies de domaine (Domain ontologies) et les ontologies de tâche (task ontologies) :

Décrivent, respectivement, le vocabulaire lié à un domaine générique (comme la médecine, ou les automobiles) ou une tâche ou une activité générique (comme le diagnostic ou la vente), en spécialisant les concepts présentés dans les ontologies de hauts niveaux. Elles donnent une représentation formelle des concepts du domaine étudié ainsi que des différentes relations qui lient ces derniers, elle ne contient pas les concepts pédagogiques, narratifs et structurels (Kaveh, 2004).

3.2.3 Les ontologies d'application :

Décrivent des concepts qui dépendent à la fois d'un domaine et d'une tâche particulière, qui sont souvent des spécialisations des deux ontologies relatives. Ces concepts correspondent souvent aux rôles joués par des entités de domaine tout en exécutant une certaine activité, comme l'unité remplaçable ou le composant disponible.

Ces différents niveaux sont récapitulés ci-dessous :

Par conséquent, une ontologie peut être vue comme une théorie qui distingue les concepts particuliers, c'est à dire les objets concrets, physiques, les événements, les régions, etc., et les concepts universels c'est à dire les propriétés, rôles, relations, états, ...etc. (Kaveh, 2004)

3.2.4. Les ontologies de représentation :

Ce type d'ontologies est un cas particulier d'ontologies supérieures qui regroupe des concepts déjà utilisés pour formaliser les connaissances. Indépendamment des domaines (GUARINO, et.al,1994) puisqu'elles décrivent des primitives cognitives communes. Parmi les ontologies de représentation, on trouve la « Frame-Ontology» qui définit, de manière formelle, les concepts utilisés principalement dans des langages à base de frames : classes, sous classes, attributs, valeurs, relations et axiomes (GRUBER T,1993).

3.2.5. Les ontologies de raisonnement :

Regroupe les processus de raisonnement appliqués aux connaissances qui forment eux-mêmes un domaine de connaissance. Nous parlons particulièrement d'ontologies développées pour représenter des connaissances génériques mises en œuvre lors de la résolution automatique de problèmes.

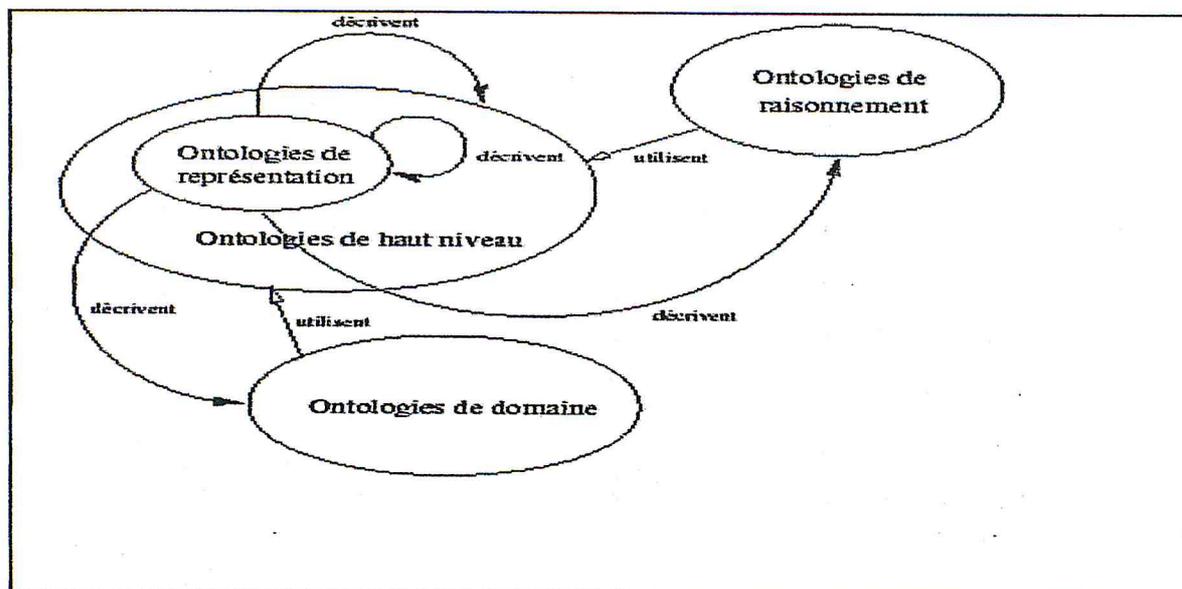


Figure 9 : Schéma représentatif des relations entre les ontologies de domaine, de raisonnement, de représentation et de haut niveau. (F. FÜRST,2004).

3.3 Selon la granularité :

La classification suivante est en fonction du degré de granularité, c'est-à-dire quel niveau de détail des objets de la conceptualisation est préconisé. En fonction de l'objectif opérationnel, une connaissance plus ou moins fine du domaine est nécessaire et des propriétés considérées comme accessoires dans un contexte peuvent se révéler indispensables dans un autre :

3.3.1 Granularité fine : correspondant à des ontologies très détaillées, possédant ainsi un vocabulaire plus riche capable d'assurer une description détaillée des concepts pertinents d'un domaine ou d'une tâche (F. FÜRST,2004).

3.3.2 Granularité large : correspondant à un vocabulaire moins détaillé. Les ontologies de haut niveau ont une granularité large, du fait que les notions sur lesquelles elles portent peuvent être raffinées par des notions plus spécifiques (F. FÜRST,2004).

4. Méthode de construction d'une ontologie :

Il existe différentes méthodes de développement d'une ontologie :

- La méthode « Enterprise Ontology » (Uschold, et. al, 1996).
- La méthode « TOVE » (Uschold, et. al, 1996).
- La méthode « METHONTOLOGY » (Gilles, 2010).
- La méthode de « Stanford » (Natalya, 2010).

Parmi les méthodes qui existent nous avons choisis la méthode développée par l'université de Stanford, car nous voyons que pour un domaine fermé elle est facile à appliquer et bien détaillée, nous montrons en détail les étapes à suivre.

4.1 La méthode développée par l'université de Stanford :

Cette méthode comporte sept étapes qui sont les suivantes :

✓ Etape 1 : Déterminer le domaine et la portée de l'ontologie

Cette étape se fait en répondant aux questions ci-dessous tout au long de la conception de l'ontologie et qui aident à définir la portée du domaine de l'ontologie :

- Quel est le domaine que va couvrir l'ontologie ?
- Dans quel but utiliserons-nous l'ontologie ?
- A quels types de questions l'ontologie devra-t-elle fournir des réponses ?
- Qui va utiliser et maintenir l'ontologie ?

✓ Etape 2 : Envisager une éventuelle réutilisation des ontologies existantes :

Dans tout domaine de recherche, il est utile de profiter de ce que les autres ont fait afin d'en tirer les informations et ainsi permettre d'élargir le travail et l'affiner pour répondre à nos propres besoins.

Il peut être intéressant d'importer des ontologies déjà existantes (dans le même domaine) et les raffiner et les perfectionner pour aboutir à une ontologie plus complète et étendue.

✓ **Etape 3 : Enumérer les termes importants dans l'ontologie :**

Il est important d'établir en premier lieu une liste complète des mots et termes concernant le domaine d'intérêt, et cela sans se soucier de la catégorisation de ces derniers dans des classes, hiérarchie, chevauchement, etc. les questions qui se posent pour établir cette liste sont :

- Sur quels termes souhaiterons-nous discuter ?
- Quelles sont les propriétés de ces termes ?
- Que veut-on dire sur ces termes ?

✓ **Etape 4 : Définir les classes et la hiérarchie des classes :**

A partir de la liste de l'étape précédente, on commence par définir les classes en sélectionnant les termes qui décrivent des objets ayant une existence indépendante. Ce sont ces termes qui constitueront les classes (appelées parfois concepts) de l'ontologie. Il faut ensuite organiser ces classes dans une taxonomie hiérarchique.

✓ **Etape 5 : Définir les propriétés des classes (les attributs ou rôles) :**

Les classes seules ne fournissent pas suffisamment d'information pour répondre aux questions de compétence de l'étape 1. Une fois que nous avons défini les classes, il faut décrire la structure interne des concepts.

✓ **Etape 6 : Définir les facettes des attributs :**

Les attributs peuvent avoir plusieurs facettes (appelées parfois restrictions de rôles). Les facettes les plus communes décrivent :

- Le type de valeur des attributs : désigne le type de valeur pouvant être affecté à un attribut. Les plus typiques sont les suivants : chaîne de caractère, nombre ou entier, booléen...
- Le nombre de valeurs ou cardinalité : désigne le nombre de valeurs qu'un attribut peut avoir. Une cardinalité peut être unique ou multiple. Il est utile de spécifier pour un attribut une cardinalité minimale et une cardinalité maximale.
- Le rang et le domaine d'un attribut : l'étendue ou le rang d'un attribut représente les classes autorisées pour les attributs de type « Instance ». Le domaine d'un attribut représente les classes auxquelles cet attribut est rattaché ou les classes dont l'attribut décrit les propriétés.

✓ **Etape 7 : Créer les instances :**

Cette étape consiste à créer les instances qui représentent des entités réelles des classes

5. Les Langages de spécification d'ontologies :

Plusieurs langages de représentation d'ontologies ont été développés pendant les dernières années. Certains d'entre eux sont basés sur la syntaxe de XML, tels que XOL (Ontology Exchange Language), SHOE (Simple HTML Ontology Extension - qui a été précédemment basé sur le HTML), OML (Ontology Markup Language), RDF (Resource Description Framework), RDF Schéma. Les deux derniers sont des langages créés par des groupes de travail du World Wide Web Consortium (W3C).

En conclusion, trois langages additionnels sont établis sur RDF(S) pour améliorer ses caractéristiques : OIL (Ontology Inference Layer), DAML+OIL et OWL (Web Ontology Language). La figure ci-dessous présente des langages de spécification d'ontologie, qui ont été récemment développés.

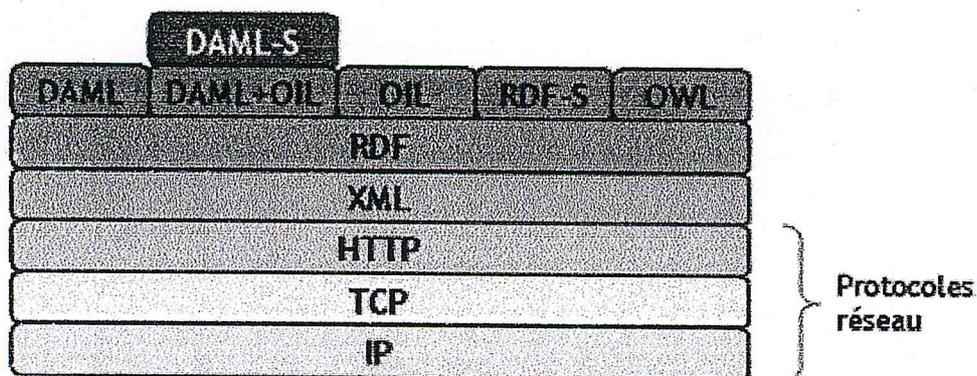


Figure 10 : Les langages de spécification de l'ontologie.

- **RDF ()** : est un langage d'assertion et d'annotations. Les assertions affirment l'existence de relations entre les objets. Elles sont donc adaptées à l'expression des annotations que nous le voulons associer aux ressources du Web. RDF est un langage formel qui permet d'affirmer des relations entre des « ressources ». Un document RDF est un ensemble de triplets de la forme < ressource, Propriétés, Valeurs >. ¹

¹<http://www.w3.org/RDF/>.

- **Ressource** : Représente la ressource décrite, i.e. tout document accessible sur le Web comme les pages HTML, les documents textuels (PDF, Ms Word) ou multimédias (images, vidéo), etc., mais aussi tout objet, abstrait ou non, du monde réel. Les ressources sont nommées en utilisant une URI.
- **Propriétés** : Représente la propriété descriptive, i.e. une caractéristique spécifique, un attribut ou une relation, utilisée pour décrire une ressource.
- **Valeurs** : Les valeurs en question sont les valeurs particulières que prennent les propriétés. Ces trois types d'objets peuvent être mis en relation par des assertions, c'est à dire des triplets (ressource, propriété, valeur), ou encore (sujet, prédicat, objet).
- **RDF(S)** : Comme son nom l'indique, RDFS a pour but de définir des schémas de métadonnées. Il définit le sens, les caractéristiques et les relations d'un ensemble de propriétés. La définition peut inclure des contraintes pour les valeurs potentielles et l'héritage des propriétés d'autres schémas. Il est, en effet, une extension sémantique de RDF afin de fournir un mécanisme pour décrire les groupes associés de ressources et les relations entre les ressources.² L'intérêt de RDFS est qu'il facilite l'inférence sur les données et renforce la recherche sur ces données.
- **DAML-OIL** : DAML est un langage qui a comme but de fournir les fondations pour la génération suivante du Web sémantique. Le langage a adopté d'abord RDFS comme langage d'ontologie pour l'interopérabilité sémantique entre projets. Comme RDFS n'est pas assez expressif relativement aux exigences du Web sémantique, un nouveau langage nommé DAML-ONT a été développé en tant qu'extension de RDF avec les capacités d'un langage de représentation du savoir : orienté objet et basé sur cadre. En même temps, un groupe des chercheurs (la plupart d'entre eux sont européens) au sein d'IST d'Union européenne, avec le même but, développe un autre langage d'ontologie appelé OIL. Ce langage a une syntaxe basée sur RDF et il est explicitement construit pour que sa sémantique puisse être spécifiée à travers une description logique très expressive, la logique de description de type SHIQ.

²RDF-Schema: <http://www.w3.org/TR/rdf-schema/>.

DAML+OIL est la combinaison de ces deux langages. Il hérite des avantages de ces deux langages. En conséquence, DAML+OIL est un langage très expressif et lisible par la machine ainsi que par un être humain avec une syntaxe basée sur RDF.³

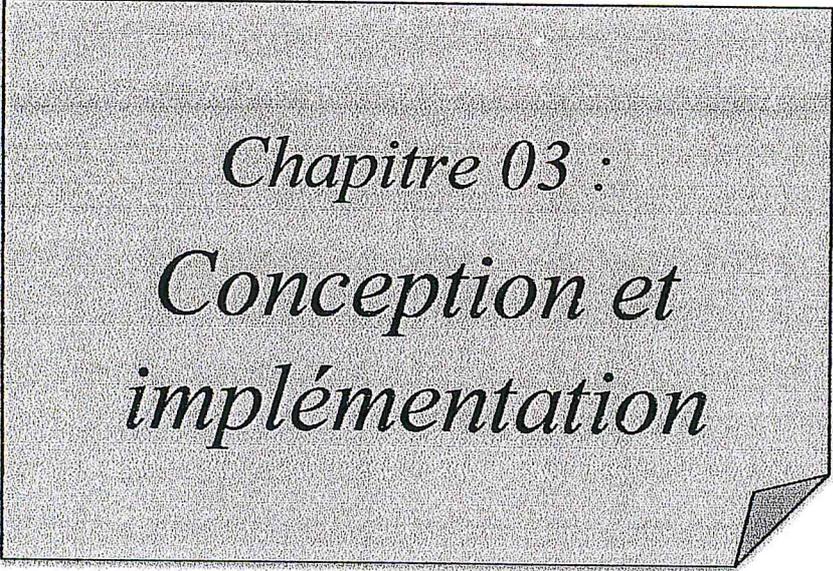
- **Owl** : OWL (Ontology Web Language) a été conçu pour être utilisé par les applications qui traitent le contenu de l'information. Il facilite grandement l'interopérabilité en fournissant plus de vocabulaire pour décrire les classes et les propriétés comme des approches orientées objets. Par exemple : les relations entre les classes, les cardinalités, l'égalité, le typage plus riche des propriétés, les caractéristiques des propriétés, etc. OWL a été conçu pour satisfaire le besoin d'un langage d'ontologie du web. Il ajoute plus de vocabulaire pour décrire les propriétés et les classes. On peut citer entre autres : les relations entre classes (par exemple la disjonction), les cardinalités (par exemple exactement un), l'égalité, typage plus riche des propriétés, caractéristiques des propriétés (par exemple la symétrie) et les classes énumérées. OWL fournit trois sous langages de plus en plus expressifs conçus pour l'usage des communautés spécifiques des utilisateurs et des développeurs : OWL Lite, OWL DL et OWL Full. Le langage OWL DL correspond à une logique de description expressive. Il convient aux utilisateurs qui demandent un maximum d'expressivité tout en maintenant la complétude (garantie de calculer toutes les conclusions) et la décidabilité (tous les calculs doivent fournir en un temps fini). OWL DL inclut tous les constructeurs du langage OWL, mais ils peuvent être utilisés seulement sous certaines restrictions.

6. Conclusion :

Dans ce chapitre nous avons défini les ontologies avec leurs caractéristiques (en remarquant que les ontologies sont partagées, évolutives, réutilisable et facile à utiliser), leurs constructions et leurs méthodes d'application, ainsi que les domaines d'application de ces ontologies. Nous avons vu aussi quelques langages de spécification d'ontologie.

Dans le chapitre suivant, nous allons présenter la conception du journal officiel avec le formalisme des ontologies, sur lequel nous développons notre système de questions réponses.

³The DARPA Agent Markup Language. <http://www.daml.org/>.



Chapitre 03 :
Conception et
implémentation

1. Introduction :

La conception des systèmes de questions-réponses nécessite une analyse profonde des questions posées. Cette tâche primordiale requiert d'être étudiée et évaluée séparément.

Dans le premier chapitre nous avons introduit les systèmes de questions réponses et présenté les différents systèmes qui existent ainsi que les approches utilisées pour développer un système de question-réponse.

Nous avons vu aussi que y'a des modules essentiels communs entre tous les systèmes, dans ce chapitre nous allons présenter l'architecture de notre système, ainsi que ses principaux modules, et l'approche que nous avons utilisée pour extraire les termes importants de la question (l'approche à base de patrons), pour entamer par la suite l'étape d'extraction des réponses. Avant ça nous allons montrer la conception et l'implémentation de l'ontologie du journal officiel.

2. Conception et implémentation de l'ontologie du journal officiel :

2.1 Structure du Journal Officiel "JO" :

À partir de l'étude du journal officiel et de sa hiérarchie, nous avons conçu et développer une ontologie pour représenter ce dernier pour bénéficier des différents avantages des ontologies.

Au sommet de la pyramide nous trouverons la constitution qui comprend ces classes :

- Lois (lois organiques et lois ordinaires).
- Ordonnances.
- Décrets (décrets présidentiels, décrets exécutifs et décrets législatifs).
- Textes internationaux (traités et conventions).
- Arrêtés (arrêtés ministériels et arrêtés interministériels).
- Décisions.
- Avis.
- Circulaires.
- Instructions.

2.2 Construction de notre ontologie (méthode de Stanford) :

En partant de la hiérarchie du journal officiel, nous avons établi selon la méthode de Stanford, nous avons conçu une ontologie pour le journal officiel en langue arabe.

Glossaire des termes :

Les termes en français	Les termes en arabe
Constitution	الدمستور
Textes internationaux	النصوص الدولية
Traités	اتفاقيات
Conventions	معاهدات
Décrets	مرسوم
Décrets exécutifs	مرسوم تنفيذي
Décrets présidentiels	مرسوم رئاسي
Décrets législatifs	مرسوم تشريعي
Arêtes	قرار
Arêtes interministériels	قرار لا وزاري
Arêtes ministériels	قرار وزاري
Avis	رأي
Décisions	قرار
Règlementation	نظام
Ordonnances	أمر
Lois	قانون
Lois ordinaires	قانون عادي
Lois organiques	قانون عضوي
Circulaires	متشور
Instructions	مقرر

Tab 1 : les classes de notre ontologie.

Les propriétés :

Nous avons basé dans la recherche sur ces propriétés : mot clé, ministère, secteur, date de publication, synonymes, signée par, pays etc...

Classes :	Définitions :	Propriétés :	Super Classe :
Traits	الإتفاقية الدولية هي إتفاق مكتوب بين دولتين أو أكثر من شأنه أن ينشئ حقوقا متبادلة في ظل القانون الدولي العام	Pays- date_de_publication- modifié-complété- signee_par-num- secteur...	Texts internationaux
Lois organiques	القانون العضوي هو قانون بشأن تنظيم السلطات	Date_de_publication -signee_par-num- ministère- secteur ...	Lois
.....

Tab.2 : les classes et les propriétés de l'ontologie.

Les relations :

Dans cette étape, nous avons relié les classes avec la relation « is-a » qui signifie « partie-de ».

Les instances :

Dans notre cas, une instance est un texte juridique.

2.3 L'implémentation de l'ontologie :

Nous avons implémenté notre ontologie dans protégé-2000

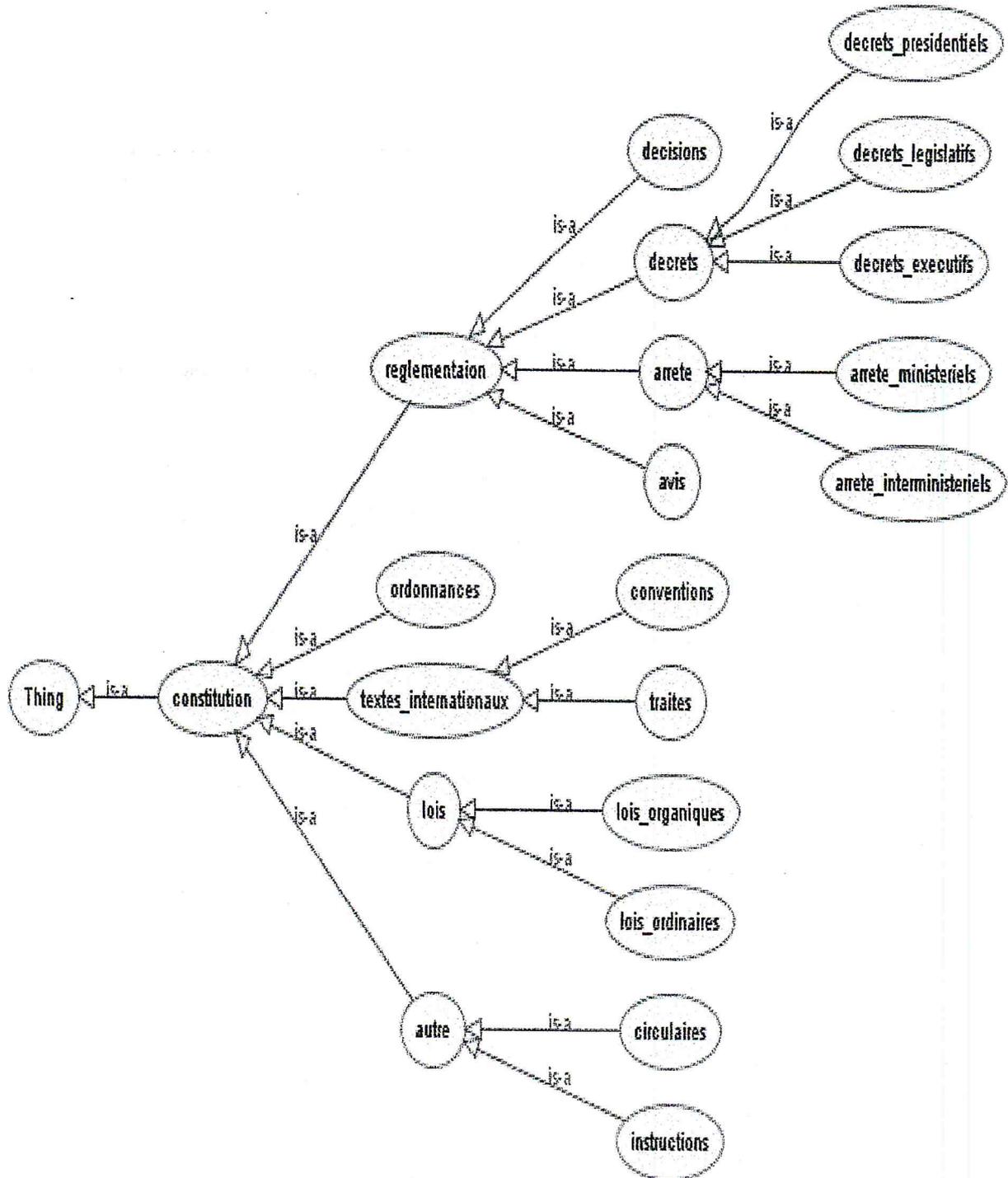


Figure11 : Taxonomie de l'ontologie du journal officiel "JORADP" 'Onto JORADP-

3. Conception et implémentation du système de question/réponse.

Nous allons maintenant présenter notre système de questions réponses qui va nous permettre d'interroger en langage naturel l'ontologie que nous avons développé pour le journal officiel arabe.

3.1 L'approche proposée :

Nous avons vu dans le premier chapitre que les systèmes de questions/réponses peuvent être différenciés selon l'approche utilisée, pour notre cas, nous proposons une approche à base de patrons pour répondre aux questions juridiques.

Cette approche se fonde sur des patrons construits manuellement pour chaque type de question. Un patron est une expression régulière décrivant un modèle de phrase où les entités sont présentes à des emplacements spécifiques (encadrés par des mots ou expressions régulières).

D'une autre façon « L'approche à base de patrons » consiste à associer à la question un ensemble de patrons définis manuellement. Elle se base généralement sur certains mots utilisés dans les questions et d'autres recherches.

3.2 Pourquoi l'approche à base de patrons :

Cette approche semble être la plus adaptée pour une solution à notre problématique, en comparaison aux autres approches concernant l'analyse des questions et l'extraction de réponses car :

- ✓ Notre système est conçu pour un domaine fermé (le domaine juridique uniquement) qui travaille sur une collection de documents restreints quant au sujet et à la quantité, donc ça nous convient pour ce domaine.
- ✓ Nous pouvons constater que le nombre de questions qui peuvent être posées est limité, et tout changement dans les éléments de la question ne peut influencer sur les patrons mis en place.

4. Typologie des questions proposées :

Nous avons proposé trois types de questions selon la réponse attendue, dont chacun a ses propres patrons. Ces types sont suivants :

- Un texte du journal officiel. (Classe 1)
- Une date de publication dans le journal officiel d'un article. (Classe 2)
- Une personne responsable sur l'exécution d'un article. (Classe 3)

Ces patrons contiennent deux parties essentielles qui sont :

1^{ère} partie désigne seulement le nom de la classe. (Elle désigne nécessairement une classe).

2^{ème} partie désigne les propriétés sur lesquelles nous faisons le filtrage dans la classe récupérée.

5. Patrons lexico-syntaxiques :

La récupération d'un texte d'une classe donnée avec des propriétés spécifiques, en ignorant le verbe utilisé par l'utilisateur.

Nous appliquons ces patrons afin d'extraire le concept et les propriétés (propriété 1 « mot clé », propriété 2, « date de publication », propriété 3 « ministère », propriété 4 « signée par », propriété 5 « pays » et propriété 6 « wilaya »), définies dans notre ontologie (voir chapitre2)

Cette typologie de questions va donner lieu à trois classes de patrons

Classe 1 : le type de réponse sera un texte du journal officiel, souvent avec un passage qui contient la réponse. Nous montrons par la suite les patrons de cette classe :

Patron 1 :

أريد المرسوم الرئاسي الذي ينص على السرقة سنة 2012 ؟

أريد < nom de la classe > < الذي > "verbe utilisé" < على > < Propriété 1 > < سنة > < Propriété 2 >

Patron 2:

أريد التعلیمة التي تنص على التعلیم سنة 2015 ؟

أريد < nom de la classe > < التي > "verbe utilisé" < على > Propriété 1 < سنة > Propriété 2

Patron 3:

أريد قانون عضوي الذي يتناول موضوع الدراسات العليا سنة 2016 ؟

أريد < nom de la classe > < الذي > "verbe utilisé" < موضوع > Propriété 1 < سنة > Propriété 2

Patron 4:

أريد تعلیمة التي تتناول موضوع الدراسات العليا سنة 2016 ؟

أريد < nom de la classe > < التي > "verbe utilisé" < موضوع > Propriété 1 < سنة > Propriété 2

Patron 5:

أريد القانون الذي يتعلق بكوسيدار سنة 2014 ؟

أريد < nom de la classe > < الذي > "verbe utilisé" < ب > Propriété 1 < سنة > Propriété 2

Patron 6:

أريد التعلیمة التي تتعلق بكوسيدار سنة 2014 ؟

أريد < nom de la classe > < التي > "verbe utilisé" < ب > Propriété 1 < سنة > Propriété 2

Patron 7:

Les mêmes patrons précédents sauf que le verbe de parole change « أريد », dans notre cas les autres verbes de demande sont : (chaque verbe est un patron différent de l'autre ' en tout, nous récolterons 30 patrons de plus')

ماهي هل يوجد، ما هو، أعطيني،

Pour ces modèle, l'utilisateur peut aussi ne pas préciser la propriété 2 « date de publication », donc nous aurons ces sous-patrons qui sont les suivants : (nous prenons pour un exemple les questions précédentes).

أريد < nom de la classe > < الذي > "verbe utilisé" < ب > Propriété 1

أريد < nom de la classe > < الذي > "verbe utilisé" < ب > Propriété 1

Propriété 1 <أريد < nom de la classe < الذي < "verbe utilisé" < على <

Propriété 1 <أريد < nom de la classe < التي < "verbe utilisé" < على <

Propriété 1 <أريد < nom de la classe < التي < "verbe utilisé" < موضوع <

Propriété 1 <أريد < nom de la classe < التي < "verbe utilisé" < موضوع <

Patron 8: dans ce cas, nous avons trois patrons différents, et les deux propriétés désignent les noms des pays (propriété définie dans l'ontologie).

أريد نص دولي بين الجزائر و الصين ؟

Propriété 5 <أريد < nom de la classe < بين < Propriété 5 < و <

ماهي الاتفاقية بين الجزائر و الصين ؟

Propriété 5 <أريد < nom de la classe < بين < Propriété 5 < و <

أعطيني المعاهدات بين الجزائر و الصين ؟

Propriété 5 <أعطيني < nom de la classe < بين < Propriété 5 < و <

Patron 9:

(أريد/ أعطيني/ ما هو/ هل يوجد) المرسوم الخاص بالشرطة ؟

Propriété 1 <أريد < nom de la classe < الخاص <

Propriété 1 <أريد < nom de la classe < الخاصة <

Propriété 1 <أعطيني < nom de la classe < الخاص <

Propriété 1 <أعطيني < nom de la classe < الخاصة <

Propriété 1 <ما هو < nom de la classe < الخاص <

Propriété 1 <ما هو < nom de la classe < الخاصة <

Propriété 1 <ماهي < nom de la classe < الخاصة <

Propriété 1 <هل يوجد < nom de la classe < الخاص <

Propriété 1 <هل يوجد < nom de la classe < الخاصة <

Patron 10: nous avons douze patrons, nous obtiendrons les autres patrons en remplaçant le verbe « أريد » par le mot « أعطيني » et le mot « ماهي », qui seront des nouveaux patrons. Et la propriété 3 désigne la propriété « ministère » dans l'ontologie.

أريد المرسوم التنفيذي لوزارة الدفاع التي ينص على الاسلحة سنة 2016 ؟

أريد جميع <nom de la classe> لوزارة < Propriété 3 > التي < "verbe utilisé" > على < Propriété 1 > سنة < Propriété 2 >

Dans ce cas, nous avons trois sous patrons que nous pouvons extraire de celui-là, ils sont :

أريد جميع المراسيم ؟

أريد جميع < Nom de la classe >

أريد جميع المراسيم لوزارة الدفاع ؟

أريد جميع < nom de la classe > لوزارة < Propriété 3 >

أريد جميع المراسيم التنفيذية لوزارة الدفاع التي نتحدث على الاسلحة ؟

أريد جميع < nom de la classe > لوزارة < Propriété 3 > التي < "verbe utilisé" > على < Propriété 2 >

Patron 11: le même cas précédent mais il est destiné pour d'autres verbes (douze patrons aussi)

أريد جميع القوانين العادية لوزارة العدل التي تتعلق بالعقوبات سنة 2010 ؟

أريد جميع < nom de la classe > لوزارة < Propriété 3 > التي < "verbe utilisé" > ب < Propriété 1 > سنة < Propriété 2 >

Patron 12: le même principe, nous avons six patrons que nous les aurons en remplaçant le verbe « أريد » par le mot « أعطيني » et le mot « ماهي », qui seront des nouveaux patrons. Et la propriété 4 désigne la propriété « signée _par » dans l'ontologie.

أريد الاتفاقية التي امضاها الرئيس عبد العزيز بوتفليقة سنة 2014 ؟

أريد < nom de la classe > التي امضاها < Propriété 4 > سنة < Propriété 2 >

Patron 1:

متى حرر القانون العادي الذي يتكلم على أمن البلاد ؟

Propriété 1 < على > "verbe utilisé" < الذي > nom de la classe < متى حرر >

متى حرر القانون العادي الذي يتعلق بأمن البلاد ؟

Propriété 1 < ب > "verbe utilisé" < الذي > nom de la classe < متى حرر >

Patron 2:

Propriété 1 < على > "verbe utilisé" < التي > nom de la classe < متى حررت >

Propriété 1 < ب > "verbe utilisé" < التي > nom de la classe < متى حررت >

Une fois, le texte candidat récupéré, nous appliquerons sur ce dernier un patron afin d'extraire l'information voulue, qui est la date de publication. Ce patron est de la forme suivante :
Propriété 2 < حرر بالجزائر في > .

Le deuxième sous type consiste à extraire le nom de la personne qui exécutera un article dans un document d'une classe et propriété donnée.

Classe 3 : le type de réponse sera une personne responsable sur l'exécution d'un article.

Patron 1:

من يكلف بتنفيذ المرسوم الذي يتحدث على المطار سنة 2016 ؟

Propriété 2 < سنة > Propriété 1 < على > "verbe utilisé" < الذي > nom de la classe < من يكلف بتنفيذ >

Patron 2:

من يكلف بتطبيق المرسوم الذي يتحدث على المطار سنة 2016 ؟

Propriété 2 < سنة > Propriété 1 < على > "verbe utilisé" < الذي > nom de la classe < من يكلف بتطبيق >

Voilà leurs sous patrons :

Patron 3:

Propriété 1 < على > “verbe utilisé” < الذي > nom de la classe < من يكلف بتنفيذ >

Propriété 1 < ب > “verbe utilisé” < الذي > nom de la classe < من يكلف بتنفيذ >

Patron 4:

Propriété 1 < على > “verbe utilisé” < الذي > nom de la classe < من يكلف بتطبيق >

Propriété 1 < ب > “verbe utilisé” < الذي > nom de la classe < من يكلف بتطبيق >

Patron 5:

Propriété 1 < على > “verbe utilisé” < التي > nom de la classe < من يكلف بتنفيذ >

Propriété 1 < ب > “verbe utilisé” < التي > nom de la classe < من يكلف بتنفيذ >

Patron 6:

Propriété 1 < على > “verbe utilisé” < التي > nom de la classe < من يكلف بتطبيق >

Propriété 1 < ب > “verbe utilisé” < التي > nom de la classe < من يكلف بتطبيق >

Une fois nous aurons le texte candidat, nous allons appliquer sur ce dernier un patron pour avoir l'information exacte, qui est le nom de la personne. Ce patron est de la forme suivante :

< يكلف > réponse < كل > , < يكلف > réponse < بتنفيذ >

5.1 L'expansion de la requête :

L'expansion de requête peut être définie comme un élargissement du champ de recherche pour cette requête. En faisant ça pour obtenir des résultats pertinents, c'est une autre manière pour augmenter le taux de précision.

أريد جميع الاتفاقيات التي امضاها الرئيس عبد العزيز بوتفليقة سنة 2014 ؟

Propriété 2 < سنة > Propriété 4 < التي امضاها > nom de la classe < أريد جميع >

Patron 13: la propriété 6 désigne celle de la « wilaya » dans l'ontologie, nous obtiendrons d'autres patrons en remplaçant le verbe « أريد » par le mot « أعطيني » et le mot « ماهي », qui seront les nouveaux patrons. En tous nous avons douze patrons.

أريد تعليمة لولاية البلدية سنة 2015 ؟

Propriété 2 < سنة > Propriété 6 < لولاية > nom de la classe < أريد >

أريد جميع التعليمات لولاية البلدية سنة 2015 ؟

Propriété 2 < سنة > Propriété 6 < لولاية > nom de la classe < أريد جميع >

Les sous patrons sont :

أريد تعليمة لولاية البلدية ؟

Propriété 6 < لولاية > nom de la classe < أريد >

أريد جميع التعليمات لولاية البلدية ؟

Propriété 6 < لولاية > nom de la classe < أريد جميع >

Classe 2 : le type de réponse sera une date de publication dans le journal officiel d'un article, nous définissons par la suite les patrons :

Pour ce type de question nous avons des patrons spécifiques pour extraire le concept et les propriétés de la question, et aussi des patrons en vue de récupérer la réponse exacte dans les documents candidats.

Le premier sous type consiste à récupérer une date de publication dans un document extrait d'une propriété et classe donnée.

- **L'expansion par des synonymes :**

L'expansion de la requête se fait en utilisant notre ontologie que nous avons défini, le système associer le mot clé de la requête aux concepts de l'ontologie, il va retourner les synonymes reliés à ce dernier en accédant à la propriété « synonymes » dans l'ontologie pour étendre et enrichir la requête.

La figure en dessous démontre le processus d'expansion de la requête.

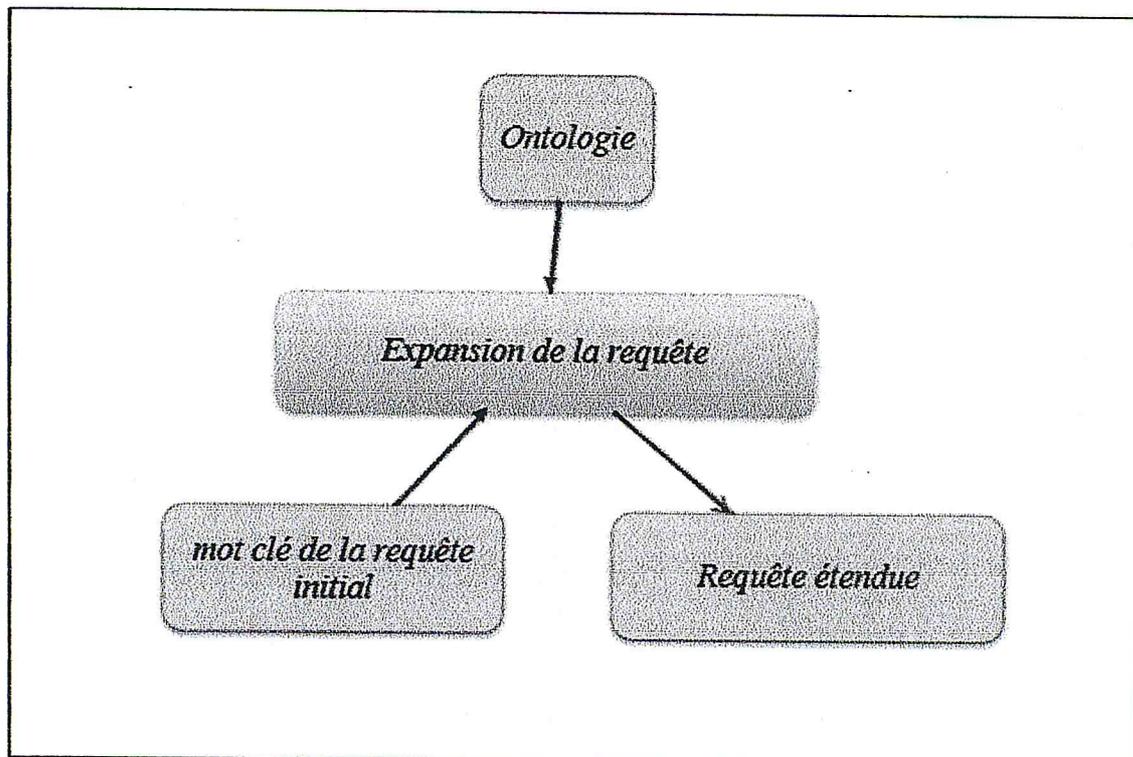


Figure 12 : Processus d'expansion de la requête

Prenons un exemple :

Si la requête initiale contient le mot clé « عقوبة », le système va chercher dans l'ontologie les synonymes de ce dernier et il va les récupérer, ces synonymes sont : « غرامة », « الإعدام », « السجن » etc..., et là il va faire une expansion de requête avec ces mots récupérés.

6. Fonctionnement du système :

Le fonctionnement du système est résumé dans les étapes suivantes :

1. L'utilisateur entre la requête de recherche.
2. Une analyse de la question est effectuée pour récupérer les mots importants en appliquant les patrons.
3. Générer les requêtes de recherche à partir des mots extraits, pour commencer la recherche.
4. Effectuer une recherche dans l'ontologie.
5. Retourner des réponses.

7. L'architecture générale du système :

L'architecture générale de notre système que nous proposons, se compose de trois parties principales :

➤ L'analyse des questions :

Dans un système de question-réponse, le module d'analyse des questions a pour rôle de fournir aux autres modules des informations afin de leur faciliter la sélection des documents et d'extraire la réponse. Donc ce module retourne les termes les plus importants de la question : ceux qui devraient idéalement se trouver dans la réponse.

➤ Processus de traduction de la requête :

La Deuxième étape consistera donc à construire une requête SPARQL, ces requêtes seront utilisées pour interroger l'ontologie « Onto JORADP ».

➤ Extraction les réponses pertinentes :

Enfin, à partir des informations de la question, les réponses possibles sont extraites des documents candidats. Nous allons détailler ces étapes (voir figure).

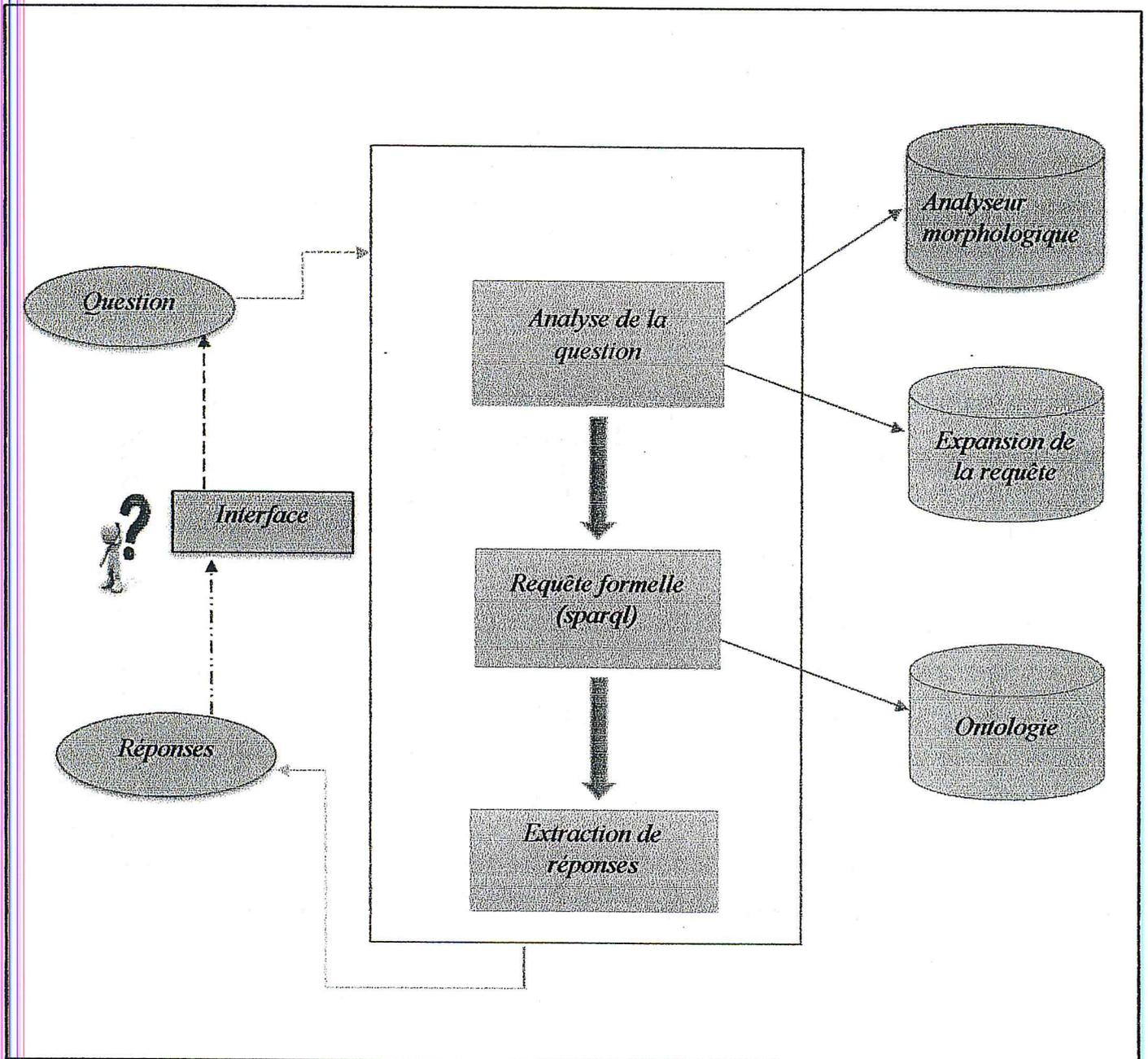


Figure13 : l'architecture générale du système

7.1 Analyse des questions :

La première étape d'un système question réponse consiste en l'analyse des questions afin d'extraire de nombreuses informations utilisées dans les étapes suivantes. Dans tous les systèmes de questions réponses, l'analyse des questions extrait un ensemble de mots importants

de la question. Lors de la recherche de documents, ces mots importants sont fournis afin d'effectuer la recherche

Dans cette partie, la question de l'utilisateur sera soumise à une analyse qui comprend les étapes suivantes :

7.1.1 Normalisation :

C'est une opération qui consiste à :

- Enlever la ponctuation (., ? ... etc.).
- Enlever les signes diacritiques.
- Enlever tout ce qui n'est pas lettre sauf les chiffres qui sont à la fin de la chaîne (qui signifient : année).
- Enlever ال.
- Remplacer آ, ا par ا.
- Remplacer ا par ا.

7.1.2 Analyse morphologique :

L'analyse morphologique est une opération qui consiste à extraire des informations de types morphologiques et/ou syntaxiques du mot à analyser : la racine, la nature morphologique et syntaxique du mot, le temps de conjugaison, le pronom, les suffixes, les préfixes, etc.

De ce fait, nous allons intégrer dans notre système, un analyseur qui a pour but de trouver la forme singulière des mots importants (notamment les noms des classes, par exemple تفرارات, مراسيم).

(L'analyseur que nous avons utilisé est : « Polyglot » qui prend en charge les applications multilingues, ses caractéristiques :

- Détection de la langue (196 langues)
- La reconnaissance des entités nommées (40 langues)
- L'analyse morphologique (135 langues)).

7.1.3 Extraction des termes par les patrons :

Dans cette partie, nous allons appliquer la notion des patrons sur le lemme obtenu de la première analyse pour extraire le nom de la classe et les propriétés sur laquelle la question est posée.

Avant de définir les patrons, nous devons savoir ce que nous allons obtenir comme résultat. En faisant quelques recherches dans le domaine juridique nous avons pu déterminer les questions possibles à poser, donc nous avons recensé toutes les façons possibles dont l'utilisateur puisse formuler une requête en langage naturel dans notre domaine d'application, nous allons donner pour chaque patron sa forme (patron générique) et ses sous-patrons avec les questions qu'il peut générer (voir l'annexe).

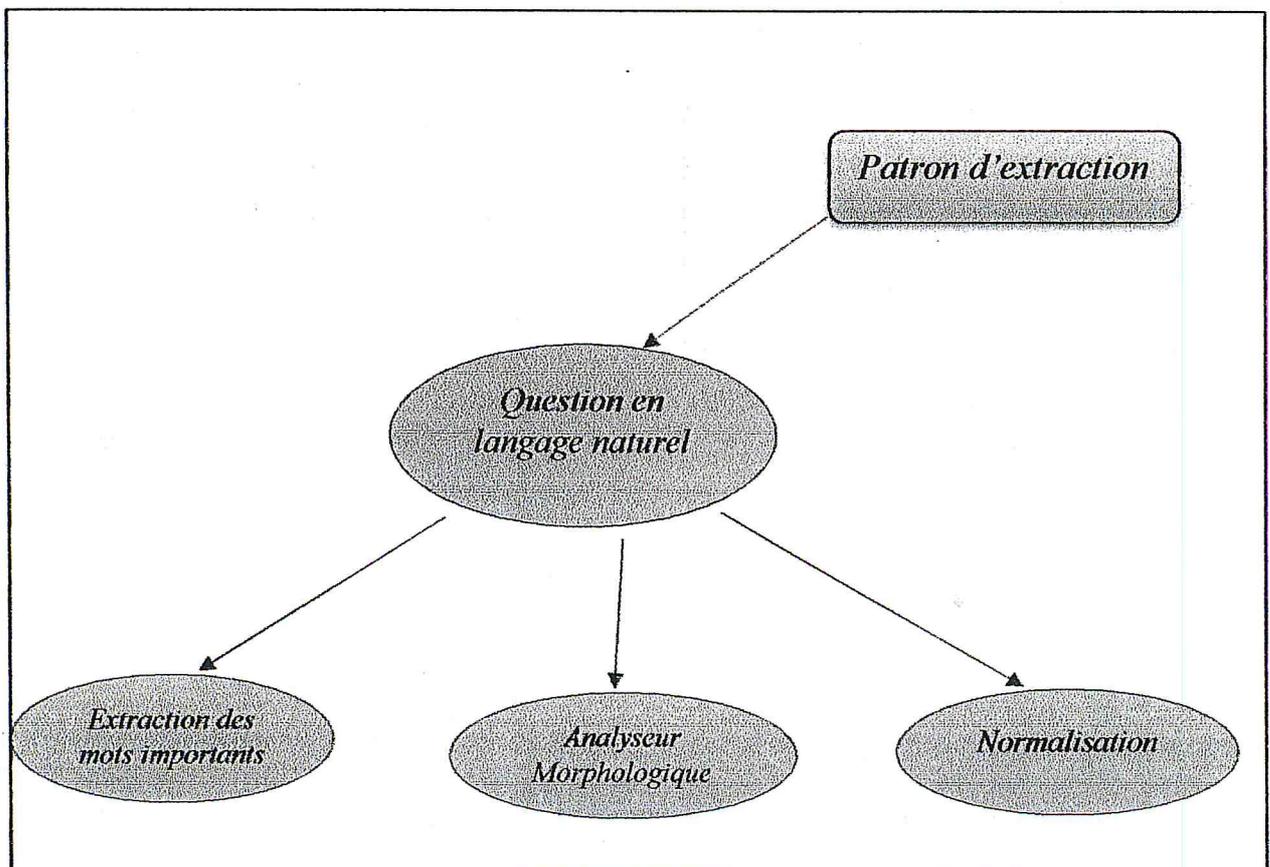


Figure14 : Détail de la phase d'analyse

7.2 Extraction de la réponse :

Une fois la question est analysée et les mots importants extraits en utilisant les patrons de la question, la Deuxième étape consistera donc à construire une requête SPARQL équivalente, représentant l'interprétation de la question avec le vocabulaire de l'ontologie utilisée. Ces requêtes seront utilisées pour interroger l'ontologie « Onto JORADP ».

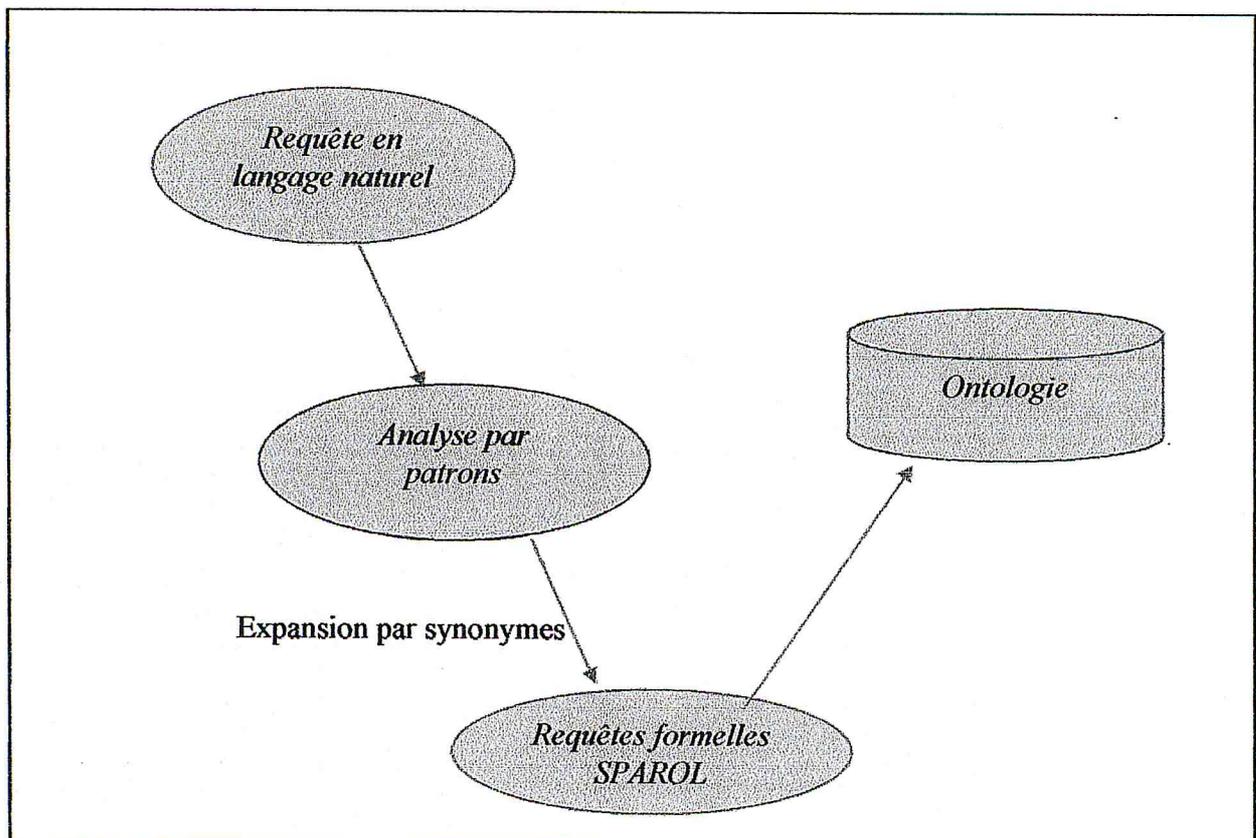


Figure15 : le processus d'interrogation d'ontologie.

7.2.1 Le langage SPARQL :

Comme pour les bases de données relationnelles ou XML, le Web sémantique a besoin d'un langage de requêtes spécifiques. SPARQL permet d'envoyer des requêtes et de recevoir des résultats, par exemple, via HTTP. Le langage SPARQL a été conçu pour interroger des bases de triplets RDF. Il est à RDF ce que SQL est aux bases de données relationnelles. Il utilise d'ailleurs une syntaxe proche de SQL, et notamment la structure générale SELECT WHERE

d'une requête. Le contenu de la clause WHERE est un motif de graphe RDF dont chaque occurrence dans la base de triplets interrogée constitue une réponse à la requête.

SPARQL définit quatre types différents de requêtes : CONSTRUCT, DESCRIBE, ASK et SELECT. CONSTRUCT vise à générer de nouveaux graphes RDF à partir des données RDF disponibles. La forme DESCRIBE est juste informative (elle retourne une sélection aléatoire des réponses). La forme ASK vérifie si un patron de graphe RDF a des correspondances dans les graphes RDF interrogés. La forme SELECT permet de faire correspondre un patron de graphe RDF et de retourner des valeurs de variables correspondant à des nœuds de ce graphe. Dans notre approche de construction de requêtes SPARQL, nous utilisons la forme SELECT afin de représenter les questions posées pour notre système.

Une requête SPARQL a deux composants principaux : un en-tête et un corps. L'en-tête indique la forme de la requête et d'autres déclarations (préfixes, noms des graphes RDF à interroger, variables à retourner pour la forme SELECT).

7.2.2 Traduction de la question analysée en SPARQL :

La forme SELECT des requêtes SPARQL est la plus appropriée pour la représentation des questions de type interrogatif (le type de questions posées pour notre système). L'en-tête de ces requêtes contient principalement le nom de la classe SELECT et les noms de variables à retourner. Le corps des requêtes SELECT contient le nom de la classe et les propriétés extraites de la question originale. Nous formalisons le processus de construction du corps de la requête comme une traduction d'une formule en logique du premier ordre (représentant la sortie des processus d'extraction d'information) en un patron de graphe RDF basique.

Chaque prédicat binaire est transformé en un triplet RDF < sujet, propriété, objet >.

Le sujet et l'objet sont définis comme des variables représentant les arguments du prédicat en logique du premier ordre.

7.2.3 Filtrage selon les propriétés :

Le filtrage d'une requête se fait selon les propriétés de l'ontologie (par exemple `date_de_publication`, `signée_par`).

L'exemple suivant représente une requête SPARQL

```
SELECT ?path WHERE
{
  ?lois_ordinaires :mot_cle ?mot_cle.
  ?lois_ordinaires :ministere ?ministere.
  ?lois_ordinaires :a_un_chemin ?y .
  ?y :chemin ?path .
  FILTER (regex (?mot_cle,'العقوبات' ))
  FILTER (regex (?ministere,'العدل' )) }

```

7.2.4 Le moteur d'inférence :

Un moteur d'inférence, que nous retrouvons également dans la littérature sous le terme de raisonneur, a notamment pour objectif de déduire de nouvelles informations à partir de la base de connaissances ou encore d'interroger l'ontologie. La plupart des moteurs existant actuellement sont conçu pour raisonner à partir des logiques de description et peuvent utiliser en entrée un fichier exprime selon le standard OWL ou RDF(S).

Le moteur SPARQL est disponible dans l'environnement Protège et donc va nous permettre de construire des requêtes sur le graphe de l'ontologie. D'autres moteurs sont disponibles, notamment le moteur ARQL du Framework Jena.

7.2.5 Extraction du texte (ou/et passage) :

La phase d'extraction de texte constitue le dernier traitement dans le processus de l'extraction de la réponse, l'objectif de cette étape est d'afficher la réponse exacte.

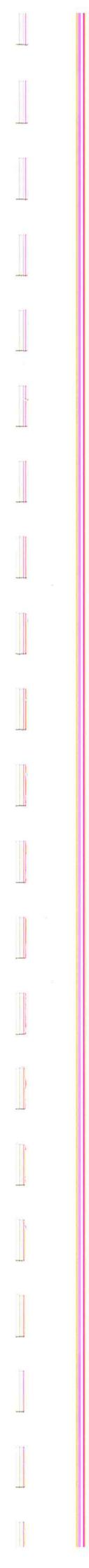
En faisant la recherche dans l'ontologie nous aurons les textes pertinents de la question, là nous avons deux manières d'afficher les réponses :

Si le résultat obtenu est un seul texte alors nous affichons la réponse exacte (passage qui contient la réponse), si c'est plus d'un texte alors nous affichons les textes en soulignant les mots/passages important dans le texte

8. Conclusion :

Ce chapitre est divisé en deux parties, dans la première partie nous avons parlé sur les étapes de la conception de l'ontologie en montrant à la fin l'ontologie résultante dans protégé-2000. La deuxième partie consiste à présenter l'architecture de notre système. Nous avons essayé d'expliquer les différentes étapes, ainsi que l'approche abordée. Tout au long du chapitre nous avons présenté l'utilité des différents modules et leurs rôles dans le système.

Dans le chapitre suivant, nous allons donner quelques exemples en discutant les résultats obtenus tout en essayant de montrer l'efficacité de ce dernier.



Chapitre 04 :
Expérimentation
Et
Évaluation

1. Introduction :

Dans ce travail, nous allons décrire les outils que nous avons utilisés et par la suite faire quelques tests et montrer la différence entre le système existant et le nouveau système, puis nous terminons par une lecture des résultats.

Dans ce chapitre, nous avons testé notre système sur une centaine de questions (saisie manuellement).

Les documents qui existent sur le site du journal officiel la plupart sont en .PDF mais sous format d'images (scannés) et aussi sous format texte mais ils sont intraitables, donc nous avons saisi quelques textes pour faire la démonstration.

Nous n'avons pas pu regrouper tous les synonymes, même tous les textes car pour faire ça, il faut un expert de la langue arabe et un expert du journal officiel.

2. Les outils utilisés :

Lors de la conception de notre système, nous avons essayé de prendre tous les cas possibles de la question posée, tout en proposant des patrons bien définis (Approche à base de patrons).

- Utilisation de l'analyseur morphologique « Polyglot » pour tester si l'utilisateur a écrit en langue arabe ou non et aussi pour avoir la forme singulière de certains mots (nom de la classe).
- Utilisation de la bibliothèque NLTK nous facilite la première tâche qui est l'analyse de la question.
Elle est une bibliothèque de traitement de texte pour la classification, tokenization, résultant, marquage, d'analyse et de raisonnement sémantique.
- Utilisation du langage de requête SPARQL pour l'extraction de la réponse dans l'ontologie. (en comparant avec le graph RDF), (Défini dans le chapitre précédent)
- Utilisation des API (Pypdf2 et python-docx) pour extraire les passages (réponses) et aussi pour souligner les textes.
- Utilisation du Django, un framework de développement web en python.
- Utilisation du python version 2.7 64bits sur windows 8.1 64bits.

3. Expérimentation et évaluation :

Nous avons pris comme corpus de 15 questions pour faire l'expérimentation et l'évaluation de notre système.

Nous utilisons cette formule pour calculer la précision.

$$\text{Precision} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents extraits}}$$

3.1. Les questions utilisées :

- أريد القرار الذي ينص على التراث الثقافي؟
- أريد القرار الذي ينص على التراث الثقافي المبني بالطين؟
- أريد القرار الذي ينص على المركز الجزائري للتراث الثقافي؟
- أريد القرار الذي ينص على إنشاء المركز الجزائري للتراث الثقافي؟
- أريد القانون الذي يتحدث على الرخصة بالنقاط؟
- أريد نص الذي يتحدث على تبييض الأموال؟
- أريد نص الذي يتحدث على تبييض الأموال سنة 2002؟
- أريد القانون الذي يتحدث على العقوبة؟
- أريد القانون الذي يتحدث على الحكم بالإعدام؟
- أريد القانون الذي يتحدث على العقوبات سنة 1996؟
- أريد المرسوم التنفيذي الذي ينص على الإستيرادات/ مبلغ الإستيرادات لوزارة المالية / لوزارة التجارة سنة 1989 ؟
- أريد جميع القوانين الممضاة من طرف الرئيس هواري بومدين؟
- من يكلف بتطبيق المرسوم الذي يتناول اعتماد ميزانية تسيير وزارة الشؤون الخارجية سنة 2016 ؟
- من يكلف بتنفيذ القانون الذي يتحدث على الحكم بالإعدام؟

Pour les mêmes questions nous aurons comme score de précision :

La précision globale obtenue par notre système est de : 0.6

La précision globale obtenu par le JORADP actuel est de : 0.1

4. Comparaison avec le JORADP :

Nous essayons de montrer dans cette partie la différence entre l'utilisation du moteur existant qui est basé sur une recherche classique et avec notre système de question réponse. Notre but est de comparer la recherche avec une requête simple et une requête (+étendu) en utilisant l'ontologie « Onto JORADP ».

Exemple 1 :

Si nous voulons avoir un texte juridique qui parle sur les "sanctions" en 1996 (malgré que le système nous donne pas la liberté de préciser l'année) en utilisant le moteur qui existe sur le site JORADP, nous aurons comme résultat :

Figure 16 : Recherche du mot « العقوبات » avec JORADP.

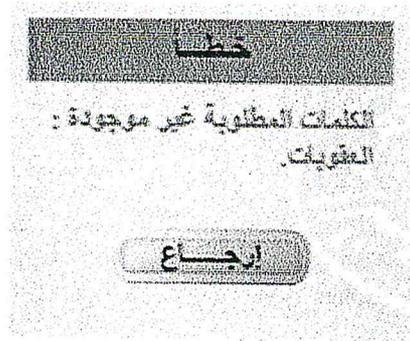


Figure 17 : Résultat du mot « العقوبات » avec JORADP.

Dans ce cas, soit il faut mettre les informations du journal qui contient ce texte (comme le numéro du journal, la date de publication etc ...), soit il faut utiliser le mot clé qui lui est associé (قانون العقوبات / code pénal), avec un non expert ça prendra du temps pour le trouver :

Maintenant, si nous utilisons notre système, l'utilisateur pose la question qui suit :

أريد القانون الذي يتحدث على العقوبات سنة 1996؟

Il peut aussi poser la question comme suit :

أريد القانون الذي يتحدث على عقوبات سنة 1996؟

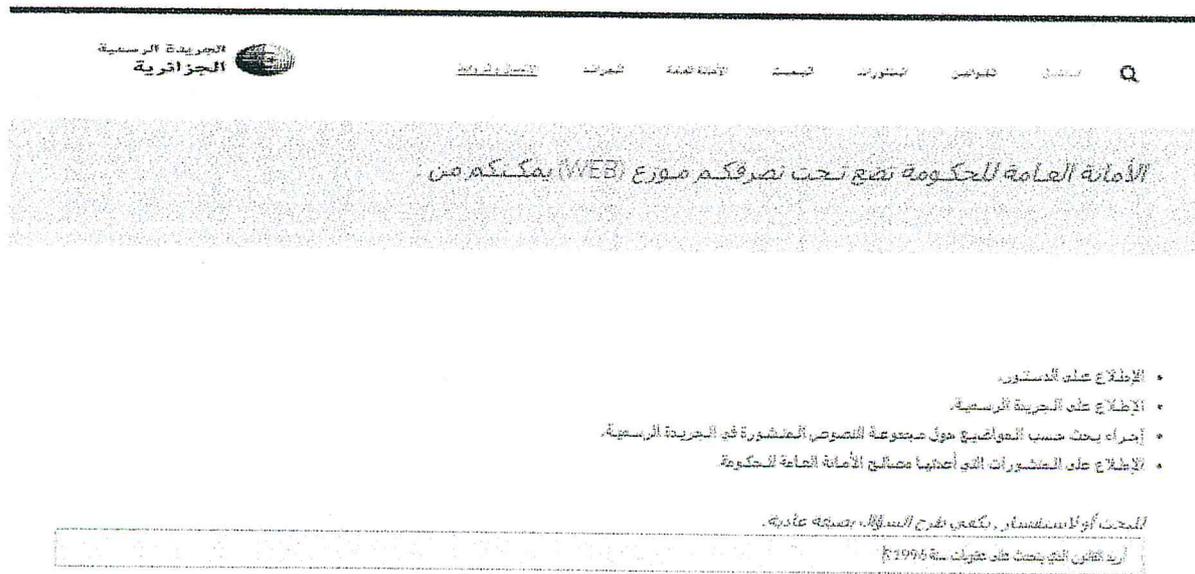


Figure 18 : Recherche avec SQR-Arabe.

Le résultat sera le texte complet, et en plus, un passage justifiant la réponse :

Ce passage est :

يعاقب بالحبس من سنة (6) أشهر إلى ثلاث (3) سنوات و بغرامة من 50.000 دج إلى 200.000 دج كل موظف عمومي في مفهوم المادة 2 من القانون رقم 01-06 المؤرخ في 20 فبراير سنة 2006 و المتعلق بالوقاية من الفساد و كماقحته, تسبب بإهماله الواضح في سرقة أو اختلاس أو تلف أو ضياع أموال عمومية أو خاصة أو أشياء تقوم مقامها أو وثائق أو سندات أو عقود أو أموال منقولة وضعت تحت يده سواء بحكم وظيفته أو بسببها.

يعاقب بغرامة من 100.000 دج إلى 500.000 دج كل من أساء إلى رئيس الجمهورية بعبارات تتضمن إهانة أو سبا أو قذفا سواء كان ذلك عن طريق الكتابة أو الرسم أو التصريح أو بأية آلية لبث الصوت أو الصورة أو بأية وسيلة إلكترونية أو معلوماتية أو إعلامية أخرى.

Remarque :

- Si l'utilisateur veut avoir tous les textes qui parlent sur les sanctions depuis 1962 jusqu'à aujourd'hui, il a qu'à poser cette question : أريد جميع القوانين التي تتناول العقوبات؟ Et il aura la liberté de choisir n'importe quel texte, la réponse sera soulignée ou bien colorée.
- Une simple recherche sur le moteur GOOGLE avec cette question :
Aucune réponse parmi les dix premières ne correspond à ce que nous cherchons.

Exemple 2 :

Si l'utilisateur veut avoir un décret exécutif qui parle sur les importations " مبلغ الإستيرادات/ الإستيرادات" publié par le ministère des finances (وزارة المالية) année 1988 en utilisant le moteur du JORADP :

Résultat :

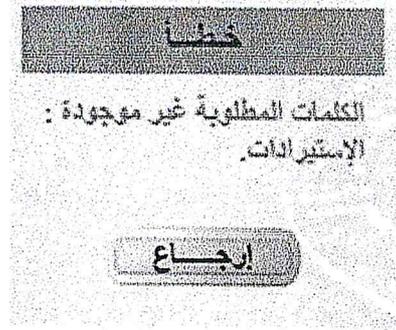


Figure 19 : Résultat du mot « الإستيرادات » avec JORADP.

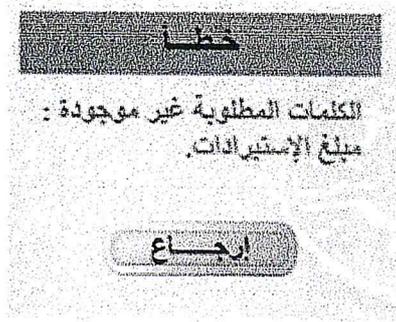


Figure 20 : Résultat du mot «مبلغ الإستيرادات» avec JORADP.

Maintenant, nous utilisons le système de question de réponse :

أريد المرسوم التنفيذي الذي ينص على الإستيرادات/ مبلغ الإستيرادات لوزارة المالية / لوزارة التجارة سنة 1989 ؟

Le résultat sera le texte complet avec une réponse justificante :

يحدد مبلغ الإستيرادات المقررة في البرنامج العام للتجارة الخارجية عن سنة 1989 بثلاثة و خمسين مليار دينار (53.000.000.000 دج).

Exemple 3 :

Si l'utilisateur veut avoir tous leslois signées par le président « هواري بومدين », il ne peut pas exécuter cette requête sur le site du JORADP car il n'a pas la possibilité de le faire, par contre dans notre système il peut poser la question :

أريد جميع القوانين الممضاة من طرف الرئيس هواري بومدين؟

Ainsi nous afficherons toutes lois (organiques et ordinaires) signée par le président Houari Boumedienne.

Exemple 4 :

أريد أمر يتحدث على الثورة الزراعية سنة 1974 ؟

Le mot « الثورة الزراعية » n'existe pas comme mot clé du texte, sur le site du JORADP il affiche qu'il n'ya pas de réponse :

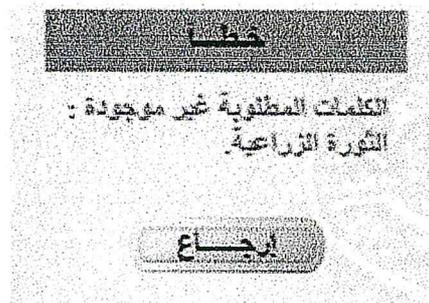


Figure 21 : résultat du mot « الثورة الزراعية » avec JORADP.

Dans notre système, nous lui affichons le texte associé à ce mot clé.

Exemple 5 :

Si l'utilisateur veut avoir un texte qui parle de « القطعة النقدية المعدنية المخصصة للثورة الزراعية سنة 1974 », là il ne pourra pas avoir une réponse pertinente en utilisant le JORADP, car le système ne traite pas une requête de ce genre, c'est à l'utilisateur de faire la recherche parmi les documents retournés pour trouver la réponse.

Par contre, nous pouvons la traiter et extraire une réponse exacte en utilisant notre système qui est :

La réponse sera le texte complet et un passage qui contient la réponse pertinente :

إن الحد الاعلى لاصدار القطعة النقدية المخصصة للثورة الزراعية من فئة 20 سنتيما و المحددة في اول الامر بأربعين ملايين دينار (4.000.000 دج) بموجب الامر رقم 72-25 المؤرخ في 25 ربيع الثاني عام 1392 الموافق 7 يونيو سنة 1972 يرفع الى ستة ملايين دينار (6.000.000 دج).

5. Lecture des résultats :

Dans cette partie, nous avons travaillé sur une quinzaine de questions, avec différentes manières en comparant les résultats en utilisant l'ontologie « Onto Joradp » avec le moteur existant des JORADP.

Si l'utilisateur est un expert du domaine juridique, qui le maîtrise bien (il connaît les propriétés du journal donné) il aura des résultats satisfaisants.

Si c'est un non-expert, là nous avons obtenus des meilleurs résultats avec le système de question réponse qui utilise l'ontologie.

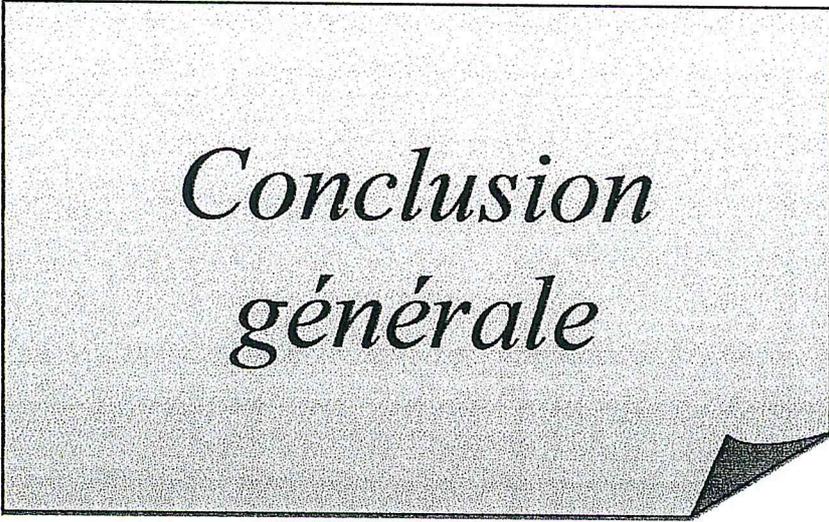
Nous avons remarqué deux choses :

La précision obtenue par le système de question réponse est assez remarquable par rapport à l'utilisation de l'ancien système.

Pour avoir une bonne précision et avoir une réponse pertinente, il faut faire un filtrage avec plusieurs propriétés (comme l'année, ministère etc ...), car la structure du journal officiel est assez compliquée, presque tous les textes sont modifiés, supprimés et complétés chaque année.

6. Conclusion

Dans ce chapitre, nous avons établi la présentation des outils utilisés pour le développement pour réaliser notre travail, nous avons effectué nos tests avec des textes qui existent vraiment sur le journal officiel, nous aurions aimé avoir une collection de textes traitables pour faire d'autres tests et avoir des résultats plus satisfaisants.



*Conclusion
générale*

Conclusion générale

L'objectif de ce travail consiste à concevoir et développer un système de questions/réponses qui vise à retourner une réponse précise à un besoin d'information exprimé en langage naturel pour le journal officiel en langue arabe dans le cadre d'un projet de recherche au sein du Centre de recherche sur l'information scientifique et technique (CERIST).

Cette étude nous a amené au développement d'un système de question-réponse pour le domaine juridique. Le système doit permettre de répondre à des questions sur les bonnes pratiques en utilisant des patrons syntaxiques.

Pour ce faire, plusieurs technologies ont été utilisées. Nous citons les études préalablement faites concernant les systèmes de recherches d'informations notamment les systèmes de questions/réponse, leurs architectures, et les différentes approches utilisées.

Nous avons procédé en premier lieu à une étude du journal officiel et de sa hiérarchie, nous avons conçu et développé une ontologie « Onto JORADP » pour représenter le journal en utilisant la méthode de Stanford.

De cette étude et recherche minutieuses effectuées, nous avons pu concevoir un système de question réponse pour le journal officiel à l'aide d'une approche à base de patrons, qui a un taux de précision assez remarquable par rapport à l'ancien système.

Enfin, de cette étude établie par nos soins, nous avons pu faire ressortir, à ce qui a été imaginé, un système qui résulte sous forme de réalité de notre travail préparatoire. Ainsi, en ce qui est des objectifs fixés au début de l'étude, nous pouvons affirmer qu'ils ont été intégralement atteints avec satisfaction. Toutefois et afin d'offrir plus de fonctionnalités, il serait souhaitable de rendre plus bénéfique ce projet en l'enrichissant par :

- L'utilisation d'une approche pour l'apprentissage des patrons, pour pouvoir répondre à d'autres questions
- Avoir une base de connaissance qui englobe tous les textes depuis 1962 à nos jours.

A la fin, nous sommes satisfaits avec le thème choisi, car la réalisation de ce projet nous a été bénéfique à tous les niveaux et nous a permis de découvrir le domaine TALN (traitement automatique du langage naturel), domaine vaste et en pleine expansion.

Références Bibliographiques :

(Abdelnasse, et.al,2008). Heba Abdelnasser, Maha Ragab, Bassant Farouk, Alaa Mohamed, Reham Mohamed, Nagwa El-Makky, Marwan Torki,(2008), Al-Bayan: An Arabic Question Answering System for the Holy Quran, Computer and Systems Engineering Department Alexandria University, Egypt.

(Ben Abacha,2004). Ben Abacha Asma, (2004), Recherche de réponses précises à des questions médicales : le système de questions-réponses, thèse de doctorat, Université de Paris-Sud.

(Benajiba,2007). Benajiba Y., Rosso P., Lyhyaoui A, (April 3-5 2007), Implementation of the ArabiQA Question Answering System's components. In: Proc. Workshop on Arabic Natural Language Processing,2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morocco.

(Benamara, et.al,2004) Benamara Farah, Patrick Saint Dizier, (2004), Cooperative question answering in restricted domains. The WebCoop experiment. In. ACL-Portability of systems, Barcelona, MIT Press.

(Birini,2009). Brini, W., Ellouze & M., Hadrich, B. L. (2009a). QASAL : Un système de question-réponse dédié pour les questions factuelles en langue Arabe. In 9th Journées Scientifiques des Jeunes Chercheurs en Génie Electrique et Informatique, Tunisie.

(BEKHTI, et.al,2010). S, M AL-HARBI – WSEAS International Conference. Proceedings, (2010). AQuASys. A Question-Answering System for Arabic SMAN BEKHTI, MARYAM AL-HARBI Department of computer science College of computer and information sciences Al-Imam Muhammad ibn Saud University (Traduit de l'Anglais -Français)

(Boucham,2009). BOUCHAM Souhila, (2009) Une approche basée Ontologies pour l'indexation automatique et la Recherche d'Information Multilingue (RIM), thèse de magistère, Université M'hamed BOUGARA de BOUMERDES.

(BOUZOUBAA et al,2010). BOUZOUBAA K., ABOUENOUR L., ROSSO P, (2010) Système de Question/Réponse dans le cadre d'une plateforme intégrée. Cas de l'Arabe, EMI.

(Borst ,1997). Brost-W,(1997), Construction of Engineering Ontologies. Center for Telematica and Information Technology, University of Tweenty, Enschede, NL.

(Chu-Carroll et al. 2002) Chu-Carroll J. & Prager J. & Welty C. & Czuba K. & Ferrucci D. (2002), A Multi-Strategy and Multi-Source approach to question answering. In. Proceedings of the 11th Text Retrieval Conference (TREC-11).

(Dominique, et.al,2005). DGA /CEP, (27 Juin 2006), synapse Développement, Dominique Laurent, Patrique Séguéla, Séminaire, QRISTAL Système de question réponse intégrant un Système de Traitement Automatique des Langues.

(Das, et.al,2002). S. Das, K. Shuster, C. (July 2002). Wu. Agent-based Complex Querying and Information Retrieval Engine, In the Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002), Bologna, Italy.

(Embark,2009). Embark Mehdi,(2009), Un système de question-réponse dans le domaine médical le système Esculape, thèse de doctorat, Université de Paris-Est.

(Foucault,2013). Nicolas Foucault, (2013) question/réponse en domaine ouvert : sélection pertinente de documents en fonction du contexte de la question, thèse de doctorat, université paris-sud.

(Ferret et al. 2000). O. Ferret, B. Grau, G. Illouz, C. Jacquemin, N. Masson,(2000), " QALC, the question answering system of the L&C group at LIMSI ", Q/A track of the Trec8 conference.

(Fluhr,2000). C. Fluhr. (2000), Indexation et recherche d'information textuel, Ingénierie des langues, Hermes.

(F. FÜRST,2004). F. FÜRST, (July 2002) « Opérationnalisation des ontologies : une méthode et un outil ». In Ingénierie des Connaissances (IC), p. 199–210, Lyon, France.

(Grau et al. 2006). Grau B. & Ligozat A. & Robba I. & Vilnat A. & Bagur M. & Séjourné K, (2006) The bilingual system MUSCLEF at QA_CLEF 2006. In: Working Notes, CLEF Crosss-Language Evaluation Forum, Alicante, Espagne.

(GRUBER,1993). T. GRUBER, (1993) « Towards Principles for the Design of Ontologies Used for Knowledge Sharing in Formal Ontology in Conceptual Analysis and Knowledge Representation », Kluwer Academic Publishers.

(Gruber, et.al,1993). Gruber T. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), 199-220.

(Guarino,1997). Guarino N, (1997b). Understanding, building and using ontologies. International J. Human-Computer Studies, 46, 293-310.

(GRUBER T,1993). GRUBER T, (1993) A translation approach to portable ontology specifications, Knowledge Acquisition 5(2), pages 199-220.

(Guarino,1998). Guarino.N, (1998) « Formal Ontology and Information Systems, amended version of a paper appeared » in N. Guarino (ed.), Formal Ontology in Information Systems, Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998, Amsterdam, IOS Press, pp. 3-15.

(GÓMEZ-PÉREZ, et.al,2004). A. GÓMEZ-PÉREZ., M. FERNANDEZ-LOPEZ et O. CORCHO (2004) « Ontological Engineering ». Advanced Information and Knowledge Processing. Madrid, Spain: Springer, 2 editions.

(GUARINO, et.al,1994). N. GUARINO, M. CARRARA et P. GIARETTA (1994) « Formalizing ontological commitments». In National Conference of the American Association on Artificial Intelligence (AAAI), p. 560–567.

(Gilles,2010). Gilles Kassel, (2010), Université de Picardie Jules Verne Association Française d'Intelligence Artificielle chez AFIA

(Hejazi et al. 2004). M.R. Hejazi, M.S.Mirian, K. Neshatian, A.Jalali, and B.R. Ofoghi,(2004), I.T. Department, Iran Telecom Research Center, Tehran Iran (Traduit Anglais-Français).

(Ihadjadene, 1999). Ihadjadene, M., (1999), La Recherche et la navigation dans un SRI grand public, thèse de doctorat en science de l'information, Université Lyon 1.

(Jian-Yun,2009). Jian-Yun Nie,(2009) cours Recherche d'information MASTER2, Université de Montréal <http://www.iro.umontreal.ca/>.

(Kaveh,2004). Kaveh BAZARGAN, (Juin 2004) « Le rôle des ontologies de domaine dans la conception des interfaces de navigation pour des collections en ligne de musées : évaluations et proposition » Mémoire de DEA en Management et Technologies des Systèmes d'Information (MATIS) Université de Genève Suisse.

(Mohammed et al. 1993). Mohammed F.A., Nasser K., Harb H.M, (1993) A Knowledge Based Arabic Question Answering System (AQAS). ACM SIGART Bull. 4(4), 21–30.

(Moldovan,et.al,2002) Moldovan D. & Harabagiu S. & Girju R. & Morarescu P. Lacatusu F. & Novischi A.Badulescu A. & Bolohan O(2002), LCC tools for question answering. In. Proceedings of the 11th Text Retrieval Conference (TREC-11).

(Mathieu-Henri,2014). Mathieu-Henri Falco, (Mai 2014), Répondre à des questions à réponses multiples sur le Web. THÈSE DE DOCTORAT, UNIVERSITÉ PARIS-SUD.

(Natalya,2010). Natalya F. Noy and Deborah L, (2010), McGuinness « A Guide to Creating Your First Ontology» Stanford University, Stanford, CA.

(Niu,et.al ,2004) Niu Y., Hirst G.,(2004), « Analysis of semantic classes in medical text for question answering »,Pro-ceedings of the ACL-2004 Workshop Question Answering in Restricted Domains.

(Neeches, et.al, 1993). Neeches, Finin T, Fikes R.E, Gruber T.R, Senator T et Swartou W.R. (1993). « Enabling technology for knowledge sharing » AI Magazine. Vol.12, no 3.

(Nyberg,et.al,2002). Nyberg E. & Mitamura T. & Carbonell J. & Callan J. & Collins-Thompson K. & CzubaK. & Duggan M. & Hiyakumoto L. & Hu N. & Huang Y. & Ko J. & Lita L. & Murtagh S. & Pedro V & Svoboda D. (2002) The Javelin question answering system at TREC 2002. In. Proceedings of the 11th Text Retrieval Conference (TREC-11).

(Roussey,2001). Roussey Catherine, (décembre 2001), Thèse pour obtenir le grade de docteur. Une méthode d'indexation sémantique adapté aux corpus multilingues.

(Strzalkowski,1993). T. Strzalkowski. Natural language processing in large scale text Retrieval tasks. In text Retrieval Conference (TREC-1), page 137, 1993.

(Smeaton,1989). A. F. Smeaton. (1998) Information retrieval and natural language processing. In proceedings of a conference jointly sponsored by ASLIB, University of York, page 2, Mars.

(SM,1983). G. Salton and McGill. (1983). Introduction to Modern Information Retrieval. Mc Graw-Hill, New York.

(Stéphane, et.al, 2004). Stéphane Chaudiron, Madjid Ihadjadene, « Évaluer les systèmes de recherche d'information. Nouveaux modèles de l'utilisateur », Hermès, La Revue 2004/2 (n°39), p.170-178. <http://www.cairn.info/revue-hermes-la-revue-2004-2-page-170.htm>

(Spink, 2001). Spink, A., Wolfram, D., Jansen, B. J., Saracevic, T,(2001) « Searching the Web : the Public and their Queries », in Journal of the American Society for Information Science, vol. 53, n° 2 , p. 226-234.

(Uschold,et.al,1996). Uschold et M. Grüninger, (1996) « ONTOLOGIES: Principles, Methods and applications» Knowledge engineering review, vol.11, N.2.

(Uschold,et.al,1996). Uschold & M.Grüninger, (1996) « ONTOLOGIES: Principles, Methods and Applications ». Knowledge Engineering Review.

(WEL, 1998) Welty, C. (1998) The Ontological Nature of Subject Taxonomies. In Proceedings of 1st International Conference on Formal Ontologies in Information Systems, FOIS'98, Trento, Italy, p. 317-327.