وزارة التعليم العالي والبحث العلمي


**Badji Mokhtar Annaba University**
**Université Badji Mokhtar- Annaba-**

**جامعة باجي مختار**
**- عنابة -**

**Faculté des Sciences de l'Ingénieur**

**Département d'Informatique**


**MEMOIRE**
Présenté en vue de l'obtention du diplôme de **MAGISTER**
**Année : 2006**

# Expansion de requête à l'aide d'une ontologie Arabe dans le domaine juridique

**Option :**
Intelligence Artificielle Distribuée (IAD)


**Par**
**Soraya Zaidi**


.

| | | | |
|---|---|---|---|
| **DIRECTEUR DE MEMOIRE** | **M.T. Laskri** | **Professeur** | **U. ANNABA** |
| | **Devant le jury** | | |
| | **Dr. Abderrezek Henni** | **MC** | |
| **PRESIDENT** : | **(**Directeur général de la modernisation de la justice**)** | | **INI. ALGER** |
| | **Dr. Nabila Nouaouria** | **MC** | **U. ANNABA** |
| **EXAMINATEURS** : | **Dr. Nadir Farah** | **MC** | **U. ANNABA** |

Sommaire

Résumé

Abstract

ملخص

Table des figures

Liste des tableaux

Abréviations et Acronymes

Introduction Générale ..... 12

## Chapitre 1: La recherche d'information sur le Web

1. La recherche d'informations sur le web ..... 14

1.1. Historique ..... 14

1.2. Les outils de recherche d'information ..... 14

1.2.1. Les moteurs de recherche ..... 14
1.2.2. Les méta moteurs ..... 16
1.2.3. Les annuaires ..... 17

1.3. Problématique de la recherche d'information sur le Web ..... 17

1.3.1. Introduction ..... 17
1.3.2. La recherche géographique ..... 18
1.3.3. La recherche thématique ..... 18
1.3.4. La recherche par mot-clés ..... 18

1.4. Les types de recherche d'information ..... 19

1.4.1. La recherche Ad Hoc ..... 19
1.4.2. La recherche multimédia ..... 20

1.5. Les techniques de recherche d'informations ..... 21

1.5.1 La catégorisation de documents ..... 21
1.5.2 Le chemin de lecture ..... 21

1.5.3. La requête par l'exemple ..... 22
1.5.4. Le résumé automatique ..... 22
1.5.5. La recherche par question-réponse ..... 22

1.5.6. Recherche par traitement du langage naturel ..... 24
1.5.7. Les méta données et le Dublin core ..... 24
1.5.8. L'expansion de la requête ..... 26

1.6. Le problème d'évaluation ..... 31

1.7. Les systèmes de recherche d'information en Arabe ..... 33

1.7.1. Les systèmes arabisés ..... 33
1.7.2. Les systèmes Arabes ..... 34

1.8. Quelques systèmes arabes de recherche d'information ..... 34

1.9. Conclusion ..... 38

## Chapitre 2: Les ontologies

2. Les ontologies ..... 39

2.1. Définition d'une ontologie ..... 39

2.2. Les méthodes de construction ..... 41

2.3. Les difficultés de construction d'une ontologie ..... 41

2.4. L'utilité des ontologies ..... 42

2.5. Outils de développement d'ontologies ..... 42

2.6. Wordnet et Eurowordnet ..... 44

2.6.1. Généralités ..... 44
2.6.2. Les relations dans Wordnet et EuroWordnet ..... 46

2.7. L'ontologie SENSUS ..... 49

2.8. Conclusion ..... 49

## Chapitre3: Le système proposé

3. Le système proposé ..... 50

3.1. Introduction ..... 50

3.2. La solution proposée ..... 51

3.2.1 Le choix du domaine ..... 51
3.2.2 La construction de l'ontologie ..... 52
3.2.3 Cohérence de l'ontologie ..... 59

.3.3 La construction de l'ontologie ..... 61

3.3.1. Le choix de l'outil utilisé ..... 61
3.3.2. La méthode de construction ..... 62
3.3.3. La base de l'ontologie ..... 62
3.3.4. La stratégie de construction ..... 62

3.4. Aspect général du système ..... 63

3.4.1. L'interface utilisateur ..... 64
3.4.2. L'analyseur de Requête ..... 65
3.4.3. L'expansion de la requête ..... 67
3.4.4. Traducteur de requête ..... 70
3.4.5. Processus de recherche ..... 71
3.4.6. L'affichage des résultats ..... 71
3.4.7. Conclusion ..... 74

## Chapitre4: Résultats et discussion

4. Résultats et discussion ..... 75

4.1. Introduction ..... 75

4.2. Les outils de recherche utilisés ..... 75

4.3. Les mesures ..... 75

4.4. Traitement de quelques exemples ..... 75
4.5. Analyse et discussion ..... 87
4.6. Conclusion ..... 89

## Conclusion et perspectives

5. Conclusion et Perspectives ..... 90

5.1. Conclusion ..... 90

5.2. Perspectives ..... 91

## Références et bibliographie

Références et bibliographie ..... 92

## Annexes

Annexes ..... 100

# RESUME

Notre mémoire traite la problématique de la recherche d'information sur le Web, celle en Arabe en particulier plus spécialement dans le domaine juridique.

Les moteurs de recherche actuels se basent, dans leur recherche, sur des mots clés, ces mots sont traités syntaxiquement, la sémantique est omise d'où un grand nombre de documents retournés non pertinents ce qu'on appelle *bruit.*

Différentes techniques ont été proposées pour réduire le taux de ce bruit et améliorer la précision de la recherche.
Nous avons présenté un grand nombre de ces techniques et nous avons opté pour celle que nous avons jugée intéressante pour notre cas, en l'occurrence l'expansion de la requête à l'aide d'une ontologie.
La raison de notre choix est que d'un coté les ontologies représentent un avenir prometteur pour le Web sémantique, qui est à son tour, l'héritier supposé du Web actuel, d'un autre coté les ontologies n'ont pas encore été expérimentées dans la recherche d'information en langue Arabe, dans le domaine juridique (du moins jusqu'à la rédaction de ce travail).

Nous traitons donc, dans ce travail le problème de la construction d'une ontologie en langue Arabe dans le domaine juridique, dans le but d'étendre une requête Arabe et la soumettre à un processus de recherche pour réduire le bruit et améliorer la précision.

Mondialisation oblige, nous tentons de permettre à un utilisateur potentiel d'effectuer une recherche avec une requête Arabe pour l'obtention de documents Anglais ou Français, ceci en soumettant sa requête à un système de traduction automatique libre d'utilisation sur le Web, tel que *Tarjim*, ainsi que l'expansion de la requête traduite à l'aide de la base lexicale qui représente une ontologie générique.
Nous décrivons les différentes étapes de la construction manuelle de notre ontologie relative au domaine juridique, à l'aide de l'éditeur d'ontologies Protégé-2000 et ce à partir d'articles publiés sur le Web concernant spécialement le domaine cité.

Nous avons essayé d'établir la hiérarchie avec la méthode descendante, nous essayons de donner à chaque concept sa définition et nous comptons lui attacher ses variantes les plus représentatives du domaine juridique.

Nous traitons à la fin quelques exemples illustratifs concernant l'expansion de la requête, pour essayer de donner une évaluation préliminaire de l'approche adoptée.

# ABSTRACT

Our work treats the problems of the information retrieval on the Web, in particular, in Arabic more especially in the legal domain.

The current search engines are based on key words, these words are treated syntactically, semantics is omitted from where a great number of documents returned, are non relevant this was called *noise*.

Various techniques were proposed to reduce the rate of this *noise* and to improve the precision of research. We presented a great number of these techniques and we chose that which we considered to be interesting for our case, the query expansion using an ontology in the legal domain.

The reason of our choice is that the ontologies represent a promising future for the semantic Web, which is in its turn, the heir supposed to the current Web, in the other hand ontologies were not tested yet in the Arabic information Retrieval on the legal domain (at least until the printing of this work).

We thus treat, in this work the problem of the manual construction of an Arabic ontology in the legal domain, in an aim of expansion an Arabic query to submitting it to a research processing in ordre to reduce the *noise* and to improve the *precision*.

Universalization obliges, we try to make it possible to a potential user to carry out a search with an Arabic query for obtaining English or French documents, this by translating its query by a machine translation system free of use on the Web, such as Tarjim, as well as the expansin of the translated query using Wordnet the lexical base which represents a generic ontology.

We describe the various steps of the manual construction of our ontology relating to the legal domain, using the editor of ontologies Protege-2000 and this based on published articles on the Web relating to the legal field. We tried to construct the hierarchy with the Top-down method, we try to give to each concept its definition and attach to it the derivatives, wich are the most representative of the legal domain.

We discuss in the end some illustrative examples concerning the query expansion to try to give a preliminary evaluation of the adopted approach.

# ملخص

يعالج بحثنا هذا إشكالية البحث و استرجاع المعلومات المخزنة على الواب، و على وجه الخصوص البحث بالغة العربية و بصفة خاصة التي تتعلق بميدان القانون.

محركات البحث الحالية تعتمد على كلمات مفاتيح بدون الأخذ بعين الاعتبار معاني هذه الكلمات، و هذا ما ينجر عنه استرجاع عدد هائل من الوثائق ليست لها علاقة البتة بموضوع البحث و هذا ما يسمى بـ "*الضجيج*" .

لقد حاول الكثير من الباحثين استعمال تقنيات مختلفة للحد من شدة هذا *الضجيج* و تحسين دقة البحث.

حاولنا تقديم عدد كبير من أمثال هذه التقنيات و وقع اختيارنا على الطريقة التي ارتأينا أنها الأمثل و هي تمديد مسالة البحث باستعمال انطولوجيا باللغة العربية في ميدان القانون الجزائري.

و قد تم اختيارنا لهذه التقنية لسببين، أما الأول فلكون الانطولوجيات تشكل مستقبلا واعدا للواب المدلولي الذي يمثل بدوره الوريث الشرعي للواب الحالي، و أما السبب الثاني فهو أنه لحد الآن و حسب علمنا فإنه لم تستعمل بعد الانطولوجيات باللغة العربية في البحث و استرجاع المعلومات على الواب و خاصة في القانون.

نتكلم إذا في هذا البحث، عن كيفية إنشاء انطولوجيا باللغة العربية في القانون، بهدف تمديد مسالة طلب استرجاع معلومات و إدخالها في معالج بحث و ذلك لأجل تخفيض *الضجيج* و تحسين مرد ودية البحث من حيث الدقة.

و لأن العولمة فرضت علينا نفسها، فإننا نحاول في عملنا هذا تمكين المستعمل من ترجمة طلبه إلى الانجليزية أو الفرنسية ترجمة آلية باستعمال النظام "ترجم" مثلا على النت، ثم تمديد الطلب بواسطة "وردنت" الانطولوجيا العامة التي يمكن استعمالها على الواب.

نشرح المراحل المختلفة لإنشاء الانطولوجيا المتخصصة في القانون الجزائري بواسطة " بروتجي 2000 " و ذلك اعتمادا على مواضيع مختلفة نشرت على الواب في الميدان المذكور.

حاولنا وضع هيكلة الانطولوجيا باستعمال الطريقة التنازلية، حاولنا ربط كل مفهوم بتعريفه من جهة و من جهة أخرى بمشتقاته المرتبطة حصريا بالميدان القانوني.

نحاول في الأخير مناقشة بعض الأمثلة المبينة لفائدة الامتداد المبني على استعمال الانطولوجيا لتقييم الطريقة المستعملة في هذا البحث.

# Table des figures

Figure 1 : Mode de fonctionnement d'un moteur de recherche ..... 15
Figure 2 : Mode de fonctionnement d'un méta moteur ..... 16
Figure 3 : Mode de fonctionnement d'un annuaire ..... 17
Figure 4: Le modele de recherche Ad Hoc ..... 20
Figure 5: Les questions-types et les réponses-types ..... 23
Figure 6: Table de mesure de pertinence d'un système de RI ..... 32
Figure 7: Paradigme d'évaluation de la recherche dans TREC ..... 33
Figure 8: structure du mot " المدرسون " (les enseignants) ..... 35
Figure 9: segmentation du mot " دراسة" (étude) ..... 36
Figure 10: Exemple de liens sémantiques ..... 37
Figure 11 : L'ontologie en tant que spécification d'une conceptualisation ..... 39
Figure 12: L'ontologie au coeur des autres disciplines ..... 40
Figure 13: quelques outils d'édition d'ontologies ..... 43
Figure 14: Architecture générale de l'ontologie EuroWordnet ..... 46
Figure 15: Les différents types de Méronymie ..... 47
Figure 16: Les types de Méronymie dans EuroWordnet ..... 48
Figure 17: les "top-types" de l'ontologie ..... 55
Figure 18: la relation d'hyponymie dans l'ontologie ..... 55
Figure 19: les propriétés des classes. ..... 56
Figure 20: Les facettes de l'attribut "الغرامة_المالية " ..... 57
Figure 21: Les facettes de l'attribut " "عقوبة_الحبس ..... 58
Figure 22: Les instances de l'ontologie ..... 59
Figure 23: Architecture générale du système ..... 66
Figure 24: le concept " جنحة" dans la hierarchie ..... 67
Figure 25: les slots herités et les slots propres au concept " جنحة" ..... 68
Figure 26: Le domaine et les facettes du concept "الغرامة" ..... 69
Figure 27: les differents champs du concept " الحقوق_العينية" ..... 72
Figure 28: Les differents champs du concept" المحكمة_العسكرية" ..... 72
Figure 29: La description de la hierarchie ..... 73
Figure 30: La presentation des concepts sous format RDF. ..... 73
Figure 31: les hyperonymes du concepts "مجلس الأمة" ..... 78
Figure 32: hyperonymes et hyponymes du concept "الحقوق العينية" ..... 80
Figure 33: " الخلع" dans la hierarchie des classes ..... 81
Figure 34: Traduction de "الخلع" avec Tarjim. ..... 82
Figure 35: Recherche de *"Removal"* dans Wordnet ..... 83
Figure 36: recherche de "*penal code*" avec Wordnet ..... 84
Figure 37: traduction de la requete avec Tarjim de Ajeeb ..... 85
Figure 38: definition du mot *children* par wordnet ..... 86
Figure 39: définition du mot "Right" par Wordnet ..... 86

# Liste des Tableaux

Tableau 1: Recherche simple de la requete"قانون العقوبات" avec Google....76
Tableau 2: Recherche étendue de "قانون العقوبات"....76
Tableau 3: Recherche simple de "مجلس الامة" ....77
Tableau 4: Recherche étendue de "مجلس الامة" ....77
Tableau 5: Recherche simple de "شروط استخدام الاجانب" ....78
Tableau 6: Recherche étendue de "شروط استخدام الاجانب"....79
Tableau 7: Recherche simple de "الحقوق العينية" ....79
Tableau 8:Recherche étendue de "الحقوق العينية":....79
Tableau 9: Recherche simple de "حقوق الطفل في الجزائر"....80
Tableau 10: Recherche étendue de "حقوق الطفل في الجزائر" ....80
Tableau 11: Recherche simple de "الخلع"....81
Tableau 12: Recherhce étendue de "الخلع"....81
Tableau 13: Recherche étendue de "*removal*" ....83
Tableau 14: Recherche étendue avec Wordnet de "*penal code"* ....84
Tableau 15: Recherche étendue de "*Children Rights*"....87
Tableau 16:Tableau récapitulatif de la précision moyenne ....87
Tableau 17: comparaison de la précision moyenne « avec » et « sans » synonymes. ....87

# Abréviations et Acronymes

**AFTDB** Arabic Full Text Data Base
**AIRSMA** Arabic Information Retrieval System based on Morphological Analysis**.**
**ASWS** Alaihi Salat Wa Salam
**ECIR** European Conference of IR
**CIRCA** Conceptual Information Retrieval and Communication Architecture
**CLAF** Cross-language Arabic forum
**CLEF** Cross Language Evaluation Forum
**CLIR** Cross Language Information Retrieval
**DAML** Darpa Agent Markup Language
**ILI** Inter Lingual Index
**IR** Information Retrieval
**IRS** Information Retrieval System
**IRSAD** Information Retrieval System for Arabic Documents
**JDK** Java Development Kit
**JESS** Java Expert System Shell
**KES** Knowledge Extraction Structure
**KRS** Knowledge Representation System
**NIST** National Institute of Standards and Technology
**OIL** Ontology Inference Layer
**OKBC** Open Knowledge Base Connectivity
**OMC** Organisation Mondiale du Commerce
**OWL** Ontology Web Language
**PRP** Probabilistic Ranking Principle
**RDF** Ressource Description Framework
**RI** Recherche d'Information
**SMI** Stanford Massachusetts Institute
**SRI** Système de Recherche d'Information
**TREC** Text Retrieval Evaluation Conference
**UMLS** Unified Medical Language System

# Introduction Générale

Durant des milliers d'années, les gens ont réalisé l'importance d'archiver l'information et de la retrouver. Avec l'avènement de l'ordinateur, il est devenu possible de stocker une grande quantité de données. Retrouver une information utile parmi la collection, est devenu nécessaire. Le domaine de la recherche d'information (RI) est né en 1950, suite à cette nécessité.

L'accroissement explosif des connaissances dans tous les domaines notamment le domaine juridique et l'inflation textuelle et multilingue, notamment sur le Web, confèrent à l'accès, à l'exploitation ou à la traduction de ces informations un enjeu important.

Le web est l'application la plus connue de l'Internet. Elle permet d'accéder à l'information en provenance de différents sites à travers le monde. En ce sens, Michard affirme que le web apporte une solution générale aux besoins d'accès à l'information à distance **[Mic99].**

La recherche d'informations sur le web se fait à l'aide d'outils de recherche automatiques notamment les moteurs de recherche (Altavista, Google, etc.). Ce sont des logiciels puissants permettant de parcourir tout le web à la recherche de nouveaux sites pour les indexer et les intégrer dans leurs bases de données. Lorsque l'internaute formule sa requête via l'interface d'interrogation du moteur de recherche, ce dernier procède à la recherche dans les sites référencés dans sa base, pour fournir en sortie les documents en rapport avec la question posée **[Oue03].**

Du point de vue documentaire, la performance de ces outils de recherche est bien inférieure à leur puissance informatique, dans la mesure où les résultats d'une requête de l'utilisateur pourraient engendrer du bruit (documents non pertinents retournés), ou bien du silence (documents pertinents non retournés). Ce phénomène est dû principalement à la pratique d'indexation réalisée par ces outils, qui est considérée comme une « indexation plein texte en aveugle » au sens de Michard **[Mic99].**

Plusieurs techniques ont été utilisées pour pallier à ces inconvénients, telles que la catégorisation de documents, le résumé automatique, le chemin de lecture et l'expansion de la requête pour ne citer que celles là.

L'expansion de la requête consiste à rajouter de nouveaux termes aux mots de la requête initiale pour élargir le champ de la recherche. Cette expansion peut se faire de différentes façons, nous pouvons citer le feedback pertinent, les relations qualia, les thesaurus et les ontologies.

Une ontologie est l'ensemble du vocabulaire d'un domaine donné, disposé sous forme d'une hiérarchie avec toutes les relations qui existent entre les termes de ce vocabulaire.

Construire une ontologie est une opération très délicate, dans le sens où nous ne pouvons avoir le même point de vue concernant le sens exact d'un terme ou de la relation le reliant à un autre terme. Cela suppose un travail considérable entre linguistes, experts du domaine et informaticiens. Cela coûte beaucoup d'argent et nécessite énormément de temps. Pour cette raison la majorité des travaux se sont focalisés sur des constructions automatiques ou semi-automatiques, leur nombre est vertigineux mais les ontologies les plus réussies restent celles qui ont commencé par une construction manuelle même si par la suite ils ont automatisé leurs méthodes pour les enrichir et les mettre à jour.

Quand nous parlons d'ontologie en droit, nous pensons directement à toutes les sources du droit, nous pouvons l'étendre à tout le processus par exemple : du rapport de police à la condamnation par un tribunal et le règlement du litige (prison, amende..) mais le nombre de concepts est interminable et plusieurs groupes de recherche en informatique juridique s'intéressent à ce domaine dont on fait souvent référence par les termes « eLaw » ou « « eJustice ». D'un autre coté, il est plutôt difficile d'imaginer une Ontologie universelle puisque le droit est a priori culturel, il peut même découler directement des textes religieux comme c'est le cas par exemple, en Arabie Saoudite.

Cependant il est très utile de conceptualiser le droit avec une ontologie du domaine, puisque l'adage que les juristes se plaisent tant à donner à ceux qui se trouvent dans le tort est : *nul n'est censé ignorer la loi* encore faut-il comprendre un peu comment elle fonctionne avant de se plonger dans le langage « obscure » des juristes. Beaucoup de ressources Web pourraient bénéficier d'un tel système. Imaginons si on pouvait comprendre les rouages des bureaucraties par de telles ontologies, parce qu'elles expliquent, montrent, comment fonctionne l'outil social qu'est le système de justice et que chaque culture pourrait tenter d'aligner leur ontologie respective et ainsi augmenter le quotient intellectuel des sociétés du monde. Nous savons tous que l'indépendance de la Justice n'a de sens que si le citoyen peut y recourir librement, pour défendre ses droits fondamentaux. Pour cela il doit être facile pour lui de consulter les différents articles de lois le concernant, le meilleur moyen de nos jours est l'Internet, cela suppose de mettre à la disposition de tout citoyen une telle ressource qu'est l'ontologie juridique.

Outre cela, le système judiciaire est le seul domaine qui reste arabisé en Algérie, en effet toute loi, tout jugement arrêté ou document, sont émis en langue Arabe. Une grande communauté travaille dans ce domaine et a besoin d'un outil comme une ontologie juridique d'un coté pour expliquer les engrenages de la loi et d'un autre coté, pour faciliter la recherche sur le Web, dans le domaine spécifié.

Pour répondre à cette dernière attente, nous avons essayé d'entamer le défrichage de ce terrain périlleux aux multiples rouages, tout en espérons que d'autres travaux suivront pour mettre en place une vraie ressource pour l'aide à la compréhension du droit.

# 1. La recherche d'informations sur le web

## 1.1. Historique

Durant des milliers d'années, les gens ont réalisé l'importance d'archiver l'information et de la retrouver, avec l'avènement de l'ordinateur, il est devenu possible de stocker une grande quantité de données, retrouver une information utile parmi la collection, est devenu nécessaire. Le domaine de la recherche d'information est né en 1950, suite à cette nécessité.

Déjà en l'an 3000 avant Jésus Christ, les sumériens ont essayé de réserver des zones spéciales pour stocker des tablettes d'argile qui portaient des inscriptions cunéiformes, de plus ils les ont organisés de telle sorte que l'accès à ces archives soit le plus optimal et le plus simple possible pour une utilisation efficace.

Le besoin de stocker des informations et de les retrouver par la suite est devenu de plus en plus important au fil des siècles, notamment avec l'invention de l'imprimerie et du papier. Juste après que l'ordinateur soit inventé, les gens se sont vite rendus compte qu'ils pouvaient l'utiliser pour stocker et retrouver automatiquement de grandes quantités d'information. En 1945, Vannevar Bush publie un article intitulé « As We May Think » qui donne naissance à l'idée d'accéder automatiquement à d'énormes quantités de connaissances stockées **[Van45].**

En 1950, l'idée a été concrétisée avec la description de la façon dont l'information doit être stockée puis retrouvée automatiquement. Beaucoup de travaux ont vu le jour, concernant le stockage et la recherche de l'information d'une façon automatique, le plus important est sans doute celui de H. P. Luhn en 1957, dans lequel il propose des mots comme unités pour indexer des documents. Plus tard en 1960 et à l'université de Harvard, Gerard Salton et ses étudiants développèrent SMART **[Sal71].**

Le système Smart a permis aux chercheurs d'expérimenter leurs idées dans le but d'améliorer la qualité de la recherche d'information.

Les années 70 et 80 ont vu le développement de modèles variés pour la RI, ces derniers ont prouvé leur efficacité sur des collections de texte de petite taille (plusieurs milliers d'articles) **[Sin01]** cependant en l'absence de large corpus, la question si ces modèles restent efficaces demeure sans réponse !

Ceci a changé en 1992 avec l'organisation de TREC (Text Retrieval Conference**)[Har93].**

TREC est une série de conférences d'évaluation, sponsorisé par des agences du gouvernement Américains sous le nom de NIST, dans le but d'encourager la recherche d'information sur de larges collections de texte, Les algorithmes développés en RI furent les premiers employés dans la recherche sur le Web entre 1996 et 1998. La plus part des systèmes de recherche d'information assignent un score à chaque document et le range selon ce score.

## 1.2. Les outils de recherche d'information

La recherche d'information consiste à retrouver une page pertinente en réponse à une requête de l'utilisateur. En se basant sur le mode de fonctionnement on peut distinguer trois catégories d'outils de recherche d'information : Les moteurs, les méta moteurs et les annuaires.

### 1.2.1. Les moteurs de recherche

Il existe à l'heure actuelle plusieurs milliers de moteurs de recherches sur internet.

Ils sont principalement constitués de trois parties : le robot (également appelé *spider* ou *crawler*), l'index (ou *base de données*) et le logiciel/interface d'interrogation.

Le robot repère les pages web par suivi récursif des liens présents dans les pages présentes sur Internet ou proposées par les auteurs. Les pages repérées sont Stockées dans l'index et indexées de façon automatique en texte intégral. Les pages qu'il ne peut pas repérer (pages orphelines ou d'accès réservé…) ne sont pas indexées. L'interface de recherche permet à l'utilisateur de faire une requête dont les termes seront recherchés dans l'index. Les résultats sont donnés selon un ordre de pertinence dépendant du type de moteur de recherche utilisé et l'utilisateur peut se rendre sur une page donnée grâce à un lien hypertexte **[Bou02]**.

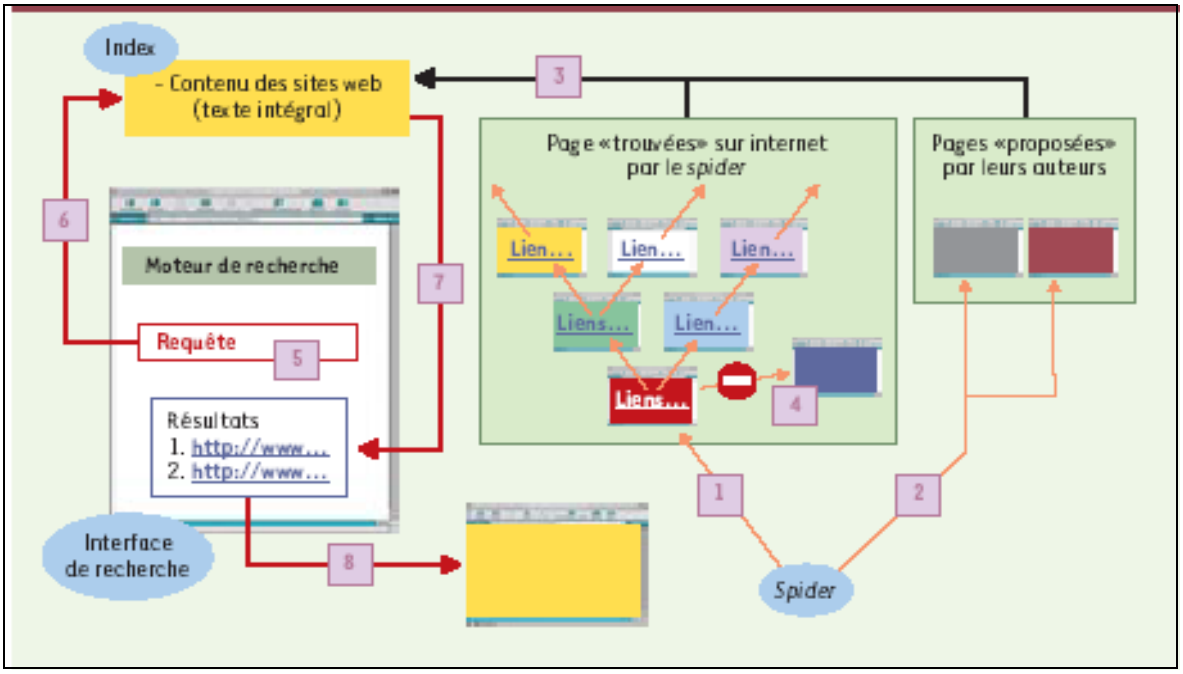



**Figure 1 : Mode de fonctionnement d'un moteur de recherche**

La plus part des systèmes de recherche se basent sur la liste inversée de structure de données, afin de permettre un accès rapide à la liste des documents contenant un terme, avec éventuellement d'autres information tels que le poids du terme dans un document, sa position relative etc.
La liste inversée peut être stockée sous la forme suivante :

$$t_i \longrightarrow \langle d_a,\ldots.\rangle, \langle d_b,\ldots\ldots\rangle,\ldots. \langle d_n,\ldots.\rangle$$

Où $t_i$ est le terme i contenu dans les documents $d_a$, $d_b$,…$d_n$.
La majorité des systèmes de recherche d'information utilisent les mots simples en tant que termes. Les mots qui sont jugés non informatifs tels que (the, in, of, a,….) en Anglais, (le, la, les, de, des, un, une, sur, sous……) en Français, ( ..،كـ،في على، عن، إلى، من ) En Arabe, encore appelés mots vides ou insignifiants (stopwords) sont souvent ignorés.

Beaucoup de systèmes s'intéressent aux diverses autres formes du mot et qui possèdent la même racine aussi appelés stemming dans le jargon de la RI, si nous cherchons par exemple, le mot « retrieval » l'idée principale du stemming est de pouvoir s'intéresser aux documents contenant : retrieve, retrieved, retrieving, retriever et ainsi de suite **[Sin01].** D'un autre coté, ceci permet

malheureusement, d'induire le système en erreur et donc, pour un utilisateur s'intéressant à « information retrieval » d'avoir en retour des documents intitulé « Information on Golden retrievers ». Nombreux stemmers ont été développés pour les différentes langues, chacun avec ses propres règles de stemming.

D'autres systèmes utilisent des phrases à mot multiples telle « Information retrieval » la phrase est ainsi indexée, cette phrase est considérée plus informative que des mots simples. Différentes techniques pour générer des listes de phrases sensées ont été explorées, ces techniques sont soit linguistiques (se basant sur l'analyse de phrases) soit statistiques (se basant sur le nombre de cooccurrences d'un mot). Les études comparatives ont échoué à montrer une nette différence entre les deux techniques, concernant la performance de la recherche **[Fag89].**

### 1.2.2. Les méta moteurs

Les méta moteurs tels que Copernic et méta crawler, permettent une interrogation simultanée de plusieurs index de moteurs de recherche différents. La requête saisie à l'aide d'une interface unique est transmise à ces différents index. Les résultats sont ensuite fournis après élimination des redondances et l'utilisateur peut ensuite se rendre sur les pages grâce à un lien hypertexte.

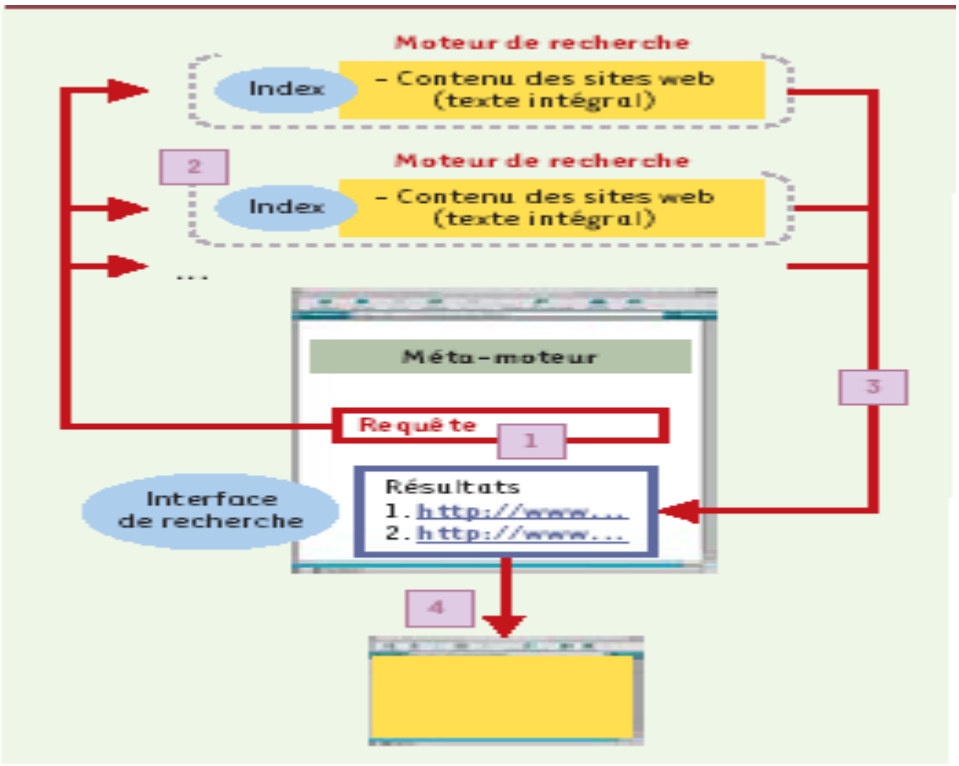



**Figure 2 : Mode de fonctionnement d'un méta moteur**

### 1.2.3. Les annuaires

La différence majeure entre les annuaires tel que Yahoo, par exemple et les moteurs ou les méta moteurs de recherche est l'intervention humaine et la sélection des pages. En effet, un indexeur examine les pages soit sur Internet, soit proposées par les auteurs. Seules les pages sélectionnées par l'indexeur seront stockées dans l'index, accompagnées d'une fiche descriptive et classées en fonction de leur contenu dans des catégories définies. En ce qui concerne la requête, l'utilisateur peut soit saisir un ou plusieurs termes qui sont recherchés dans les fiches descriptives de l'index et consulter la liste des sites correspondant à sa requête, soit naviguer lui-même dans les catégories grâce à des liens hypertextes. Dans les deux cas, l'accès au contenu des pages web s'effectue *via* un lien hypertexte.

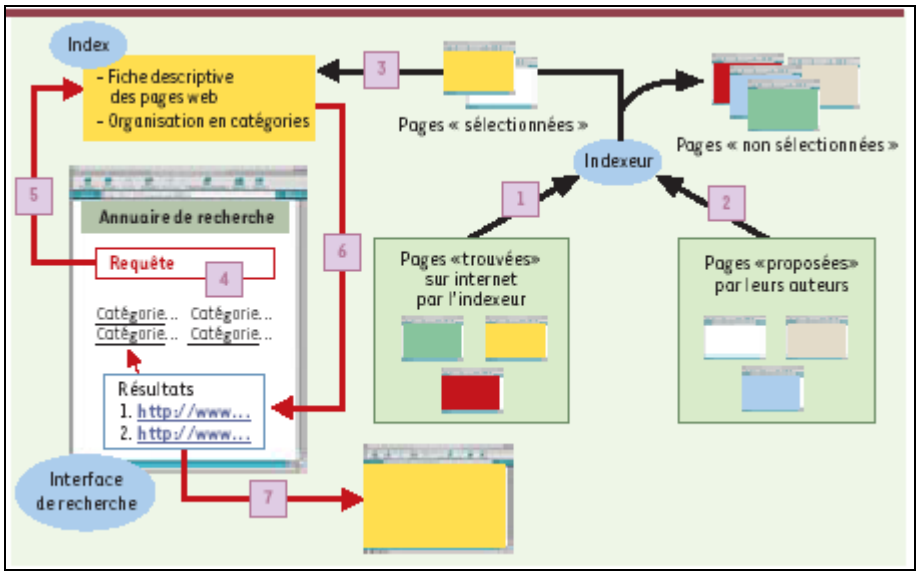



**Figure 3 : Mode de fonctionnemen4t d'un annuaire**

## 1.3. Problématique de la recherche d'information sur le Web

### 1.3.1. Introduction

Aujourd'hui, un grand capital des connaissances humaines dans les différents domaines se trouve sur le Web, et le volume de documents crées, stockés pour y être gérés ne cesse de croître, on estime à environ 7 millions le nombre de pages mis chaque jour sur le web.

Beaucoup de personnes croient que parce qu'il existe des moteurs de recherche le problème de la recherche d'information est résolue: loin de là nous savons que sur le Web, il y a une quantité impressionnante d'informations mais comment accéder à l'information pertinente : c'*est chercher une aiguille dans une botte de foin.*

En effet, la majorité des moteurs de recherche tels que Google, AltaVista, Hahooa (version Arabe de Voilà) et Ajeeb (proposé par Sakhr) se basent sur des mots clés, sans prendre en compte la sémantique de ces mots ni du contenu des différents documents indexés.

L'utilisateur se trouve submergé par un flot d'informations en majorité non pertinentes c'est ce qu'on appelle le *bruit*. D'un autre coté si une toute petite erreur d'orthographe glisse dans la requête le système ne répond pas. Il arrive aussi que le mot contenu dans la requête soit d'utilisation rare et que les documents indexés, contiennent plutôt ses synonymes, ces documents, pourtant pertinents, ne seront pas retournés c'est le *silence*.

Ces moteurs ne répondent pas aux besoins pointus de certains utilisateurs.
Certaines applications nécessitent la mise au point d'une technologie avancée de recherche d'informations telles les veilles technologiques, scientifiques, économiques, la gestion des connaissances dans l'entreprise, la protection de la propriété intellectuelle etc.….
Face à la grande prolifération de l'information électronique, le défi consiste à créer des moyens pour faciliter l'accès aux documents afin d'y rechercher et d'en extraire *rapidement* l'information pertinente.

Les documents sont indexés soit manuellement soit automatiquement, avec des mots choisis et jugés importants. Lors d'une recherche, un document est alors retrouvé et retourné s'il contient les mots-clés utilisés dans la requête, soit dans la liste des mot-clés le caractérisant, soit dans le corps du texte.
Avec Internet, l'utilisateur a, à sa disposition, une quantité impressionnante d'informations jamais encore archivée. Cependant, ces informations sont disséminées sur des systèmes informatiques hétérogènes. Aucun organisme officiel n'est chargé d'indexer les documents disponibles sur le réseau Internet. De plus, dans cet environnement hypertextuel et hypermédia, un document peut être de différente nature : texte, image, son, film vidéo. Tout le problème consiste à découvrir où se trouve l'information recherchée, et quelle est son adéquation par rapport à la question posée **[Roui00]**. Pour cela, des outils d'aide à la recherche ont été développés, notamment pour le Web. Ils suivent trois logiques de recherche distinctes : la recherche géographique, la recherche thématique, et la recherche par mots-clés (dans un index ou un méta index).

### 1.3.2. La recherche géographique

Elle se fait à l'aide d'un moteur de recherche géographique qui permet de se focaliser progressivement sur la région désirée en procédant étape par étape sur des détails de plus en plus fins (pays, région, villes).
Il existe plus de 80 outils de recherche géographique sur Internet. La recherche peut démarrer à partir d'une carte du monde ou de manière moins large à partir d'un pays visé. Sur *City.net* Par exemple, pour installer quelqu'un en écosse pour la prochaine année universitaire **[Roui00].**

### 1.3.3. La recherche thématique

Les moteurs de recherche thématique ou annuaire permettent d'avancer de thème en sous thème, l'utilisateur choisit une catégorie initiale parmi celles proposées puis approfondit la recherche afin d'aboutir au résultat recherché. Avec Yahoo, nous pouvons procéder à une recherche thématique.

### 1.3.4. La recherche par mot-clés

Un moteur de recherche par index compare les mots-clés donnés par l'utilisateur avec ceux stockés dans une base de données, qui caractérisent les pages Web référencées par ce moteur.
Les bases de données des moteurs de recherche sont mises à jour régulièrement, soit avec des informations obtenues par des robots informatiques, qui parcourent le Web continuellement, soit grâce à des données fournies volontairement par une personne qui veut voir son site référencé par ce moteur. À partir du titre, des mots-clés et du contenu d'une page Web, le moteur de recherche indexe cette page, de manière automatique. Quand un utilisateur soumet une requête, le moteur lui présente

une liste ordonnée d'hyperliens, pointant vers les pages Web censées contenir l'information recherchée.
La recherche sera d'autant plus efficace, si nous formulons les requêtes en utilisant des opérateurs booléens (ET, OU, SAUF, etc.), afin d'ajouter ou d'exclure certains mot-clés. Une requête avec comme mots-clés *fibre* ET *optique* donne des résultats plus proches du domaine de la télécommunication que de ceux du tissu ou de la lunetterie **[Roui00].**

Avec certains moteurs de recherche, d'autres éléments sont à prendre en compte pour obtenir un résultat optimum : l'ordre de saisie des mots-clés, les caractères accentués, les différences entre majuscules et minuscules, la langue, le type de document recherché. Nous nous intéresserons plus particulièrement aux moteurs de recherche supportant la recherche en Arabe tels que: Hahooa[1], Ajeeb[2], Ayna[3], Fattich[4], Konouz[5], Arabia[6], Google[7], Maktoob[8] , MSN Arabia [9]etc.

## 1.4. Les types de recherche d'information

### 1.4.1. La recherche Ad Hoc

L'utilisateur, avec un besoin particulier d'information, soumet une requête au système de recherche, en considérant une collection, supposée statique, de documents. Après traitement de la requête, le système retourne alors un ensemble de documents à l'utilisateur. Les moteurs de recherche sur le Web sont maintenant les systèmes de recherche Ad Hoc les plus utilisés, mais nous pouvons trouver ce genre de service dans quelques sites Web tels que les services d'information commerciale **[Haw00].**

[1] http://www.Hahooa.com

[2] http://www.Ajeeb.com

[3] http://www. Ayna. com.

[4] http://www.Fattich.com

[5] http://www.Konouz.com

[6] http://www.Arabia .com

[7] http://www.Google.com

[8] http://www.Maktoob.com

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

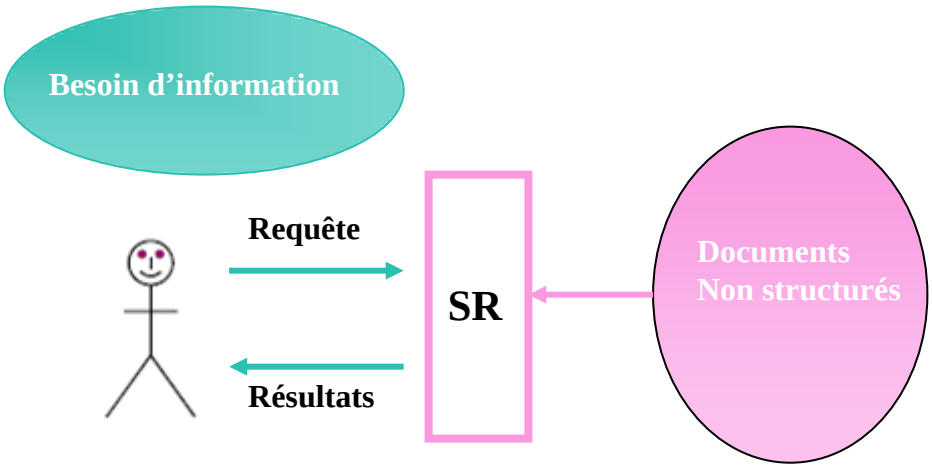



**Figure 4: Le modele de recherche Ad Hoc**

### 1.4.2. La recherche multimédia

Elle consiste à rechercher une information qui ne se limite pas à du texte mais elle peut être un document contenant des informations sous forme d'image, son, vidéo ou musique. Le problème qui se pose est comment établir une similarité entre la requête et un document. Certain moteur de recherche fournissent le service « *image search* » basé sur ce type d'information **[Haw00].**

### 1.4.3. La recherche interlingue

Avec l'expansion vertigineuse de l'Internet, le World Wide Web est devenu le moyen le plus populaire pour la diffusion de l'information multilingue **[Alj01].**
Les usagers des réseaux actuels d'information et des bibliothèques électroniques ne sont plus limités par les frontières géographiques ou spatiales, ils veulent pouvoir trouver, retrouver et comprendre l'information pertinente, où quelle soit et quelle qu'en soit la langue. Pour cette raison, on a donné beaucoup d'attention ces dernières années à l'étude et au développement d'outils et de technologies pour l'accès multilingue à l'information **[She97], [Pet01].**
La recherche d'information interlingue consiste à trouver et extraire une collection de documents, pour une requête donnée dans une langue différente de celle des documents **[Sad02]**. La requête est faite dans une langue pour retrouver des documents écrits dans une autre langue **[Haw03]**.
Dans CLIR, aussi bien la requête que les documents sont traduits. Pour cela il existe trois approches : La traduction automatique, les corpus parallèles, les dictionnaires à lecture automatique.
La traduction automatique, consiste à traduire une requête d'un langage humain à un autre. La désambiguïsation se base sur l'analyse syntaxique, cependant généralement la requête est composée de quelques mots sans structure syntaxique**[Pir98]**, pour cette raison la désambiguïsation n'est pas encore très efficace !**[Hul96],[Oar98].**

Dans la méthode des corpus parallèles, la requête est traduite en se basant sur les termes extraits des documents parallèles de la collection et un programme pour la traduction de la requête **[Dun93].**

Dans la méthode basée sur le dictionnaire, la traduction de la requête est obtenue en cherchant les termes dans un dictionnaire bilingue et en construisant une requête dans le langage cible, en ajoutant une partie ou la totalité des mots retrouvés **[She96].**

## 1.5. Les techniques de recherche d'informations

Le nombre très important de conférences et de publications relatives à la recherche de textes rend impossible une présentation exhaustive des méthodes. Néanmoins nous allons présenter ce que nous avons trouvé de plus important.

### *1.5.1 La catégorisation de documents*

La catégorisation de documents a émergé en réponse au problème posé par la recherche intelligente de documents, pour satisfaire les besoins des utilisateurs finaux des systèmes de recherche d'information. Elle consiste à affecter des documents dans des catégories prédéfinies selon leur contenu. La catégorisation automatique est utilisée dans les moteurs de recherche, les bibliothèques numériques et les systèmes de gestion de documents. Elle est utilisée pour que, lors d'une recherche d'information, un ensemble de documents similaires est retourné. Concernant la langue Arabe, un système de catégorisation automatique a été mis au point par Sakhr **[Sakhr04].**

Le traitement des documents se fait de la manière suivante :
On élimine les mots vides (mots insignifiants), les signes diacritiques (voyelles) sont enlevés s'il s'agit d'un texte avec voyéllisation, la racine est alors extraite de chaque mot du document, le module de classification est responsable de la classification d'un document dans une classe cible avec un calcul de probabilité à posteriori utilisant l'estimation obtenue avec un ensemble de test (avec des documents étiquetés) quand un document non marqué est présenté, une probabilité à posteriori est calculée pour chaque classe et le document est classé dans celle où la probabilité est la plus grande **[Elk04].**

### 1.5.2 Le chemin de lecture

**Le principe**
Les systèmes de la RI classiques ne sont plus adaptés aux spécificités du Web, ils considèrent que ce dernier est constitué d'un ensemble de documents indépendants. Ils omettent le fait que ces documents sont liés par des liens hypertextes. Plusieurs méthodes ont vu le jour depuis essayant de prendre ce coté en considération, prenant en compte aussi bien le type de page que le type de lien. L'une de ces méthodes est le chemin de lecture.
Dans cette méthode le plus petit élément considéré est non plus la page Web, mais la zone de texte.
Un document est constitué par plusieurs zones de texte, certaines sont pertinentes d'autres ne le sont pas. La réponse à une requête sera un document virtuel constitué d'un ensemble de zones de textes jugées pertinentes : c'est le chemin de lecture.
L'indexation d'un chemin de lecture $ch_j$ en un vecteur $CH_j$ consiste à extraire les termes ti représentatifs du chemin et à leur affecter une pondération $w_{i,chj}$ comprise entre 0 et 1, qui se base sur la fréquence locale du terme dans le chemin et la fréquence documentaire qui est le nombre de chemins dans lesquels le terme apparaît.

Pour extraire des chemins de lecture, l'algorithme utilisé se base sur les valeurs de similarité calculées entre les zones de texte (une page Web est constituée de zones de texte, celle ci étant la granularité la plus fine utilisée et elle peut contenir un ou plusieurs paragraphes).
L'article est supposé contenir une introduction, l'état de l'art, les travaux similaires, l'expérimentation etc. l'article est découpé en zones de texte, un algorithme est utilisé pour essayer de reconstruire un article à partir de zones de texte similaires**[Rad02]**

**Avantages et inconvénients**
**Avantage** : le fait de prendre la zone de texte comme granularité permet de se débarrasser des zones non pertinentes mais qui se trouvent sur la même page Web et laisser l'utilisateur chercher seul, l'information qu'il veut.
**Inconvénients**: si la zone de texte est constituée par un seul paragraphe (vecteur très petit) la mesure de leur similarité est difficilement utilisable et les résultats sont médiocres, toujours d'après les travaux de Radouani **[Rad02]**.

### 1.5.3. La requête par l'exemple

**Le principe**
Cette méthode se fait en deux étapes, dans la première, une requête est présentée au moteur de recherche, un certain nombre de documents est retourné par le système, une fois qu'un document est retourné, l'utilisateur peut copier une partie, un bloc de texte d'une taille quelconque et le coller dans le champs de recherche en tant que requête, c'est la deuxième étape. Cette technique est utilisée dans verity[10], car à la base elle a été destiné aux bases de données relationnelles.

**Avantages et inconvénients**
L'avantage est que cette technique peut-être considérée comme une extension de la requête. Son inconvénient est que le texte recopié va être une grande source d'ambiguïté.

### 1.5.4. Le résumé automatique

**Le principe**
Certains outils de recherche génèrent un résumé des documents présents dans la liste retournée des résultats, pour que l'utilisateur ne perd pas du temps dans le téléchargement d'un document qui s'avèrera non pertinent, après la lecture complète du fichier. Le résumé est donné avec les mots clés en surbrillance.
**Avantages et inconvénients**
Le résumé va aider peut être à gagner un peu de temps, mais un bon résumé nécessite néanmoins un traitement du langage naturel, qui reste, encore, fastidieux pour tout traitement informatique.

### 1.5.5. La recherche par question-réponse

**Le principe**
La requête est une question simple en langage naturel, la réponse par le système est aussi précise que la question **[Wan01].** Le système se compose de trois parties :

a. L'analyseur de question

[10] http://www.verity.com

b. L'étiqueteur des entités nommées
c. L'extracteur de réponse.

### a. L'analyseur de questions

Après avoir soumis La question, celle-ci est analysée et une liste de mots est générée, dans l'étiqueteur des entités nommées, les premiers documents retournés sont analysés et les entités nommées sont extraites à partir de ces documents. Finalement l'extracteur de réponse, détermine les réponses pertinentes à partir des entités nommées, utilisant des questions-types et une liste de mots clés.
Exemple :

| **Réponse-type** | **Question-type** | **Exemple** |
|---|---|---|
| Personne | Qui/que/quelle personne/laquelle (personne) | Mr Mohamed |
| Lieu | Où/quel lieu/lequel (lieu) | Annaba |
| Organisation | Quelle organisation/laquelle (organisation)/qui | ONFA |
| Argent | combien | 2 millions DA |
| pourcentage | Combien/quel pourcentage | 6% |
| Date | Quand/quelle date | 15/12/1994 |
| nombre | combien | 3 |
| Durée | Depuis quand | 9 ans |
| Distance | Combien/ quelle distance/quelle longueur | 15 KM |
| Surface | Quelle surface/ quelle grandeur | 10m2 |
| Estimation | Quelle mesure/ quel poids/quelle vitesse | 10k m/H |
| Devise | Quelle devise | Dinar |
| Nationalité | Quelle nationalité/ quelle langue | Algérienne/Arabe |
| Raison | Pourquoi/ comment | - |
| Nom | Qui/ quel est | - |
| Pas de réponse | tout | nil |

**Figure 5: Les questions-types et les réponses-types**

### b. L'étiqueteur des entités nommées

C'est un outil qui permet l'extraction des entités nommées telles que PERSONNE, LIEU, ARGENT, POURCENTAGE etc.

Avant d'identifier les entités nommées, nous avons besoin de construire des passages candidats en utilisant les documents retrouvés en tête, lors de la recherche.

Généralement les étapes de l'algorithme sont :

1. Analyser les phrases des documents
2. retrouver les phrases contenant les mots clés de la question
3. construire un passage candidat pour toutes les deux phrases (si une phrase est assez longue, elle devient elle-même un passage candidat).
4. assigner à chaque passage candidat un score initial égal au score du rang du document.
5. calculer le nombre de mots clés associés dans chaque passage candidat.

6. calculer la taille de la fenêtre d'association, qui est définie en tant que nombre de mots clés dans un passage candidat entre le premier mot clé associé et le dernier.
7. réarranger les passages candidats par leur score final puis afficher les premiers passages des entités nommées.

**c. L'extracteur de réponse**

L'extracteur de réponse compare la question-type avec chaque entité nommée dans les passages candidats, si un passage candidat contient une entité nommée associée à la question-type le score du passage est augmenté de 100. Les passages sont réarrangés, les premières entités nommées (en général 5) sont retournées en tant que réponse finale.

### 1.5.6. *Recherche par traitement du langage naturel*

Il fait partie des outils proposés pour améliorer la recherche d'information, mais son succès reste très limité **[Str97].** Cette interrogation dite " en langue naturelle ", où l'expression du thème de recherche prend la forme d'une demande. Cela est souvent présenté comme une prouesse technique ( la machine comprend, dans la langue de tous les jours ") et comme un supplément de convivialité. Mais à l'usage, on préfère entrer au clavier quelques mots en relation avec le thème de recherche, que de rédiger une demande. De plus, la difficulté principale n'est peut-être pas d'exprimer le thème de la recherche sous une forme recevable par le moteur de recherche, que de maîtriser la manière dont le moteur va interpréter la requête et procéder à la recherche, autrement dit comment bien lui faire comprendre ce que l'on cherche. Quant à l'analyse linguistique de la demande formulée, d'une part elle est effectivement très complexe (tolérance aux erreurs, portée des négations, résolution des anaphores…) et donc met durement à l'épreuve les performances et la robustesse des analyseurs, d'autre part si l'analyse consiste à transposer la demande sous la forme d'une équation booléenne élémentaire, alors il serait plus sûr, plus puissant et plus efficace de l'écrire directement.

### 1.5.7. Les méta données et le Dublin core

**Les Meta données**

Les méta données sont des données sur les données ou des données décrivant d'autres données. Le terme de méta donnée est constitué de deux parties : "méta" et "donnée". La composante "méta" révèle une volonté d'abstraction à un niveau supérieur et introduit aussi la notion de réflexivité. Les méta données doivent donc compléter l'information relative aux données à un niveau d'abstraction supérieur tout en étant capables de se décrire elles-mêmes. La composante "donnée" indique que les méta données sont aussi des données et peuvent donc être structurées et interrogées.

A partir de l'usage des méta données, il convient de préciser que la description des documents web par ces éléments n'est pas un objectif final mais plutôt un moyen pour faciliter l'usage de ces documents dans une perspective de recherche d'informations. Dans ce cadre, plusieurs projets ont été engagés dans un objectif d'unification de la description des documents web. Parmi ces projets, on trouve le Dublin Core.

**Le Dublin Core**

C'est en 1995, que des chercheurs en informatique avec des spécialistes venant des bibliothèques et du domaine du codage des textes se sont réunis à Dublin (Ohio - USA) avec comme objectif de définir des propriétés pour la description des documents électroniques conservés en réseau.

Ce groupe a retenu un ensemble d'éléments susceptibles d'être intégrés aux documents électroniques afin de les identifier automatiquement. Ils constituent le DUBLIN CORE META DATA ELEMENT SET. Ces éléments sont connus sous le nom de Dublin Core.
Le Dublin Core vise, depuis sa création, à résoudre le problème de la description unifiée des ressources d'information électroniques et de leur localisation dans un contexte réseaux **[Ben99].** C'est dans cette perspective qu'il est devenu une norme ISO 15836 depuis février 2003.

**Le schéma descriptif de Dublin Core**
Selon la norme ISO 15836, le Dublin Core propose une quinzaine d'éléments descriptifs. Ces éléments sont les suivants :
**Title :** nom donné à la ressource par son auteur ou son organisme. Le titre c'est le nom par lequel la ressource est officiellement connue.

**Creator :** entité (personne physique ou morale) responsable de la création du contenu intellectuel de la ressource. Il s'agit de l'auteur principal dans le cas d'un document écrit.

**Subject :** description du domaine sémantique par des mots clés ou des phrases ou un code de classification précisant le sujet de la ressource. L'usage de vocabulaires contrôlés pour cet élément est encouragé.

**Description :** Description textuelle du contenu. Généralement cet élément contient un résumé descriptif sur le contenu de la ressource.

**Publisher :** entité responsable de l'édition et la publication de la ressource.

**Contributor :** personne (physique ou morale) qui a collaboré à la production du document, exemple : illustrateur, traducteur...

**Date :** date de création ou de publication de la ressource conformément au format ISO 601 (AAAA-MM-JJ) ex : 2002-12-25, ou simplifiée 2002.

**Type :** catégorie à laquelle appartient la ressource : roman, poème, thèse, etc.

**Format :** c'est la matérialisation physique ou digitale de la ressource (texte, son, image). Cet élément peut être utilisé pour préciser le logiciel ou autre équipement nécessaire pour afficher la ressource.

**Identifier :** identification unique de la ressource par un URI (Uniform Resource Identifier) qui peut inclure l'URL (Uniform Resource Locator) ou l'ISBN (International Standard Book Number).

**Language** : langue du contenu intellectuel de la ressource sous forme d'un code. La valeur de l'élément langue doit respecter les directives en vigueur, c'est pour cela qu'il est recommandé d'utiliser les codes définis par le schéma du RFC 3066[11] (ISO 15836). Ce schéma donne un code à chaque langue à deux ou trois caractères selon la norme ISO 639, et dans certains cas il sera suivi d'un code à deux caractères pour le pays. Par exemple « ar » pour l'arabe, « fr » pour le français et « en-GB » pour l'anglais utilisé en Grande Bretagne.

[11] Étiquettes pour l'identification des langues, RFC 3066 d'Internet. (http://www.ietf.org/rfc/rfc3066.txt)

**Relation :** identificateur d'une seconde ressource ayant une relation avec la première.

**Coverage :** la couverture spatiotemporelle de la ressource. Il est recommandé d'utiliser un vocabulaire contrôlé pour choisir la valeur de cet élément.

**Rights :** cet élément couvre les droits de propriété intellectuelle (copyright). Cet élément doit être mentionné pour préserver tous les droits des créateurs de la ressource. Si cet élément est absent de la description, aucune hypothèse ne peut être faite sur l'état des droits des différents créateurs.

**Source :** référence à une source à partir de laquelle le document est dérivé. Il est recommandé de référencer cette source par une chaîne de caractères.

Il est à noter que chaque élément est optionnel et répétitif. De plus, ces éléments peuvent apparaître dans n'importe quel ordre. Nous pouvons constater aussi que la définition des éléments est purement sémantique : elle ne fait aucune hypothèse sur les langages formels et sur les outils logiciels qui peuvent être employés pour créer des descriptions, les associer aux ressources et les exploiter dans les moteurs de recherche comme l'affirme par ailleurs Michard **[Mic99].**
Par exemple, l'élément méta "DC. Subject " fournit au moteur de recherche les mots clés selon lesquels le contenu du document pourrait être indexé. Ce qui permettrait de décrire aussi fidèlement que possible le contenu du document, pour assurer une pertinence lors de la réponse à la requête de l'utilisateur, et d'éviter ainsi au maximum, le bruit ou le silence dans les résultats.
Les méta données ont un rôle naturel d'aide à la structuration et à la recherche d'informations. De manière générale, les méta données sont très répandues dans les systèmes multimédias et dans le monde des sciences environnementales. Les applications s'appuyant sur les méta données en biologie (UMLS par exemple) restent encore marginales **[Dui99].**

### 1.5.8. L'expansion de la requête

En utilisant un simple mot-clé pour rechercher une information, il est supposé implicitement qu'il existe une correspondance unique entre le mot-clé et le sens, ce qui est bien évidemment faux, parce qu'en réalité un mot peut avoir plusieurs sens, et un sens peut être exprimé par plusieurs mots différents.
Pour traiter ce problème, on propose de considérer des mots ou termes reliés pour étendre la requête.
L'utilité de cette extension dépend de deux facteurs :

- Comment choisir les mots pour étendre la requête ?
- De quelle manière rajouter ces mots à la requête ?

La majorité des techniques utilisées pour étendre la requête et qui se basent sur la génération automatique de thesaurus, n'ont pu montré une amélioration des résultats obtenus, ceci peut s'expliquer par le fait que cette extension ne se fait pas dans le contexte de la requête, en effet, si le mot « machine » est fortement relié au mot « moteur » une éventuelle extension serait insignifiante dans le cas de « moteur de recherche ».
En 1965, Rocchio propose l'utilisation du *feedback pertinent* « relevance feed-back » **[Roc71]**. Le principe est simple : une requête utilisateur est soumise au système de recherche, qui retourne une liste ordonnée de documents, l'utilisateur va juger de la pertinence des documents retournés et le système va alors générer automatiquement une nouvelle requête à partir des documents les plus pertinents de la collection.

De nouvelles techniques pour l'expansion de la requête en l'absence du feed-back utilisateur ont été développées dés les années 90. La plus importante est *le pseudo-feedback* **[Buc95]**.

Une expansion de requête peut être définie comme un élargissement du champ de recherche pour cette requête. Ce traitement est souvent vu comme un moyen augmentant le taux de rappel, cependant en considérant les documents contenant les termes reliés, nous avons de forte chance d'augmenter la précision.
La requête étendue contiendra en plus des mots initiaux de la requête, ceux rajoutés soit qu'ils sont des synonymes, hyperonymes, hyponymes, soit qu'ils sont reliés par la règle de cooccurrence.
L'expansion de la requête peut-être effectuée de différentes façons, dans la méthodes dite du *pseudo feedback pertinent* **[Sad01]** une première recherche est effectuée, on extrait des 50 premiers documents retournés, jugés pertinent par le système de recherche, les 10 termes qui cooccurrent en conjonction avec les mots de la requête initiale pour établir une nouvelle requête.

Claveau et Sebillot proposent des liens sémantiques nom-verbe acquis sur corpus pour étendre la requête **[Cla04].**
L'idée est d'extraire automatiquement des couples nom-verbe en relation *qualia*, comme *disque dur-stocker* ou *lettre-communiquer*, de la base documentaire à l'aide du système ASARES **[Cla03**]. Une relation *qualia* décrit à l'aide de prédicats essentiellement verbaux les différentes facettes sémantiques des noms (fonction, mode de création…) par exemple, le nom *couteau* et le verbe *couper* sont en relation qualia (*couper* représente la fonction de *couteau*) Ces couples sont ensuite utilisés pour étendre les requêtes d'un système de recherche.

Qiu et Frei ont expérimenté l'expansion avec un thesaurus, l'idée sous jacente sur laquelle ils s'appuient est que tout terme étroitement lié à un terme d'indexation peut lui-même être utilisé comme terme d'indexation. En pratique, ces termes « étroitement liés » sont calculés à partir des cooccurrences fréquentes des mots, par des méthodes essentiellement numériques, Un thésaurus est ainsi construit.
Lors d'une interrogation, aux termes de la requête sont alors rajoutés les éléments du thésaurus qui leur sont proches, soit en considérant chaque mot de la requête indépendamment, soit en considérant l'ensemble de la requête **[Qiu95].**
L'efficacité de ces approches d'extension de requêtes par cooccurrence est variable selon les travaux, mais aucune amélioration franche des résultats ne semble se dégager, quelle que soit la collection de documents.
Peat et Wilett **[Pea91]** expliquent ce phénomène, par le fait que les méthodes utilisées, pour l'extraction des cooccurrences, favorisent l'acquisition de termes approximativement de même fréquence. Or si les termes de la requête sont très fréquents, les termes ajoutés sont eux aussi trop fréquents pour être discriminants.

L'expansion peut aussi être faite en utilisant un thesaurus. Un thesaurus est un vocabulaire de termes contrôlés d'indexation, structuré de manière à ce qu'il mette en évidence les relations *a priori* entre les concepts. Comme une liste de mots-clés, c'est un instrument qui utilise une terminologie normalisée et contribue à aider l'utilisateur à sélectionner de manière organisée des occurrences dans une base de données. Cet outil documentaire offre en outre un certain nombre de développements et d'enrichissements propres à l'organisation des thesaurus monolingues.
Un thesaurus se distingue d'une liste de termes par les points suivants :

- Il permet de regrouper les termes d'un même domaine à l'intérieur d'une hiérarchie, et de les mettre en relation avec des termes d'autres domaines ;

- La relation hiérarchique permet d'accéder à des concepts plus larges ou plus étroits à l'intérieur d'un même domaine ;
- Lorsque plusieurs termes peuvent rendre compte d'un même concept, l'utilisateur est guidé vers le terme préférentiel choisi par l'indication des autres termes possibles dans le champ " employé pour "
- Le thesaurus est un outil dynamique qui peut être mis à jour par ajout, modification ou suppression de termes ou relations entre termes.

Nous pouvons donner l'exemple du thésaurus d'EDF, qui est constitué de :

- Une liste alphabétique des descripteurs, suivis de leur traduction anglaise.
- Un ensemble de liaisons internes entre les descripteurs :
  -Liaisons associatives : **VA** (voir aussi) ;
  -Liaisons hiérarchiques : **TG** (a pour terme générique, **TS** a pour terme spécifique) ;
  -Liaison de synonymie : **EP** (employé pour) ;
- EM employer
- E2 employer conjointement

Le thésaurus d'EDF contient 13 852 descripteurs et 6 868 synonymes soit 20 720 racines (Descripteurs + synonymes).
Les descripteurs et synonymes sont répartis en 328 champs avec 306 schémas fléchés et 22 listes, eux mêmes regroupés dans 44 thèmes. Le nombre moyen de descripteurs est d'environ 40 par schéma fléché, certains champs cependant, sont plus denses, dans des domaines où les Préoccupations d'EDF nécessitent un vocabulaire plus pointu.

Une autre manière d'étendre la requête, consiste à rajouter les mots déduits, après association des mots clés de la requête initiale avec les concepts d'une ontologie.
L'expansion de requête est souvent utilisée dans le modèle booléen et le modèle vectoriel.

Rappelons d'abord les différents modèles utilisés dans la recherche d'information, nous pouvons citer le modèle booléen, le modèle probabiliste, le modèle réseau d'inférence et le modèle vectoriel.

**Le modèle Booléen**

Le modèle booléen est basé sur la théorie des ensembles et l'Algèbre de Boole, où le document est considéré comme un ensemble de termes et la requête est une expression booléenne, nous pouvons le considérer ainsi:
D: est un ensemble de mots (les termes indexés) présents dans le document.
Q: est une expression booléenne.
Les mots sont les termes indexés, reliés par les opérateurs AND, OR, NOT.
F: Algèbre de Boole sur les ensembles de termes et les ensembles de documents.
R (q, di): Le rangement ou fonction de similarité qui ordonne les documents selon leur pertinence par rapport à la requête.
Cette fonction est déterminée ainsi (pour un document donné d, $t_i$ appartenant à q):

$$R(t_i , d) = 1 \text{ si } t_i \in d;\ 0 \text{ sinon.}$$

Le modèle booléen est un modèle simple à comprendre et à implémenter, mais les expressions booléennes complexes sont difficiles à gérer. Tous les termes ont la même importance.

**L'expansion dans le modèle booléen**

Le processus d'expansion se fait ainsi : en considérant qu'une relation forte existe entre mot1 et mot2, si mot1 apparaît dans une requête booléenne, alors on remplace mot1 par (mot1 $\vee$ mot2). Mot2 est donc considéré comme un synonyme. Si on considère que mot2 est un bon substitut de mot1, alors la requête étendue ne change pas la sémantique de la requête initiale. En général, une requête booléenne n'est pas pondérée. Il n'y a donc pas de question de pondération pour les nouveaux mots **[Qiu93].**

**Le modèle probabiliste**

Ce modèle se base sur le principe général que les documents dans une collection sont rangés selon leur probabilité décroissante de leur pertinence, ceci est souvent appelé PRP (probabilistic ranking principle) **[Rob77]**. Les systèmes basés sur le modèle probabiliste, estiment la probabilité de pertinence des documents pour une requête, cette estimation est la partie essentielle du modèle. La première idée de la recherche probabiliste a été proposée par Maron et Kuhns dans un papier publié en 1960 **[Mar60]** depuis, plusieurs modèles probabilistes ont été proposés, chacun utilise une technique différente d'estimation de probabilité. La probabilité de pertinence d'un document D est notée P(R|D) **[Sin01]**, le critère de rangement peut aussi être calculé par

$$\mathbf{Log}\frac{P(R/D)}{P(\vec{R}/D)}.$$

**formule 1: Probabilité de pertinence d'un document**

Où $P(\overline{R}/D)$ est la probabilité pour que le document soit non pertinent.

**Le modèle réseau d'inférence**

Dans ce modèle le processus de recherche est modélisé tel un processus d'inférence dans un réseau d'inférence, dans la plus simple des implémentations un document instancie un terme d'une requête donnée avec une certaine force, le crédit des différents termes est accumulé donnant à une requête l'équivalent d'un score numérique pour le document. La force de l'instanciation du terme pour le document peut être considéré comme le poids d'un terme dans le document et le rangement devient similaire au rangement dans le modèle vectoriel et le modèle probabiliste. La force de l'instanciation n'est pas définie par le modèle et n'importe quelle formulation peut être utilisée.

**Le modèle vectoriel**

Dans le modèle vectoriel, le texte est représenté sous forme de vecteurs de termes. Les termes sont des mots ou des phrases. Pour assigner un score à un document, le modèle mesure la similarité entre le vecteur requête et le vecteur document. Cette similarité permet d'ordonner les documents en fonction de leur ressemblance avec la requête. L'angle compris entre les deux vecteurs est utilisé comme une mesure de divergence entre ces vecteurs et le cosinus de l'angle est utilisé comme mesure de similarité (puisque le cosinus a la propriété d'avoir la valeur 1 pour deux vecteurs identiques et la valeur 0, pour deux vecteurs orthogonaux).

Si $\vec{D}$ est un vecteur document et $\vec{Q}$ un vecteur requête, alors la similarité entre un document

D et une requête Q (ou score du document D pour la requête Q) peut être représentée par :

$$Sim(\vec{D},\vec{Q}) = \sum_{t_i \in Q,D} w_{t_{iQ}} \cdot w_{t_{iD}}$$

**Équation 1: Similarité entre un document et une requete**

Où $wt_{iQ}$ est la valeur de la i ème composante du vecteur $\overrightarrow{Q}$ et $wt_{iD}$ est la i ème composante du

Vecteur $\overrightarrow{D}$ (puisque tout mot qui ne se trouve pas dans la requête ou le document, a son $wt_{iQ}$ , $wt_{iD}$ respectivement égal à 0, $wt_{iD}$ désigne souvent le poids du terme i dans le document D)

**L'expansion dans le modèle vectoriel**

Dans le cas de l'expansion, toujours en considérant qu'une relation forte existe entre mot1 et mot2, si un mot apparaît dans une requête vectorielle, nous allons simplement ajouter mot2 dans le vecteur de la requête (s'il n'y est pas déjà). Cette méthode d'ajout est généralement utilisée, le système de recherche Smart **[Sal71]** utilisé dans les conférences pour l'évaluation des systèmes de recherche d'information telle que TREC utilise le modèle vectoriel.

La question qu'on se pose est plutôt sur la pondération des nouveaux mots dans le vecteur.

Elle peut se faire ainsi:

- Un mot ajouté mot2 est pondéré comme le mot initial mot1 en relation.
- Un mot ajouté mot2 est pondéré comme la pondération de mot1 multiplié par un facteur.

Dans le dernier cas, ce facteur peut être fixe (par exemple, 0.5), ou bien déterminé selon le nombre de mot2 en relation avec mot1. L'idée est que si mot1 donne un grand nombre de mot2, ceux-ci doivent être pondérés plus faiblement, cette méthode a été expérimentée par Voorhees qui a travaillé sur l'expansion de requête avec Wordnet, la performance a été dégradée **[Voo94]**.

La raison de cette dégradation peut trouver une explication dans le fait que l'expansion n'est pas uniforme pour tous les mots de la requête. Seulement certains mots seront étendus, et certains sont étendus plus fortement (par plus de mots) que d'autres. En conséquence, un concept qui est fortement étendu sera renforcé dans le vecteur obtenu, car il est maintenant représenté plusieurs fois par le mot initial et par tous les mots ajoutés. Est-ce que ces concepts renforcés sont réellement importants dans la requête? Pas nécessairement. Étant donné la manière d'évaluer la similarité, il est possible qu'un document retrouvé ne concerne qu'un seul concept, il contient plusieurs représentations de celui ci, mais aucun autre concept. Ce document pourrait être mieux classé qu'un autre document qui concerne tous les concepts de la requête (selon la pondération des mots). Salton parle de la spécificité et de l'exhaustivité. La spécificité d'un document détermine si tout le contenu du document est concentré sur le thème de la requête, alors que l'exhaustivité veut mesurer si tous les aspects de la requête ont été abordés dans le document.

Les mots utilisés pour faire l'expansion de la requête doivent être fortement reliés à celle-ci. Généralement, on utilise un dictionnaire de synonyme, un thésaurus ou une ontologie. Les mots reliés avec des mots de la requête par certains types de relation ( IS_A) sont choisis pour étendre la requête, ce sont les hyponymes et les hyperonymes, dans le cas d'ontologies.

Il y a aussi des études qui essaient de trouver automatiquement les mots fortement liés. La plupart de ces approches exploitent les cooccurrences: Plus deux mots cooccurrent dans un texte, plus on suppose qu'ils sont fortement liés. Une fois ces relations statistiques choisies, on peut les utiliser dans un processus d'expansion de requête.

Ces méthodes statistiques se basent sur la fréquence des cooccurrences, elle utilisent généralement *l'information mutuelle* **[Gal91]** définie par :

$$MI(w1,w2) = Log_2\left[\frac{N\ f(w1, w2)}{f(w1)\ f(w2)}\right]$$

**Équation 2: L'information mutuelle**

Où, N est la taille du corpus, f(w) est le nombre de fois où le mot w est cité dans le corpus et f(w1,w2) est le nombre de fois où les mots w1 et w2 sont cités ensemble dans une phrase**[Sad01]**.
La plupart des techniques d'expansion considèrent chaque mot de la requête isolé des autres mots. Certains pensent qu'il vaut mieux choisir des mots qui sont reliés à la requête qu'aux mots individuels de la requête **[Qiu93].** Autrement dit, ils calculent la relation entre un mot et la requête dans son ensemble, et optent pour les mots les plus fortement reliés. Ils montrent que cette approche est meilleure.
Il est aussi suggéré que le processus d'expansion soit interactif: L'usager peut filtrer les mots proposés par le système. Cette approche est utilisée dans certains systèmes, par exemple, *Medline* qui intègre un thésaurus du domaine médical.
Avantages ou inconvénients de cette méthode dépendent étroitement de la manière dont nous étendons cette requête, si par exemple nous utilisons les synonymes, c'est simple à réaliser mais ne donnent pas toujours de bons résultats.
Si par contre nous utilisons un thesaurus, où les mots sont reliés et font partie de la même catégorie, le coté sémantique n'est pas vraiment exploité, enfin les ontologies c'est une méthode efficace mais reste coûteuse.
C'est cette méthode que nous avons choisie pour l'amélioration de notre recherche d'information en Arabe sur le web. Vus les résultats encourageants et prometteurs obtenus un peu partout dans le monde, même si à chaque fois elle sont utilisées d'une manière différente**.**

## 1.6. Le problème d'évaluation

Au début des études s'intéressaient à comparer la recherche sur le titre, le résumé ou le contenu des documents. Les résultats ont montré que la recherche sur le contenu était supérieure à celle sur le résumé, mais Salton a précisé que cette supériorité n'était pas assez significative pour conclure d'une manière univoque que la recherche sur le contenu était toujours supérieure que celle sur le résumé **[Ten90].**

Une évaluation objective d'une recherche efficace est la pierre angulaire de la RI, Les tests de Cranfield, ont conduit en 1960 à établir un ensemble de caractéristiques requis pour les systèmes de recherche, après un long débat les propriétés retenues par la communauté de la RI étaient : le *Rappel* qui représente la proportion de documents pertinents retrouvés par le système et la *Précision* représentant la proportion de documents pertinents **[Cle67].**

Il est admis qu'un bon système de recherche doit d'un coté retourner le plus grand nombre possible de documents pertinents (augmenter le Rappel) d'un autre coté il doit retrouver le moins possible de documents non pertinents (avoir une haute précision). Malheureusement, ces deux facteurs sont plus au moins contradictoires, les techniques qui tentent d'améliorer le Rappel ont tendance à détériorer la Précision et vice versa. Si les concepteurs du système de RI jugent que la précision est plus importante pour leurs utilisateurs, ils peuvent utiliser la précision en priorité pour les dix ou vingt premiers documents lors du rangement, d'un autre coté si le rappel est plus important pour les

utilisateurs, certains mesurent la précision, disons 50% du rappel indiquant combien de documents non pertinents l'utilisateur doit lire pour avoir une moitié de documents pertinents **[Sin01].**

La recherche d'information s'évalue donc, en terme de *Rappel* et de *Précision* ce sont les plus importants et les mieux connus des mesures pour l'évaluation des systèmes de RI. De là, découle les notions de *bruit* (qui est le nombre de documents non pertinents retournés sur le nombre de documents retournés par le système) et de *silence* (qui est le nombre de documents pertinents non retournés sur le nombre de documents pertinents existant dans la collection).
Ces facteurs sont calculés de la manière suivante :

| | pertinent | Non pertinent | total |
|---|---|---|---|
| retrouvé | a | b | a+b |
| Non retrouvé | c | d | c+d |
| total | a+c | b+d | a+b+c+d |

**Figure 6: Table de mesure de pertinence d'un système de RI**

**1) Le rappel**

Le rappel est le nombre de documents pertinents retrouvés sur le nombre de documents pertinents dans la collection.

**Rappel = a/ (a+c)**

**2) La précision**

La précision est le nombre de documents pertinents retrouvés sur le nombre de documents retournés.

**Précision = a/ (a+b)**

**3) Le bruit**

Le bruit est le nombre de documents non pertinents retournés sur le nombre total de documents retrouvés

**Bruit = b/ (a+b)**

4) **Le silence**

Le silence (aussi appelé facteur d'omission) est le nombre de documents non retrouvés sur le nombre de documents pertinents présents dans la collection.

**Silence = c/ (a+c)**

Dans TREC la plus connue des conférences, qui s'est spécialisée dans l'évaluation de nouvelles techniques pour la recherche d'information, le Parcheur est souvent confronté à la précision **[Haw00].**

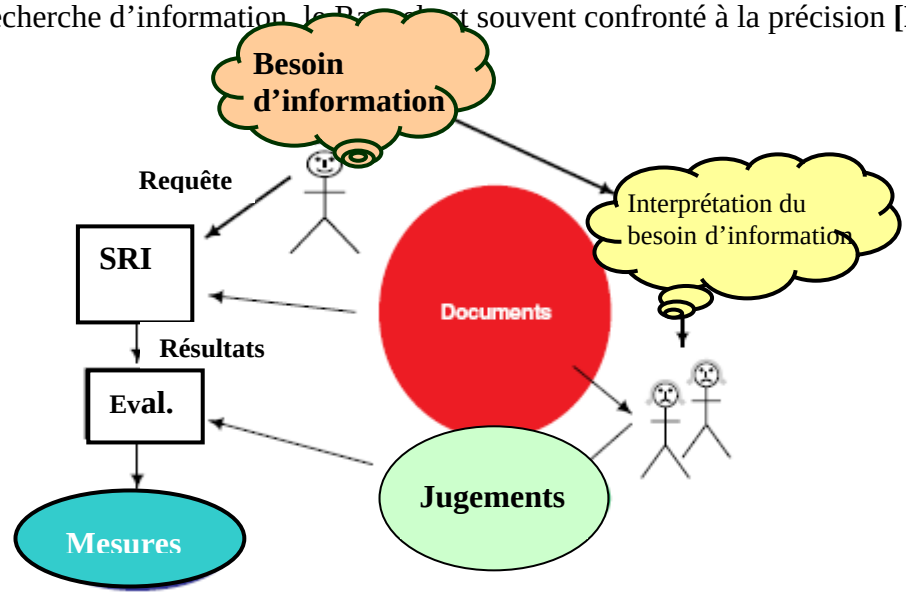



**Figure 7: Paradigme d'évaluation de la recherche dans TREC**

## 1.7. Les systèmes de recherche d'information en Arabe

La langue Arabe est l'une des six langues officiels de l'ONU, langue mère de plus de 300 millions de personnes, **[EDC01]** elle a connu un développement rapide, les statistiques montrent que depuis 1995 date à laquelle le premier journal Arabe « www.asharqalawsat.com »a été mis en ligne le nombre de sites web ne cesse d'augmenter exponentiellement. En 2000 il y a eu environ 20 mille sites, ce qui représentait 7% de l'ensemble des sites sur le Web.
Le nombre d'internautes Arabophones en 2002 était environ de 4,4 millions, ce qui représente 1,5% de la population du monde Arabe. Les standards pour l'évaluation d'outils de recherche d'information sur le web n'ont été introduits qu'en 2000 par TREC et CLAF.
L'Arabe est une langue sémitique qui s'écrit de droite à gauche, son alphabet est constitué de 28 lettres pouvant être étendu à quatre vingt dix éléments en rajoutant les formes, les voyelles, les marques.
Les mots Arabes sont classés en Noms(Adjectifs et adverbes), Verbes et Particules, tous les verbes et la majorité des noms sont dérivés d'une racine généralement trilittère, cependant elle peut être formée de deux, quatre ou plus rarement de cinq lettres**[Tay90].**
Il existe deux approches pour designer un système de RI en Arabe : la première est d'arabiser un système existant à partir d'un autre langage, généralement l'Anglais, dans un format capable de supporter le texte Arabe. Ce sont les systèmes arabisés.
La seconde approche consiste à construire un système Arabe manipulant des textes Arabes **[Abd04].**

### 1.7.1. Les systèmes arabisés

Les premiers systèmes arabisés ont été utilisés dans les bibliothèques numériques Arabes, tels que DOBIS/LIBIS (utilisé en Arabie Saoudite) , STAIRS , MINISIS utilisé dans le centre Saleh Kamel à

l'université d'El Azhar pour la recherche d'information dans les 16 livres d'Essirah :tradition du prophète Mohamed ASWS.
Ces systèmes sont faciles à implémenter au prix d'abandonner quelques caractéristiques de la langue Arabe. Seulement ces systèmes n'ont pas eu de succès à cause de la différence due à la nature de la langue s'écrivant de droite à gauche, la structure des mots et des phrases Arabes. De plus ces systèmes ne permettent pas de représenter les signes diacritiques Arabes ni certains symboles et caractères du script Arabe. Ces systèmes ne considèrent pas l'analyse linguistique de la langue Arabe, ce point est très important dans la recherche d'information en Arabe. L'insuffisance de ces systèmes résident dans le fait qu'ils aient été conçu pour manipuler des caractères non Arabes, ils ne sauront donner pleine satisfaction pour la manipulation des caractères Arabes.

### 1.7.2. Les systèmes Arabes

Vu l'échec des systèmes arabisés nombre d'organisations ont pensé à développer des systèmes de recherche d'information entièrement en Arabe.
Nous pouvons citer les travaux de Kaseem **[Kas88]**, le système *Al Bukhary* ce nom vient du texte intégral de Sahih el Bukhary (environ 7000 Hadiths du prophète Mohamed ASWS) développé par Al Naim**[AlN89]** intitulée « text analysis and automatic indexing for Arabic based automated information retrieval system », le système *Bayan* de Morfeq **[Mor91],** D'autres systèmes ont été le fruit de projet de thèses doctorales tels que : microcomputer based Arabic information retrieval system comparing words, stems and roots as index terms, d'Al Kharashi **[AlK91**] ainsi que celui Abu Salem **[Abu92].**
Le système IRSAD (Information Retrieval system for Arabic Documents), développé en 1993 par la compagnie ASSET (**Computer Guide, 1993**). Hmeidi **[Hme95],**
Le système AFTDB (Arabic Full Text Data Base) développé par Al Alamiyah de Sakhr Software company **(Sakhr97**), il a été essayé sur le Coran et le Hadith.

## 1.8. Quelques systèmes arabes de recherche d'information

Pour rechercher une information sur le web, l'utilisateur va soumettre une requête constituée de mots clés, cette requête est ensuite analysée pour supprimer les mots vides et éventuellement extraire la racine des mots contenus dans la requête, En Arabe, la recherche peut être basée sur le mot, la racine ou le stem le stemming (desaffixage, desuffixage) est l'action qui consiste à ôter du mot les préfixes et les suffixes.
Des études expérimentales ont été réalisées pour essayer de dégager la meilleure méthode des trois citées plus haut pour la langue arabe.
Al Kharashi décrit un système de recherche d'information où il a essayé de comparer l'utilisation des mots entiers, des stems et de la racine pour l'indexation des termes**.** Les résultats de son étude révèlent que l'utilisation de la racine et du stem donnent de meilleurs résultats que la méthode de recherche par les mots. la méthode de la racine est aussi bonne que celle du stem pour une recherche à bas niveau de rappel, mais nettement meilleure pour une recherche à haut niveau de rappel **[AlK91]**.

Morfeq a conçu El Bayan un système de recherche manipulant des textes en Arabe et en Anglais, le système est basé sur une analyse morphologique, mais il n'existe pas d'évaluation du système dans son étude**[Mor91].**

Dans son étude Abu Salem obtient des résultats conduisant à des conclusions presque similaires à celles d'Al Kharashi alors qu'il a travaillé sur le titre et le résumé au moment où Al Kharashi travaillait seulement sur le titre **[Abu92]**.

Hmeidi a aussi travaillé sur la comparaison entre les trois méthodes celle basée sur la racine, le stem et le mot. Les résultats qu'il obtient confirment ceux obtenus par Al Kharashi et Abu Salem, la racine étant meilleure que le mot **[Hme95].**
Dans sa Thèse Hasna **[Has96],** expérimente la recherche de passage de texte Arabe, Hasna conclue que la recherche de passage améliore la précision.
Nous pouvons citer comme autre travail, celui de Al Tayyar qui a mis au point **AIRSMA** (Arabic information retrieval system based on morphological analysis)**.**
Le prototype réalisé a pour but de comparer quatre méthodes de recherche d'information en arabe (la quatrième étant proposée par lui) en l'occurrence celle basée sur le mot, le stem, la racine et celle utilisant l'analyse morphologique, l'étude a été faite avec 32 requêtes sur une base contenant 590 documents dont la pertinence a été jugée par des spécialistes. L'étude a révélé que pour le rappel la méthode de l'analyse morphologique et celle de la racine sont les plus performantes notamment grâce aux variantes morphologiques. Les méthodes basées sur le stem et le mot ont donné de meilleurs résultats concernant la précision, parce qu'elles n'utilisent pas de variantes morphologiques **[AIT98]**.
Pour expliquer brièvement la différence entre ses quatre méthodes prenons par exemple le mot Arabe «المدرسون» (AL moudarrissoun=les enseignants) ce mot va être segmenté de la manière suivante :

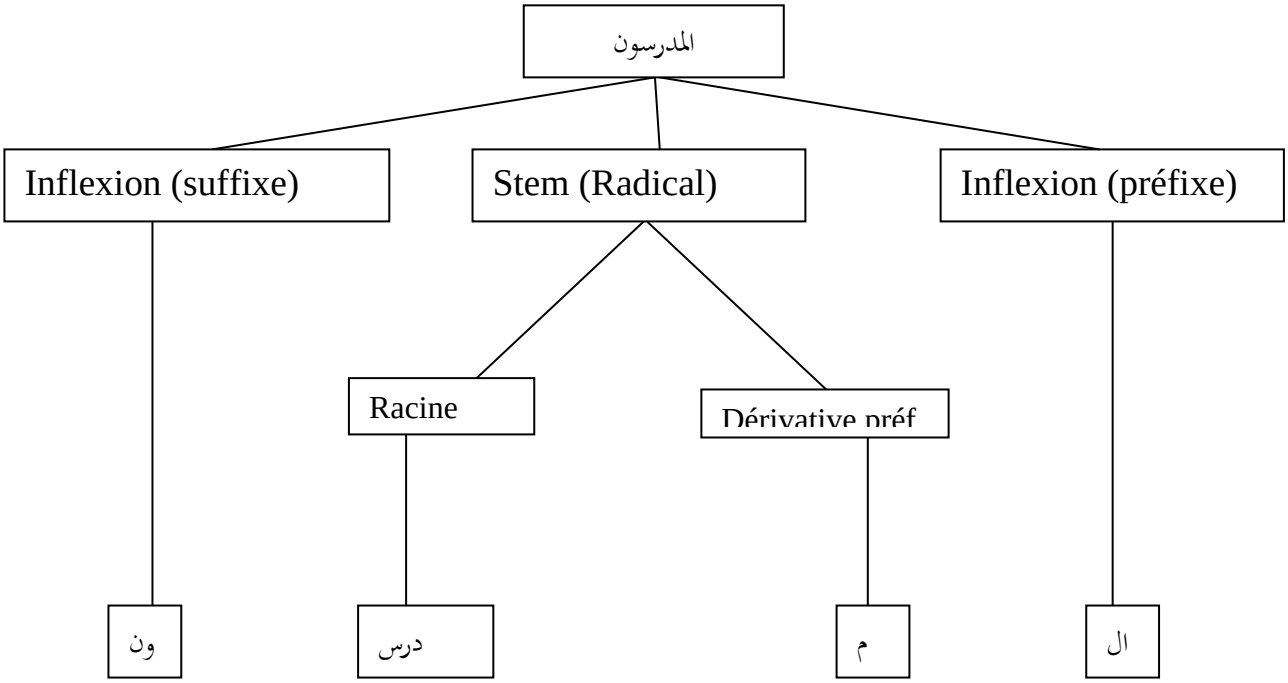
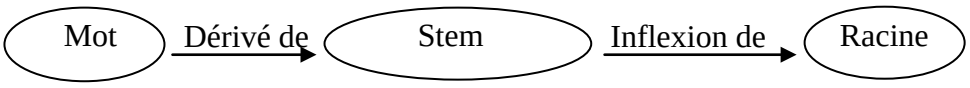



**Figure 8: structure du mot " المدرسون " (les enseignants)**



Si la requête contient le mot " المدرسون " (Al moudarrissoun) le système utilisant la méthode se basant sur « le mot » cherchera dans la base d'index " المدرسون " (Al moudarrissoun=les enseignants) alors qu'il cherchera "مدرس" (moudarris=enseignant) si la méthode est basée sur le stem, et enfin il va chercher "درس" (darasa = enseigner) avec toutes les variantes morphologiques, si la méthode est basée sur la racine.

Le principal inconvénient des méthodes basées sur la racine et l'adjonction de mots non pertinents mais partageant la même racine, par conséquent un grand nombre de documents n'ayant pas une relation avec le sens réel de la requête sera retourné ce qui générera un taux élevé de bruit.
Prenons par exemple le mot **جامع** (mosquée), **جامعة** (université), **مجموعة** (groupe ou ensemble) et **إجتماع** (réunion) tous ces mots possèdent la même racine **جمع** (réunir), d'une manière générale tous ces mots partagent le même sens de la racine (réunir), parce que les gens se réunissent dans une mosquée, dans une université, dans un groupe ou une réunion. Mais pour la recherche d'information, si nous avons besoin de réduire un mot à sa racine, toutes ces variantes ne sont pas utiles, d'après les travaux de Al Karashi **[AlK91]** et ceux de Al Tayyar **[AlT00]** ces différentes variantes du mot peuvent être utile seulement si un fort rappel est exigé mais, sans haute précision.

Pour la quatrième méthode basée sur l'analyse morphologique et afin d'extraire la racine d'un mot, nous devons connaître la forme morphologique appropriée, par exemple le mot « دراسة» DeRaSah « étude » peut être segmenté ainsi :

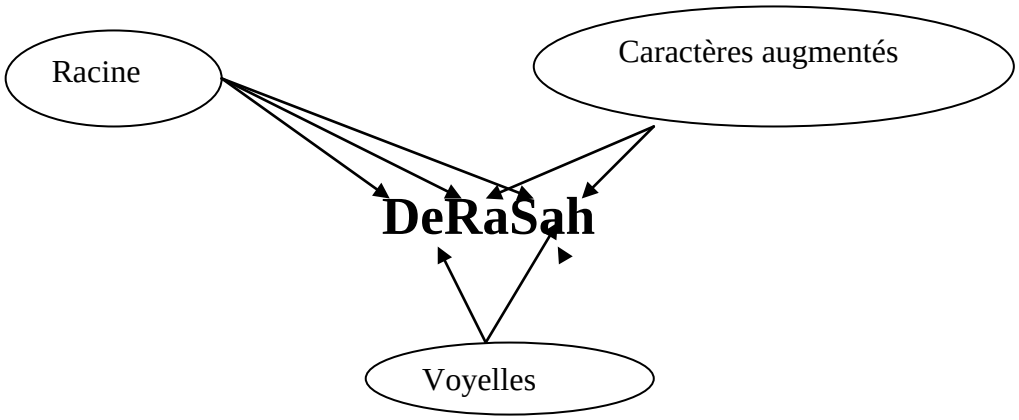
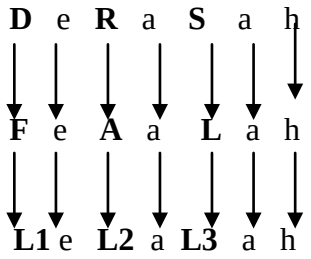
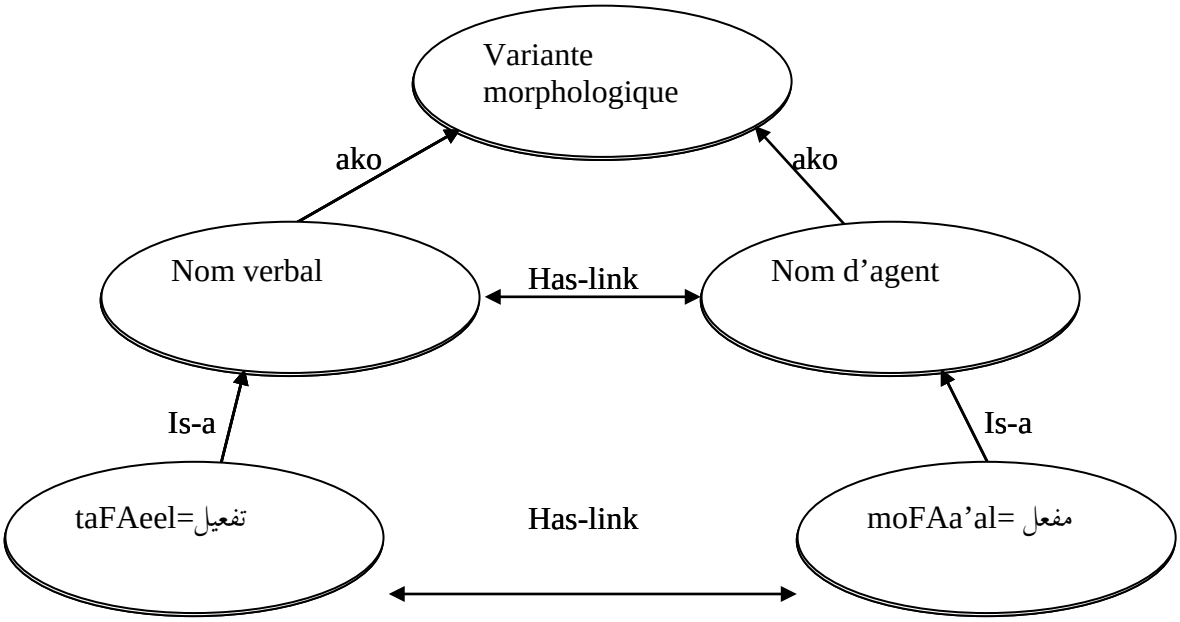



**Figure 9: segmentation du mot " دراسة" (étude)**

La forme morphologique du mot ci-dessus est "فعالة" FeAaLah




Les variables L1, L2, L3 représentent la racine de la forme morphologique, les autres caractères sont des constantes, une telle représentation permet non seulement d'analyser un mot donné, mais aussi de générer plusieurs autres mots à partir de la même racine. Mais pour bénéficier des avantages de la racine et du stem, la méthode développée par Al Tayyar **[AlT00]** consiste à lier sémantiquement, les différentes variantes morphologiques similaires, les unes aux autres, tels que les formes du nombre (singulier, dual, pluriel), le nom verbal et le nom d'agent. Un exemple de liens sémantiques entre *nom verbal* et *nom d'agent* est donné par le te schéma suivant :




**Ako** : a kind of : *est un type de*
**Is-a** : *est un exemple de*.

**Figure 10: Exemple de liens sémantiques**

Les résultats d'une telle approche selon Al Tayyar, révèlent que la méthode basée sur la forme morphologique donne une nette amélioration au niveau du rappel par rapport aux autres méthodes, quant à la précision elle est juste meilleure que la méthode basée sur la racine mais reste inférieure à celles basée sur le mot et le stem.

Nous pensons que le rappel se trouve amélioré parce qu'en plus des formes morphologiques, l'auteur a utilisé les liens sémantiques, mais la précision reste faible, justement à cause de ces variantes qui dévient, malgré les liens utilisés, la recherche.
Nous ne pouvons omettre le travail de T.Rachedi et al**.** **[Rac03]** qui travaillent depuis deux ans sur la réalisation d'un moteur de recherche multilingue distribué baptisé *Barq*, focalisé sur l'Arabe, *Barq* indexe tous les documents qui se trouvent sur le Web, incluant les documents Word et XML contenant au moins un seul mot Arabe, les documents eux-mêmes peuvent contenir des caractères latins. Les auteurs décrivent les différentes parties du système en l'occurrence le stemming, l'extraction de la racine l'indexation, la catégorisation etc…
Le traitement de la requête se fait d'abords en éliminant les mots vides, pour étendre la requête, Barq utilise trois thesaurus : le premier construit manuellement, le deuxième construit à partir de l'analyse de documents XML et le dernier construit à partir du traitement de la catégorisation des documents. En outre, le système étend la requête par l'adjonction de la racine de chaque mot de la requête, après avoir étendu la requête, un système assigne un poids à chaque mot de celle-ci pour privilégier les mots exacts de la requête sans négliger ceux ajoutés lors de l'expansion.
Cependant l'utilisateur garde le choix d'opter pour une recherche avec mots clés exacts, une recherche basée sur les concepts ou une recherche d'images.
Cette recherche est booléenne, les documents retrouvés doivent impérativement contenir au moins un des mots/mots clés ou leur extension dans la requête utilisateur.
Après le processus de recherche, le système affiche une liste de documents, ceux qui contiennent le plus grand nombre de mots clés exacts de la recherche originale se trouvent toujours en tête de liste. D'après les résultats préliminaires obtenus, le rappel se trouve amélioré de 75%.
Les auteurs ne mentionnent toutes fois pas la précision. Nous savons, comme nous l'avons cité plus haut, que l'expansion par la racine améliore effectivement le rappel mais cela se fait au détriment de la précision qui s'en trouve altérée, à cause des variantes morphologiques du mot.
C'est pour cette raison que nous avons choisi pour notre système d'éviter d'étendre la requête par la racine, mais de choisir parmi les variantes morphologiques du mot, celles qui sont très inhérentes au domaine juridique, et donc nous pouvons gagner du coté du rappel, car la requête est étendue, nous gagnerons du cote de la précision car nous n'allons pas dévier la recherche avec de nouveau mots qui peuvent n' avoir aucune relation avec le domaine cible, puisque nous choisissons justement ceux qui représentent le plus, le domaine et qui sont choisis après une étude statistique !

## 1.9. Conclusion

La recherche d'information a fait un grand pas les quarante dernières années et a permis un accès plus rapide et plus facile à l'information. Mais le chemin est encore long, cependant les méthodes statistiques ont montré leur efficacité dans ce domaine. Avec l'accroissement vertigineux du volume d'information disponible sur le Web, la recherche d'information jouera sans aucun doute un rôle capital dans l'avenir.

# 2. Les ontologies

## 2.1. Définition d'une ontologie

Le mot ontologie puise sa source dans la métaphysique où il est défini comme étant *la science de l'être en tant qu'être indépendamment de ses déterminations particulières* **[Roc03]**.
L'ontologie en ingénierie des connaissances est l'*ensemble des objets reconnus comme existant dans le domaine* **[Aus02].** Construire une ontologie c'est aussi décider de la manière d'être et d'exister des objets.
Les ontologies dans la recherche d'information peuvent être utilisées à différents niveaux:

- Représenter les textes (indexation) : Étiqueter les textes ;
- Exprimer la requête : Reformuler la requête ou la traduire ;
- Visualiser les résultats : Explorer les résultats.

Pour pouvoir faire des requêtes sur un domaine donné, il faut une conceptualisation de ce domaine. Cette conceptualisation consiste à nommer et décrire toutes les entités Qui peuvent exister dans ce domaine, ainsi que les relations existant entre ces entités. Donc cela revient à fournir tout le vocabulaire pour représenter et communiquer la connaissance de ce domaine **[Far96].**

Une ontologie décrit donc, les concepts, les attributs ou slots des concepts et les relations entre ces concepts, par exemple, la taxonomie des espèces en biologie est un type d'ontologie qui classifie tous les organismes biologiques dans des classes, ordres, familles, genres et espèces **[Das02].**

Techniquement, les ontologies sont des réseaux sémantiques comme on en connaissait voici vingt ou trente ans. La nouveauté réside dans leur échelle sans précédent (par dizaine de milliers de concepts) et dans leur utilisation pour servir de base de connaissances multilingues **[Ras04]**

Nous pouvons dire tout simplement qu'une ontologie est l'ensemble du vocabulaire d'un domaine et les relations sémantiques associées existant entre les mots de ce vocabulaire.
Mais nous ne pouvons parler d'ontologie sans passer par la première définition, qui fut donnée par Gruber en 1993 **[Grub93]** : *Une ontologie est une spécification d'une conceptualisation,* Gruber s'est intéressé à la réutilisation et au partage de connaissances entre applications, il considère que toute représentation de connaissances est basée sur une *conceptualisation.*
Une conceptualisation est l'ensemble des entités existantes dans le domaine et les relations existantes entre ces entités, cependant il ne précise pas la façon dont cette conceptualisation est obtenue **[Fur01].**

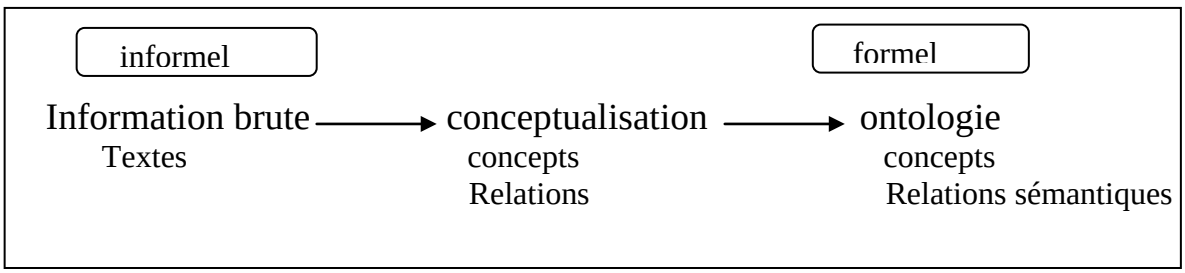



**Figure 11 : L'ontologie en tant que spécification d'une conceptualisation**

Gruber pose cinq critères pour guider la construction d'une ontologie :

1. **la clarté et l'objectivité :** les définitions doivent être indépendantes de tout choix d'implémentation ;
2. **la cohérence** ou consistance logique des axiomes ;
3. **l'extensibilité** : elle doit pouvoir être étendue sans la modifier ;
4. **la minimalité des postulats d'encodage** : garantissant une bonne portabilité ;
5. **la minimalité du vocabulaire** par le choix des termes d'expressivité maximum.

Pour mieux comprendre la définition de Gruber, Guarino explicite la notion de conceptualisation. Il la définit comme *l'identification par des termes et/ou des symboles, des concepts du domaine et des relations existantes entre ces concepts.* **[Gua96]** .

Les ontologies sont au croisement de plusieurs disciplines, en l'occurrence l'ingénierie de la connaissance, l'intelligence artificielle et les sciences de la langue.

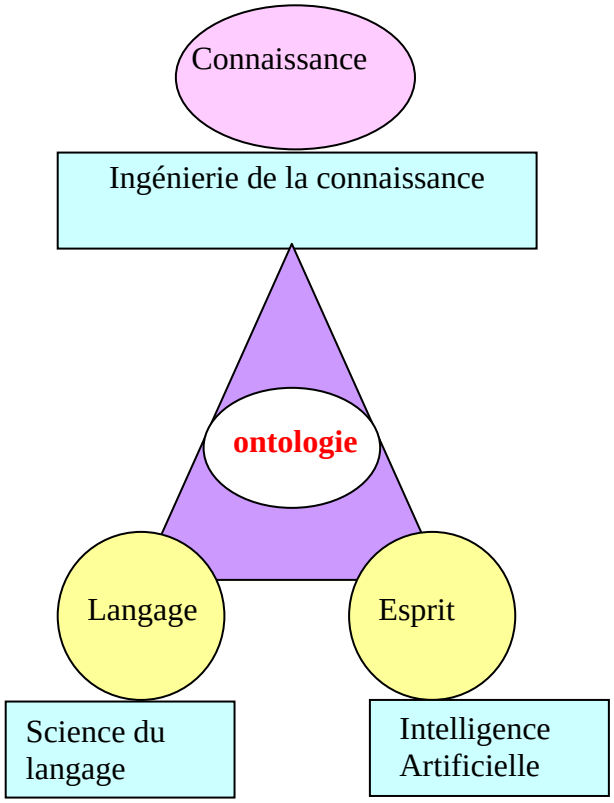



**Figure 12: L'ontologie au coeur des autres disciplines**

## 2.2. Les méthodes de construction

Une ontologie peut être construite soit :

1. manuellement ;
2. semi automatiquement ;
3. automatiquement.

**La construction manuelle** : Elle fait intervenir des experts de plusieurs disciplines, elle reste la plus fiable mais de loin la plus chère.
**La construction automatique** se fait à l'aide d'outils d'extraction de candidats termes, et de classification de ces termes. Elle se fait à partir de textes suivant l'enchaînement suivant :

Texte ⟶ normalisation ⟶ terminologie ⟶ méthode de regroupement des mots ou des termes.

Il existe de nombreux outils de catégorisation, fondés sur l'hypothèse de Harris. (qui dit que « qui se ressemble s'assemble »)

**La construction semi-automatique** quant à elle, nécessite une mixité de techniques

- techniques de NLP, linguistique et informatique
- apprentissage automatique
- intervention de l'expert.

Il existe différentes méthodologies pour la construction d'ontologies notamment :

**La méthode linguistique**
Dans cette méthode on annote les corpus, puis on apprend les annotations sur un corpus (training) et on étend à un autre corpus (test), l'expert peut largement intervenir ici.

**La méthode statistique**
Les textes sont quasiment bruts et on se sert de ressources expertes pour annoter et pour trouver des classes. L'expert ne fait pas partie de la boucle.

## 2.3. Les difficultés de construction d'une ontologie

Bien que les ontologies constituent un moyen très prometteur, pour l'amélioration de la recherche d'information, il existe cependant quelques problèmes de taille, soulevés par les ontologies du domaine, nous pouvons citer :

- **La distribution du sens** : les domaines mettant en jeu différents savoir-faire et devant partager et échanger des connaissances, nécessitent la prise en compte d'ontologies locales. Comment fondre ces ontologies locales, est un problème qui reste ouvert.
- **le partage et la réutilisation des ontologies** : l'un des objectifs majeurs des ontologies est le partage et la réutilisation. L'utilisation d'un même langage de représentation ou d'un formalisme d'échange tel que KIF (knowledge interchange format) ne permet pas de résoudre encore ce problème, bien que des efforts aient été fournis dans ce sens avec par exemple, l'outil *Chimaera*[12] utilisé dans *ontolingua* et permettant l'intégration d'ontologies existantes dans une ontologie que l'on construit pour la première fois.

[12] http://www.ksl.stanford.edu/software/chimaera/

Il existe d'autre outils, ayant le même but, tels que: *FCA-Merge* **[Gan99]** ; PROMPT[13] **[Noy00]** WebODE[14].

Toutes les applications mettent en œuvre des connaissances du domaine. Ces connaissances doivent être structurées en ontologies. L'ontologie est importante en entreprise ou dans un organisme tel que le système judiciaire, car elle peut aussi servir de référentiel métier et sera au cœur d'autres applications comme l'apprentissage des lois au grand public puisque nul n'est sensé ignorer la loi.**[Tai04]**.

## 2.3. L'utilité des ontologies

L'utilité des ontologies est capitale, dans beaucoup de disciplines, notamment dans :

- La gestion des ressources humaines : analyse automatique du contenu des CV permettant une présélection en réponse à un profil recherché.
- La gestion documentaire : classement sémantique automatisé des documents (jugements), reçus ou émis par un service (veille, documentation…).
- La recherche d'information : recherche d'information guidée par une ontologie du domaine.
- La création de bases de connaissances: notamment dans les systèmes à base de connaissances, et ce dans les différents domaines.
- La traduction automatique : pour la désambiguïsation.
- L'indexation automatique des documents : nous pouvons aller jusqu'à une indexation selon le contenu des document.

## 2.4. Outils de développement d'ontologies

Nous présentons dans le tableau ci-dessous (figure13)quelques outils de développement d'ontologies, avec leurs caractéristiques.

Notons qu'il existe des outils d'évaluation d'ontologies tels que ontoclean et One-T (Ontology Evaluation Tool).

[13] http://protege.stanford.edu/plugins/prompt/prompt.html
[14] http://delicias.dia.fi.upm.es/webODE/index.html

| Outils | caractéristiques | langage | url |
|---|---|---|---|
| Apollo | • Utilisé en industrie<br>• Supporte classes, instances, fonctions et relations | Java | http://apollo.open.ac.uk |
| LinkFactory | • Stockage en bases de données<br>• Différentes plateforme : Windows, Solaris, Unix, Linux. | Java | http://www.landc.be |
| OILEd | • Permet la construction d'ontologies avec DAML+OIL<br>• Développé à Manchester | Java | http://oiled.man.ac.uk |
| OntoEdit | Un concept peut avoir plusieurs noms (synonymes)<br><br>Version libre et professionnelle | | http://www.ontoprise.de |
| Ontolingua | • Fournit un accès à une bibliothèque d'ontologies<br>• Traduit aux langages Prolog, Corba, Clips, Loom.<br>• Développé à Stanford | Lisp | http://ontolingua.stanford.edu/ |
| OntoSaurus | • Développé en Californie<br>• Traduit en ontolingua, KIF, C++ | Loom | http://www.isi.edu/isd/ontosaurus.html |
| Protege-2000 | • Le plus récent des outils<br>• Développé à Stanford<br>• 7000utilisateurs<br>• Jusqu'à 150000 frames | c | http://protege.stanford.edu |
| WebOnto | • Developpé par KMI (knowledge Media Institute)Angleterre<br>• Bibliotheque de 100 ontologies. | OCML | http://webonto.open.ac.uk |

**Figure 13: quelques outils d'édition d'ontologies**

## 2.6. Wordnet et Eurowordnet

### 2.6.1. Généralités

WordNet[15], fut conçu à l'université de Princeton par le psychologue George Miller, c'est un dictionnaire électronique de l'anglo-américain, développé depuis 1985 et initialement conçu pour tester les déficits lexicaux dans des expériences de psychologie cognitive. Sa structure est celle d'un thésaurus. Il a été transposé à une dizaine de langues, du Basque au Bulgare. En outre, il sert d'index interlingue (*ILI* ou *Inter Lingual Index*), et donc de représentation conceptuelle indépendante des langues, dans le projet EuroWordNet[16], développé depuis 1996. Chacune des langues décrites (I'Italien, Néerlandais, Anglais, Espagnol, l'Allemand, le Français, l'Estonien, le Tchèque, etc.) développe son propre lexique à l'image de WordNet.

Depuis sa construction, Wordnet a été utilisé dans une multitude de projet, dans de différents domaines.
Dans la recherche d'information, domaine qui nous intéresse dans ce travail, nous pouvons citer quelques travaux en l'occurrence l'expansion de la requête, notamment celui de Smeaton et Berrut**[Sme95]** qui a essayé d'étendre la requête en affectant un poids aux termes rajoutés, avec des techniques de désambiguïsation manuelle et automatiques, malheureusement la stratégie a dégradé la performance de la recherche.

Toujours en 1995, Richardson **[Ric95]** se sert de Wordnet pour calculer la distance sémantique entre les concepts et utiliser ces distances pour calculer la similarité entre une requête et un document. Il propose deux méthodes pour le calcul des distances sémantiques, les deux ont amélioré la performance de la recherche **[Ric95].**

En 1997 Stairmand **[Sta97]** utilise Wordnet pour la calcul de la cohésion lexicale selon la méthode proposée par Morris **[Mor91]** et l'appliqua à la recherche d'information. Il conclu que cette méthode ne peut être appliquée à un système de recherche pleinement fonctionnel. Mais il mena une étude d'analyse cherchant à expliquer pourquoi Wordnet dégrade la performance de la recherche. Les résultats montrent que la recherche avec Wordnet améliore le Rappel mais altère la Précision. La raison en est que deux mots tels *stochastic* (adjectif) et *statistic* (nom) n'ont pas la même base dans Wordnet, il n'est pas possible de trouver une quelconque relation entre ces deux mots dans Wordnet, de plus nous ne pouvons trouver certaines relations dans Wordnet par exemple comment saurons nous que Sumitomo Bank est une compagnie Japonaise ? Enfin quelques mots ne sont pas inclus dans Wordnet notamment les noms propres.
Pour pallier à ces inconvénients, certains auteurs ont proposé d'enrichir Wordnet avec un thesaurus construit automatiquement dans le but de compléter Wordnet. Les mots polysémiques détériorent la précision de la recherche d'information.
L'idée du thesaurus basé sur la cooccurrence considère que deux mots qui cooccurrent dans un document, ont tendance à avoir une forte relation entre eux **[Man97]** et donc nous pouvons étendre la requête avec ces nouveaux mots.
Le but du projet Eurowordnet, qui a débuté en Mars 1996 et devait durer 3 ans, était de créer des Wordnets similaires mais avec d'autres langages Européens, les premiers en sont l'Allemand (à l'université d'Amsterdam), l'Italien (CNR, Pise), l'Espagnol (Fondation des Universités de

15 (http://www.cogsci.princeton.edu/~wn/)

16 **(**http://www.illc.uva.nl/EuroWordNet/)

Barcelone et Madrid), L'Anglais (Université de Sheffield USA) plus tard le Czech, l'Estonien, le Germanique et le Français seront rajoutés. Des Wordnets sont actuellement développés pour le Suédois, le Norvégien, le Danois, le Grec, le Portugais, le Basque, le Catalan, le Romain, le Lithuanien, le Russe, le Bulgare, le Slovène et l'Arabe.

Comme dans Wordnets (devenu un terme générique) chaque Eurowordnet est une hiérarchie de 1024 concepts de base, où chaque nœud est un *synset* (synonymous set ou ensemble de synonymes) un mot auquel sont associés un ou plusieurs synonymes ou phrases.
Les *synsets* sont reliés par des relations d'hyponymie(sous-concept), hyperonymie(super-concept), méronymie(une partie de) et antonymie(le contraire).
Des mots tels que *student* et *teach* sont reliés par la relation ROLE-PATIENT, la relation inverse est appelée INVOLVED-PATIENT.
L'équivalence interlingue est généralement prise comme une synonymie dénotée EQ-NEAR-SYNONYM ou lié à un concept avec EQ-HAS-HYPERONYM ou EQ-HAS-HYPONYM si c'est plus spécifique ou moins spécifique, respectivement, que le concept dans ILI.
Le développement le plus intéressant est le multilinguisme, les différents Wordnets sont reliés à une seule base d'index interlingue ILI (Inter-Lingual-Index) basée sur Wordnet à Princeton, via cet index les langages sont interconnectés ce qui rend possible le passage d'un mot dans un langage au mot équivalent dans n'importe quel autre langage **[Vos98], [Fel98],** et donc peuvent aussi bien être utilisés dans la recherche d'information que dans les applications de traduction automatique.

Pour le moment l'accent est mis sur les noms et les verbes. Chaque nom ou verbe est représenté par un *frame*, regroupant tous les mots de même sens. Le but principal des bases de données Eurowordnets est l'expansion de la requête dans la recherche d'information interlingue.

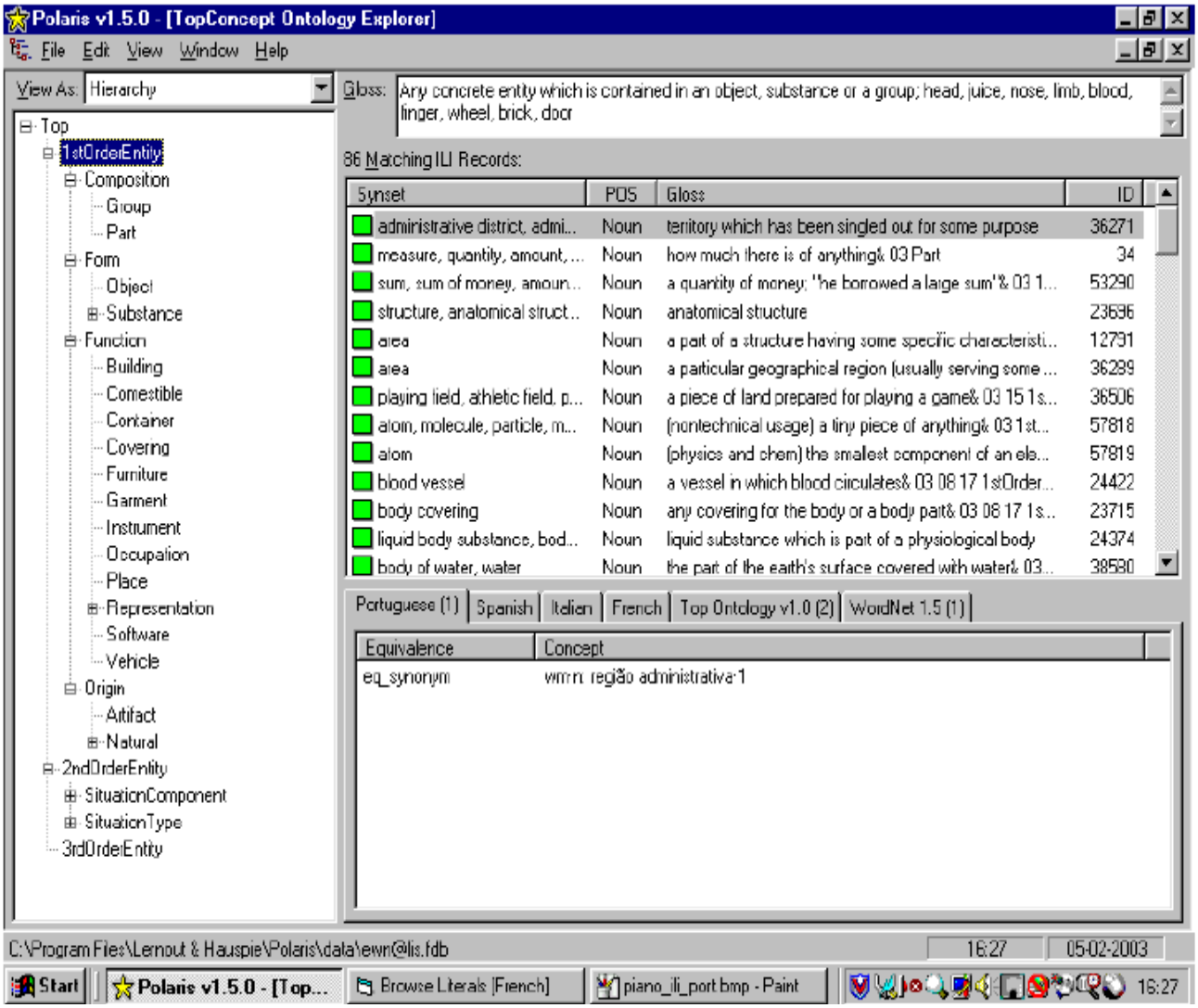



**Figure 14: Architecture générale de l'ontologie EuroWordnet**

## 2.6.2. Les relations dans Wordnet et EuroWordnet

Nous allons expliciter les principales relations utilisées dans Wordnet et EuroWordnet

**1) Hyponymie** : c'est la relation entre un terme spécifique et un terme générique exprimée par l'expression « is-a »

**Dans Wordnet**

L'hyponymie est exprimée par les relations « is-a » ou « is-a-kind-of ».

**Dans EuroWordnet**

La relation d'hyponymie est exprimée par « has-hyponym » et la relation réciproque « has-hyperonym »

**2) Méronymie** : c'est la relation exprimée par « est-une-partie-de »

Sa relation réciproque est l'Holonymie si un concept C1 est un méronyme d'un concept C alors C est un Holonyme de C1.Winston a distingué six types de méronymie **[Win87].**

| **Types de méronymie** | **Exemple** |
|---|---|
| Composant -objet complet | Pédalier/bicyclette |
| Membre -collection | Arbre/forêt |
| Portion –masse | Tranche de pain/pain |
| Matière –objet | Métal/voiture |
| Sous activité –activité | Payer/acheter |
| Lieu -Zone | Oasis/désert |

**Figure 15: Les différents types de Méronymie**

Il a définit en plus trois critères afin de différencier entre ces types de relations :
_ La relations entre la partie et le tout, est fonctionnelle ou pas,
_ Les parties et le tout sont homéomères ou pas (c à d ils sont du même type ou nature que le tout ou pas),
_ La partie et le tout sont séparables ou pas.

***Dans Wordnet***
La relation de méronymie est représentée par la relation « meronym » et elle est interprétée de trois manières différentes :
X est un méronyme de Y Si :

- X is -a -part of Y (Composant - objet complet),
- X is -a -substance of Y (Matière – objet),
- X is -a -member of Y (Membre – collection).

***Dans EuroWordnet***
Les auteurs ont développé plus en détail la relation de méronymie et sa relation réciproque et ils ont représenté les six types de la relation de méronymie qui ont été évoqués précédemment.

| **Les différents types de la relation de méronymie** | **Relation réciproque** |
|---|---|
| HAS_MERONYM | HAS_HOLONYM |
| HAS_MERO_LOCATION | HAS_HOLO_LOCATION |
| HAS_MERO_MADEOF | HAS_HOLO_MADEOF |
| HAS_MERO_MEMBER | HAS_HOLO_MEMBER |
| HAS_MERO_PART | HAS_HOLO_PART |
| HAS_MERO_PORTION | HAS_HOLO_PORTION |

**Figure 16: Les types de Méronymie dans EuroWordnet**

**3) La synonymie :**

Le synonyme est un mot qu'on peut substituer à un autre sans changer le sens.

***Dans Wordnet***

La relation de synonymie est à la base de la structure de Wordnet, les unités lexicales sont regroupées en ensemble de synonymes « synsets ». Un synset regroupe tous les termes utilisés pour dénoter un concept.

Exemple : [person, individual, someone, somebody, mortal, human, soul]

***Dans EuroWordNet***

Il y a l'ajout de la relation *NEAR_SYNONYM* pour les concepts qui ne sont pas identiques.

***4) Antonymie***

Ce sont des couples de mots dont la négation de l'un implique l'affirmation de l'autre.

Exemple : Pair / Impair, marié / célibataire.

**Dans Wordnet**

C'est une relation lexicale entre deux mots et non une relation sémantique entre deux synsets. Elle représente la complémentarité.

Exemple : chance/ malchance, riche/pauvre.

**Dans EuroWordnet**

Il y a l'ajout de *NEAR_ANTONYM* qui permet de définir les quasi-antonymes afin de compenser le degré d'antonymie d'une langue à une autre.

Exemple : construire et détruire sont des antonymes, construire et destruction sont des quasi_antonymes.

## 2.7. L'ontologie SENSUS

SENSUS est une ontologie de concepts constituée de 70,000 noeuds reliés entre eux par des Relations d'hyponymie (is-a) et des relations de méronymie (part-of) essentiellement.

SENSUS est la fusion de WordNet et upper model de Penman **[Bat90].**

SENSUS est considéré comme une bonne ontologie pour démarrer (de base) car :

- Elle contient un très grand nombre de termes ;

Les termes couvrent la plus part des secteurs généraux des expériences humaine ;

- Elle adopte une notation simple et facile à lire.

On peut explorer SENSUS[17] avec le browser ONTOSAURUS[18] qui a été construit à ISI par Ramesh Patil et Tom Russ.

## 2.8. Conclusion

Beaucoup de définition ont été données au mot ontologie, autant de définition que de méthodologie de construction, *ce qui est important c'est le but pour lequel l'ontologie est faite.* L'ontologie est donc, définie pour un objectif donné et exprime un point de vue partagé par une communauté. La recherche, le partage, et la réutilisation de ressources du Web (ou entrepôts) sont les principaux usages attendus du Web dans beaucoup de domaines tel que le domaine juridique. Pour répondre à ces besoins il faudra construire des ontologies *partagées*, *conformes aux standards* et à l'expertise du domaine, *évolutives,* aisément enrichissables et modifiables *réutilisables,* et faciles à utiliser. De nombreuses questions qui restent encore largement ouvertes, quant à la compatibilité des ontologies, entre elles, et avec les standards existants.

[17] (http://mozart.isi.edu:8003/sensus2/)

[18] (http:// www.isi.edu/isd/ontosaurus)

# 3. Le système proposé

## 3.1. Introduction

Avant de penser à une solution nous nous sommes posées les questions suivantes :

1. Qui va utiliser ce système ? Dans quel but ?
2. Comment ces utilisateurs potentiels ont-ils l'habitude d'effectuer leurs recherches sur le Web ?
3. Que va apporter ce système de nouveau ? en d'autres mots à quelles questions va-t-il répondre ?
4. Quelles sont les attentes et les aspirations des utilisateurs potentiels ?
5. Dans le cas où ce système serait implémenté et opérationnel, qui assurera sa maintenance ?
6. Est-ce que la requête est un ensemble de mot-clés ou des phrases en langage naturel ?
7. Si la requête est en langage naturel quels sont les meilleurs outils d'analyse à utiliser.
8. Est-ce qu'on réalise ces outils ou est-il préférable d'utiliser ceux déjà existants (pour ne pas réinventer la roue).
9. Si la requête est un ensemble de mot-clés, est-il préférable d'extraire la racine du mot avant le processus de recherche ou après, pour l'expansion de la requête.
10. comment doit-on éliminer les mots insignifiants (ou mots vides) comme les articles, les prépositions etc.
11. Comment va-t-on étendre la requête pour améliorer la précision du résultat sans altérer le rappel ?

Concernant le premier point, le système pourrait être utilisé par une large communauté composée de juristes, avocats, juges, enseignants chercheurs et étudiants dans la recherche d'information spécifique à leur domaine sur le web.

Pour le deuxième point, nous avons essayé d'établir un questionnaire et nous avons invité quelques juristes (avocats, enseignants, étudiants..) à y répondre, pour avoir une idée sur la manière avec laquelle ils utilisent l'outil Internet. (voir Annexe A).
Les résultats obtenus après dépouillement, ont montré que la majorité n'utilisent pas encore l'outil informatique, à l'exception des chercheurs postulant pour des mémoires de Magister ou des theses de doctorat, pour les utilisateurs occasionnels, ils sont plutôt satisfaits des services offerts par Internet, puisque leur utilisation se limitent à l'utilisation des emails.
Les réponses recueillies auprès des chercheurs pour leurs projets, ont laissé comprendre qu'ils aspirent à plus de facilité dans la recherche et à de meilleurs résultats fournis par les systèmes outils spécifiques.

Pour répondre à la troisième question, nous espérons qu'une fois le système mis au point, il puisse rendre compte des différentes modifications et mises à jour apportées aux différentes lois.

Concernant le cinquième point, nous pouvons imaginer une autorité telle que le ministère de la justice ou un centre informatique juridique spécialisé, qui aurait pour fonction la formalisation de la loi, dans le but de sa vulgarisation.
Pour le sixièmement, nous avons choisi de traiter une requête avec des mot-clés, d'une part parce que selon les statistiques faites sur Internet, les utilisateurs soumettent le plus souvent des requêtes de

deux mots, toutes les études effectuées montrent que les requêtes formulées sont en moyenne inférieures à deux mots **[Jen00].**
D'autre part tous les systèmes de recherche utilisant le langage naturel donnent encore des résultats peu performants dus aux difficultés et à l'ambiguïté du langage naturel **[Str97].**

Concernant le huitième point, bien que la requête ne traite pas du langage naturel puisqu'il s'agit de mots clés, mais l'utilisation d'outils existant pourraient s'avérer nécessaire tels que le système de recherche, l'analyseur morphologique, l'extracteur de racine etc. notre étude se veut focalisée sur la construction de l'ontologie et son exploitation dans l'expansion de la requête et donc pour ne pas se perdre dans un travail trop large pour être cerné dans un mémoire de Magister, nous ne sommes pas attardés sur ces points.

Le problème de l'extraction de la racine se pose avec une certaine acuité, en effet il est presque impératif de représenter toutes les variantes par un même mot, d'un autre coté inclure toutes ces variantes dans la requête peut induire la recherche en erreur et dévier ainsi les résultats obtenus **[AlT00] [AlK95].**
Dans notre ontologie, nous avons préférés attacher à chaque concept, dans la mesure du possible leurs définitions et les variantes qui sont fortement liées au domaine juridique. Nous ne parlons pas de l'extraction de la racine lors de la soumission de la requête au système de recherche, puisque nous utilisons les moteurs de recherche disponibles sur le Web et qui supportent la langue Arabe.

Pour le dixième point, les mots insignifiants sont généralement regroupés dans une liste prédéfinie, mais ce point aussi ne nous concerne pas ici, puisque nous utilisons des moteurs de recherche existants, ce traitement est interne au moteur utilisé.

Pour étendre la requête nous avons choisi l'utilisation d'une ontologie en Arabe dans le domaine juridique, d'un coté parce que les ontologies en langue Arabe sont presque inexistantes d'un autre coté les ontologies sont un moyen très efficace pour l'expansion de la requête afin d'améliorer le Rappel et la Précision en réduisant le silence et le bruit, respectivement. C'est une technique qui a fait ses preuves dans multiple domaine, nous pouvons citer en particulier l'ontologie générique Wordnet **[Mil90], [Vos99], [Gat00],** ainsi que Circa (CIRCA est l'acronyme de « Conceptual Information Retrieval and Communication Architecture » (architecture d'extraction d'information conceptuelle et de communication) **[Car04].**

## 3.2. La solution proposée

La recherche simple avec une liste de mots clés, est une description de la requête qui ne dit rien sur les relations sémantiques entre ces mots, nous pouvons facilement avoir un synonyme valide d'un terme de la requête dans un document, qui ne sera jamais retourné, et qui peut par conséquent conduire à un échec de la recherche. Le mot dans ce type de recherche est une simple séquence de code binaire représentant ce mot **[Haa01]**.
L'utilisation d'une ontologie, comme moyen de reformulation de la requête va pourvoir ces mots d'un certain sens, dû essentiellement aux différentes relations entre concepts.

### 3.2.1 Le choix du domaine

Pourquoi le domaine juridique ?

Une grande communauté est liée d'une façon directe ou indirecte à ce domaine, notamment des juges, des procureurs, des avocats, des notaires, des huissiers de justices, des traducteurs, des greffiers, des commissaires priseurs, des experts, des enseignants chercheurs, des étudiants etc.

C'est le seul domaine, à notre connaissance, où la langue Arabe est officiellement la seule à être utilisée pour l'émission de jugements, de décisions, ou de tout autre document.

De plus dans le projet de la reforme du système judiciaire, on préconise l'utilisation d'un réseau intranet pour faciliter l'extrait du casier judiciaire, et certains autres documents, ainsi que la construction d'un site Internet qui doit aider à se mettre à jour, avec les changements économiques, sociopolitiques, culturels et afin de faire face à la mondialisation et son impact sur la justice notamment concernant l'adhésion de l'Algérie à l'organisation mondiale du commerce l'OMC et le traité de coopération avec l'union européenne l'UE, le respect des engagements envers d'autres pays, la coopération avec des organisations mondiales et régionales, la coopération avec les ONG (organisations non gouvernementales) et la coopération avec la croix rouge.

Alors n'est-il pas nécessaire de conceptualiser le droit puisque nul n'est censé ignorer la loi !

### 3.2.2 La construction de l'ontologie

**a)Choix de l'outil**

Avant de commencer la construction de l'ontologie, nous avons tenu à effectuer une petite étude comparative pour la sélection du meilleur outil pour l'édition de l'ontologie.

Notre premier choix c'est d'abord porté sur Ontolingua vus les différents outils mis à la disposition de l'utilisateur notamment pour l'intégration d'autres ontologies existantes et disponible sur le site Web.

Deux inconvénients ont fait que notre choix se reporte sur un autre outil qu'est Protégé2000. Le premier étant que le logiciel ne peut être télécharger, on ne peut qu'y travailler constamment et exclusivement en ligne, le second inconvénient, et il était de taille, c'est que Ontolingua ne supporte pas la langue Arabe.

Protégé2000 est un outil de développement d'ontologies et d'acquisition de connaissances. Développé par SMI à l'université de Stanford, s'exécute sur n'importe quelle plate-forme qui supporte la version 1.3 de JDK.

L'éditeur d'ontologies Protégé2000 est un éditeur qui permet de construire une ontologie pour un domaine donné et d'acquérir des connaissances sous forme d'instances de cette ontologie.

Protégé est aussi une librairie JAVA qui peut être étendue pour créer de véritables applications à bases de connaissances en utilisant un moteur d'inférence pour raisonner et déduire de nouveaux faits par application de règles d'inférence aux instances de l'ontologie .

Dans le contexte du web sémantique des « plugin » pour les langages RDF, DAML+OIL, et OWL ont été développés pour Protégé. Ces « plugin » permettent d'utiliser Protégé comme éditeur d'ontologies pour ces différents langages, de créer des instances et les sauvegarder dans les formats respectifs.

Il est également possible de raisonner sur les ontologies en utilisant un moteur d'inférence général tel que JESS (Java Expert System Shell).

**b) Module de connaissance dePprotégé2000**

Le module de connaissance de Protégé2000 est à base de frame, Les frames sont les principaux modules de la base de connaissance.

Une ontologie de protégé se compose de classes, slots, facettes et d'axiomes.

- Les classes : définissent les concepts du domaine.
- Les slots : définissent les propriétés des classes.
- Les facettes : définissent les propriétés des slots.
- Les axiomes : ce sont des contraintes additionnelles (Exemple : une valeur appartenant à un intervalle).

**c) La construction de l'ontologie**

Nous avons essayé de développer une ontologie, en suivant la démarche proposée par N. Noy **[Noy00b].** Pour construire une ontologie nous devons passer par sept étapes essentielles :

**Etape1*:*** elle consiste à se poser quatre questions capitales (et bien sûr à y répondre !).

**1) Quel est le domaine couvert?**

Pour notre cas c'est le système judiciaire Arabo-Musulman en général et Algérien en particulier.

**2) Quel est le but de l'utilisation de cette ontologie?**

Pour faciliter la recherche d'information en langue Arabe dans le domaine juridique, sur le Web.

**3) À quelles types de questions l'ontologie devra t-elle fournir des réponses** ?

Comme il est important de limiter un peu le domaine et qu'il y a des processus qui sont très longs et énormément documentés, nous allons commencer par les concepts les plus génériques, une perspective serait éventuellement d'étendre une ontologie à tout le processus, par exemple: du rapport de police à la condamnation par un tribunal et le règlement du litige (prison, amende, etc), mais le nombre d'artéfacts crées par une telle chaîne d'événements légaux est interminable et plusieurs groupes de recherche en informatique juridique s'intéresse à ce domaine dont on fait souvent référence par les termes «eLaw» ou «eJustice», bien que la création d'une ontologie publique n'est pas publicisé.

Il est plutôt difficile d'imaginer une Ontologie universelle puisque le droit est a priori culturel, voire même dans certains pays, religieux. Il faut aussi déterminer quelle genre d'interrogation sont utiles une fois l'ensemble de la hiérarchie modélisé. Ces question tourneraient autour de deux ordres: l'état de la loi à tout moment et donc son historique, ainsi que l'organisation des différentes règles et leur champ d'«applicabilité» par exemple : Ce jugement est-il toujours valide, ou en vigueur? Renversé par une cour supérieure? Contredit par un autre jugement ultérieur?

**4) Qui utilisera et maintiendra l'ontologie?**

On pourrait imaginer une autorité centrale tel que le ministère de la justice ou encore par un centre de recherche sur l'informatique juridique.

Puisque l'ontologie a pour but l'amélioration de la recherche d'information il est très important d'inclure les synonymes de chaque concept, mais comme protégé2000 ne supporte pas les synonymes (il n'y a pas de champs spécialement pour les synonymes, comme pour d'autres outils), nous avons penser au début, à utiliser des méta-classes ou les sous classes seront tous les synonymes de cette métaclasse. Mais réflexion faite et après quelques exemples, nous avons pu remarquer que l'ontologie devenait trop lourde à gérer, ce qui va à l'encontre à sa propriété principale d'être

aisément modifiable et facile à gérer. Alors nous avons penser à insérer ces synonymes dans le champs *documentation,* ce qui se fait normalement même dans d'autres outils d'édition d'ontologie.
Nous devions établir une liste de questions qu'on appelle des questions de compétence, qui sont juste une ébauche et serviront de test plus tard, la question que nous devions nous poser est la suivante: l'ontologie contient-elle suffisamment d'information pour répondre à ces type de requêtes. Voici un exemple de ce genre de question.
L'ontologie va t-elle pouvoir nous informer des lois en vigueur, de celles abolies ou corrigées par un quelconque décret présidentiel, ou une loi exceptionnelle ?
Ces questions sont à priori vagues, mais le raffinement successif saura peut-être les préciser ou aider à y répondre.

**Etape2 : Envisager une éventuelle réutilisation des ontologies existantes**
Pour notre cas il n'existe pas encore d'ontologie en Arabe dans le domaine juridique. En Anglais il y a eu quelques travaux qui avaient surtout comme objectif d'organiser les différents domaines, concepts du Droit. Par exemple: le droit criminel, le droit des finances. Il y a aussi le travail que fait actuellement Guillaume Blain sur le droit Québécois **[Bla].**
Pour notre travail, nous supposons donc qu'il n'existe pas d'ontologies appropriées et nous commençons le développement de l'ontologie à zéro.

**Étape 3 : Énumérer les termes importants dans l'ontologie**

Il serait intéressant d'établir sous forme de liste tous les termes à traiter et de ne pas se soucier de l'éventuel chevauchement entre les concepts ni des relations qui peuvent exister entre ces termes ni s'ils sont des classes, attributs(slots) ou facettes, puis établir les propriétés de ces termes. Une liste préliminaire, dont les termes ont été extraits d'articles publiés sur Internet par l'ONU. La liste fut ensuite enrichie au fur et à mesure de l'avancement de la construction de l'ontologie.

**Étape 4 : définir les classes et la hiérarchie des classes**

Nous commençons la hiérarchie par des concepts dits « top-types », ils ne sont couverts par aucun autre concept du domaine en général, mais dans Protégé ce sont des sous classes de la classe système THING. Chacun de ces concepts constitue une hiérarchie à lui seul.

Les différents concepts sont reliés par (**is-a**) qui détermine une relation d'hyponymie et d'hyperonymie. Et la relation « **a –part- of** » (Est- une -partie -de)

Exemple1 : «المحكمة_الابتدائية» est un hyponyme de «قضاء_عادي »

«قضاء_عادي » est un hyperonyme «المحكمة_الابتدائية»

Exemple2 : « يسلم_ الأوسمة_ الشرفية » est- une –partie- de «مهام_رئيس_الجمهورية»

Nous avons commencé notre construction de la hiérarchie, en usant avant tout, de noms simples ou syntagmes nominaux tels que "المحاكم الابتدائية" ou "المحاكم" . Nous avons utilisé quelque fois des syntagmes verbaux comme : « يسلم_ الأوسمة_ الشرفية » qui fait partie des fonctions du président de la république. Une étude future prendra en charge les verbes avec leurs différentes classes ainsi que le schéma thématique de chaque classe.

La stratégie de construction utilisée est le top-down ou (de haut en bas) : nous avons commencé par les concepts les plus généraux et nous poursuivons par la spécialisation des concepts. Par exemple nous commençons par «النظام_القضائي_الجزائري », Nous spécialisons par «القضاء_العادي » Puis «المحاكم_الابتدائية». Puis «القسم_المدني» puis « الدعاوى_المدنية » etc.

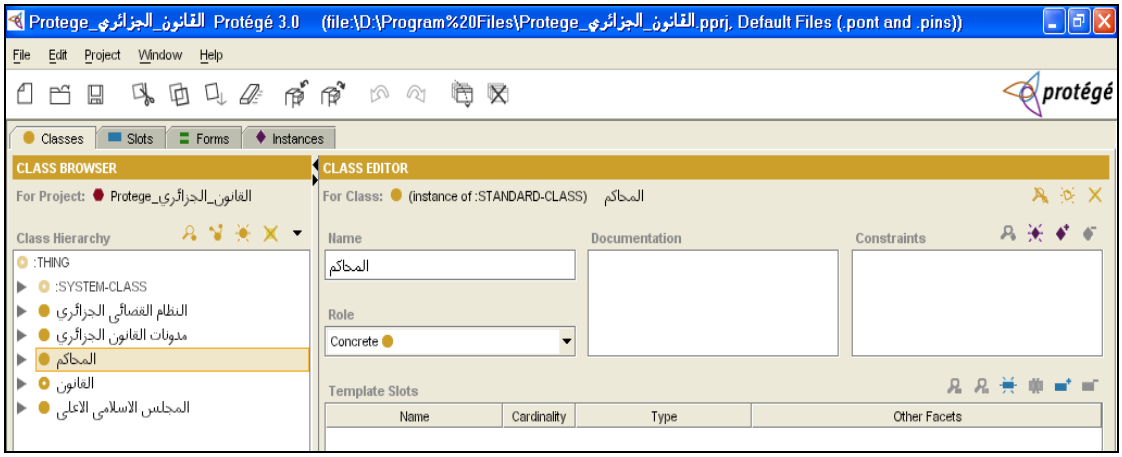



**Figure 17: les "top-types" de l'ontologie**

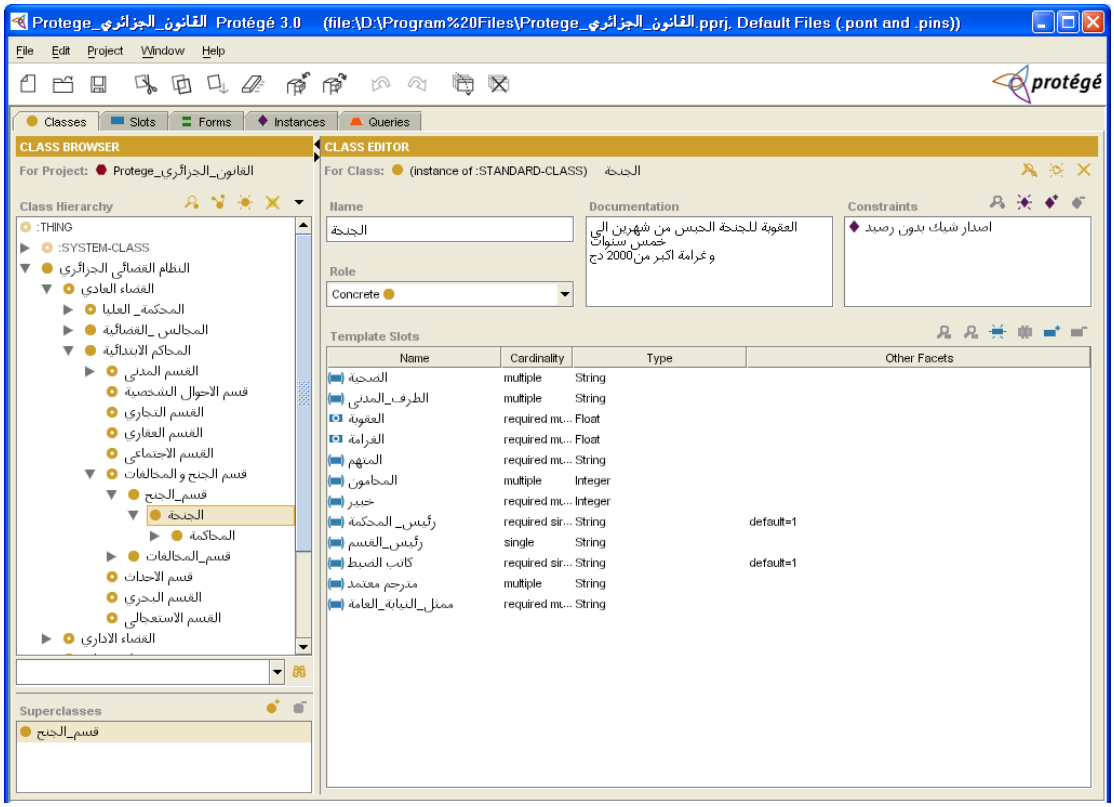



**Figure 18: la relation d'hyponymie dans l'ontologie**

**Étape 5 : Définir les propriétés des classes (attributs, slots)**

Les classes seules ne fournissent pas assez d'information, après avoir défini quelques classes nous devons décrire les propriétés de ces classes. Des parties (physiques ou abstraites) si l'objet est structuré ou en relation avec d'autres objets. Toutes les sous-classes héritent les attributs de cette classe.

Il faut noter que nous avons rencontré un problème de construction concernant l'héritage des slots. Dans Protégé, les sous classes héritent toutes les propriétés de leur super classe, cependant, il arrive qu'une sous classe ne doit pas hériter une propriété de son hyper-classe. Considérons l'exemple suivant :

« المجالس القضائية » a parmi ses propriétés « رئيس–المجلس », « غرفة_الأحوال_الشخصية » est une sous classe de « المجالس القضائية » mais ne doit pas avoir « رئيس–المجلس » comme propriété, puisqu'elle a un slot « رئيس_الغرفة». Nous n'avons pu résoudre ce problème, nous avons accepté l'attribut « رئيس–المجلس », en continuant à chercher une éventuelle solution au problème.

Les étapes 4 et 5 sont généralement entrelacées.

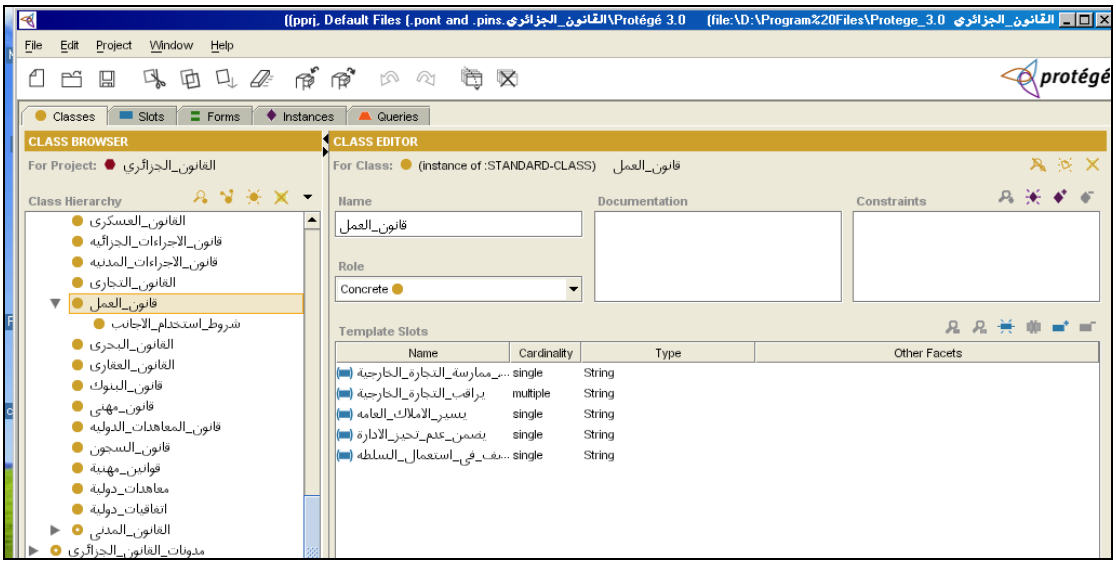



**Figure 19: les propriétés des classes.**

**Étape 6 : définir les facettes des attributs**

Les attributs ou slots peuvent avoir plusieurs facettes décrivant la valeur du type ou les valeurs autorisées, la cardinalité (nombre de valeurs) etc.

Exemple : la valeur de l'attribut : *Nom* est *string* (une chaîne de caractères). La classe: «جنحة » a pour attribut « غرامة_مالية»، qui a pour facette « دج 2000 أكبر من » (supérieur ou égal à 2000 DA) Alors que la facette de l'attribut « عقوبة_الحبس » de la classe « مخالفة » est (entre un jour et deux mois).

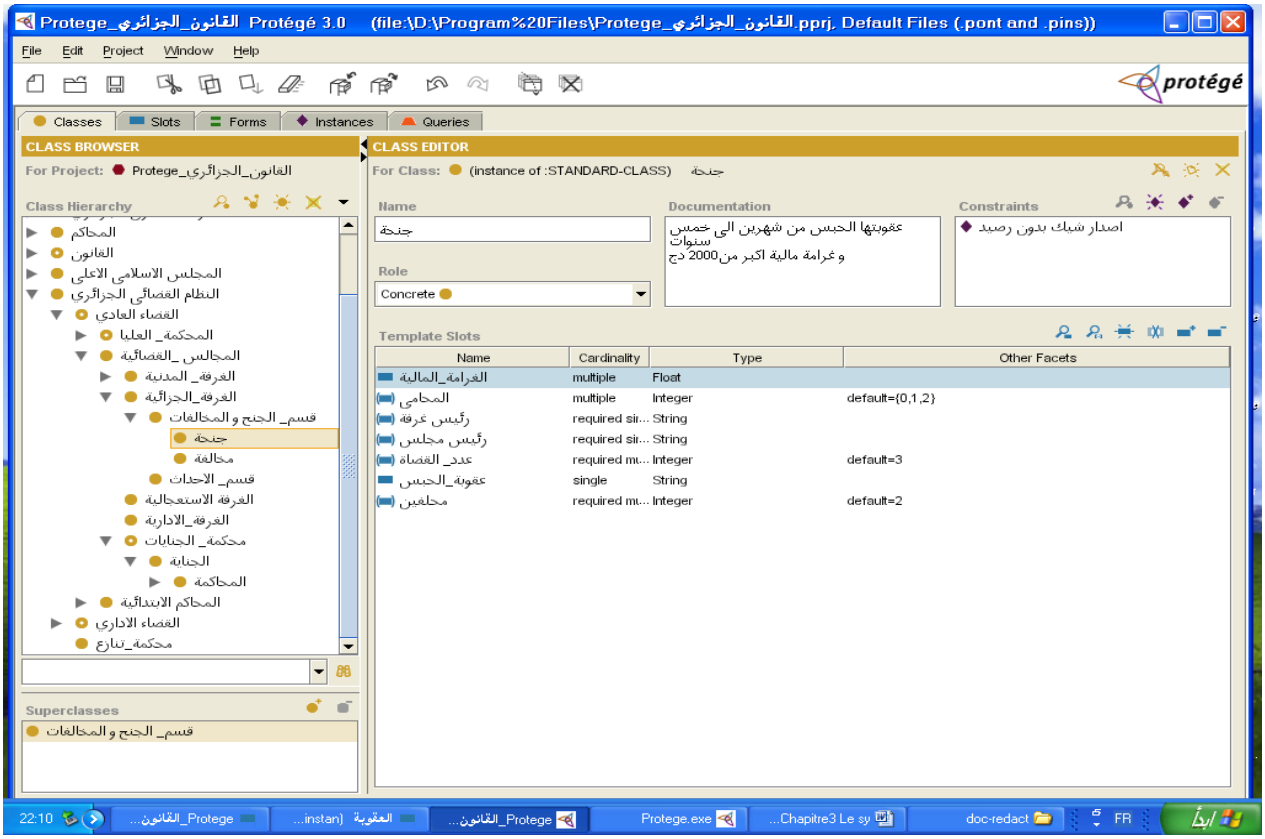



**Figure 20: Les facettes de l'attribut "الغرامة_المالية "**

***La cardinalité* :** elle définit le nombre de valeur d'un attribut, elle peut être unique ou multiple. Elle peut aussi être minimale ou maximale.

***Le type de valeur des attributs* :** elle décrit les types de valeurs pouvant être affectées à l'attribut comme : integer pour « عدد_القضاة ».

***String*** *:* la valeur est une simple chaîne de caractères (Exemple : « اسم_القاضي » Nom du juge ).

***Float, Integer*** : décrit les attributs ayant une valeur numérique Exemple « الغرامة_المالية ».

***Booleen* :** de type vrai ou faux, Exemple « جنسية_الزوجة_جزائرية » comme condition pour être président de la république.

***Énuméré*** : on spécifie une liste de valeurs énumérées (samedi, dimanche ….).

***Instance* :** permettent la définition des relations entre individus. Elle impose la définition d'une liste de classes autorisées desquelles les instances sont issues.

***Le domaine* :** Les classes auxquelles un attribut est rattaché ou les classes dont l'attribut décrit les propriétés, sont appelées le domaine d'un attribut.

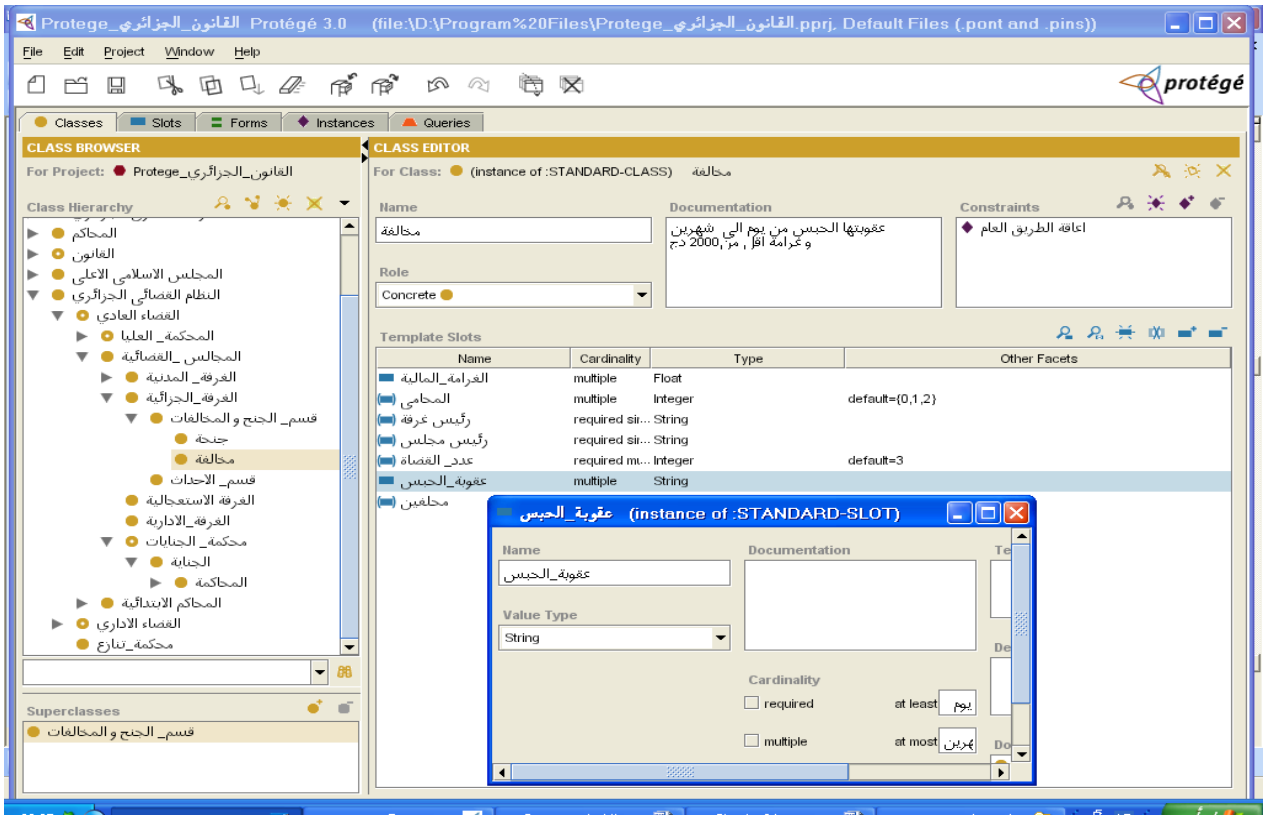



**Figure 21: Les facettes de l'attribut " عقوبة_الحبس"**

### Étape 7 : créer les instances

Cette étape consiste à créer des instances pour les classes de la hiérarchie, cela revient à choisir une classe, créer une instance pour cette classe, définir des attributs et leurs valeurs.

Par exemple :
Pour la classe « المحاكم_الابتدائية » (Tribunaux), nous pouvons créer l'instance "محكمة_ عنابة" (Tribunal d'Annaba) « محكمة_الذرعان » (Tribunal de Drean), « محكمة_الحجار » (Tribunal d'El-Hadjar), « محكمة_برحال » (Tribunal de Berrahal).

Nous définissons aussi les attributs du concept « المحاكم_الابتدائية» tels que:

« المدعي» (le plaideur)

«المدعى_عليه » (L'accusé, L'inculpé)

«رئيس_المجلس» (président du tribunal)

«المحامي » (avocat)

«خبير » (expert)

«النيابة_العامة» (représenté par le procureur de la république)

«كاتب_الضبط » (Greffier)

«عدد_القضاة» (nombre de juges).

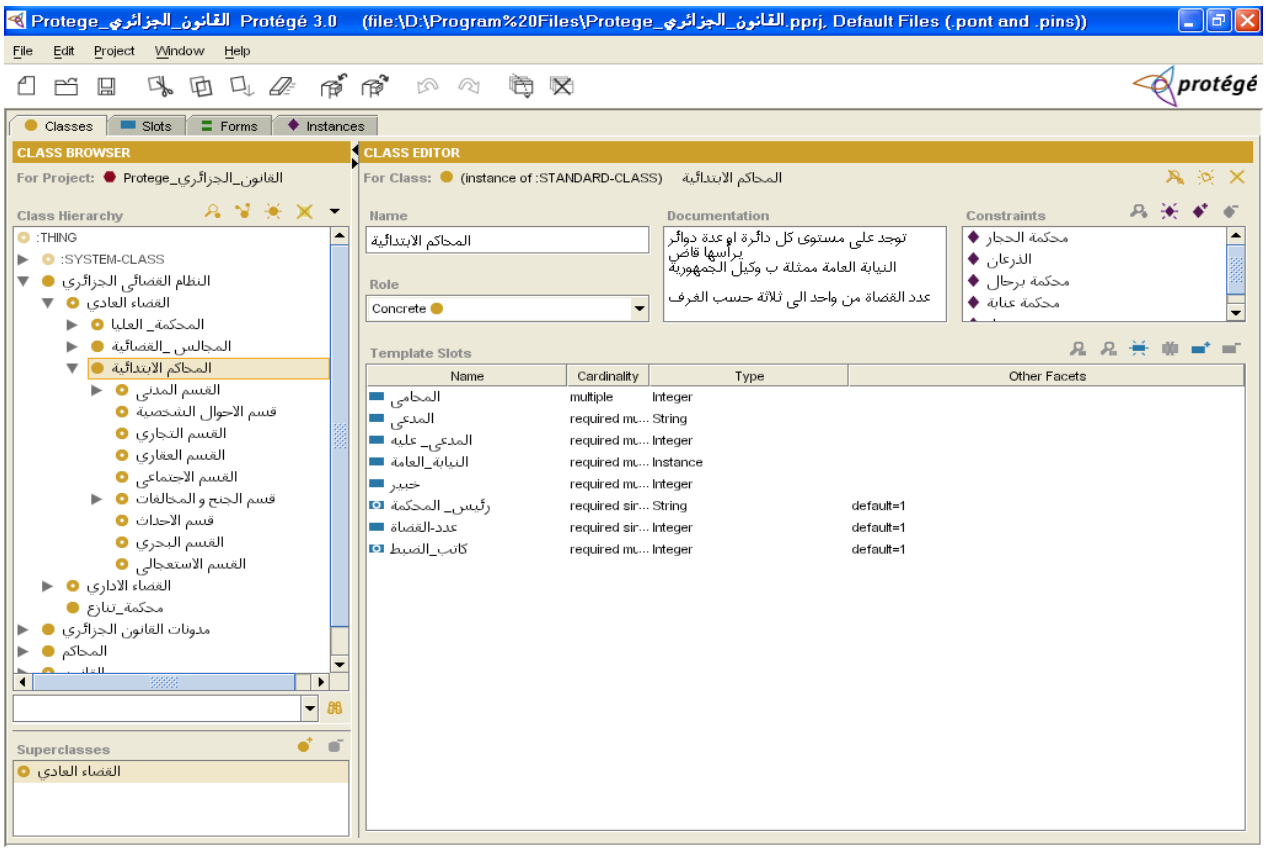



**Figure 22: Les instances de l'ontologie**

### 3.2.3 Cohérence de l'ontologie

Notons qu'il n'existe pas une seule hiérarchie correcte pour un domaine, puisque celle-ci dépend des utilisations possibles, du niveau de détail nécessaire pour l'application et même des préférences personnelles, de plus les points de vues de différents experts peuvent ne pas se rencontrer, dans notre cas l'ontologie est vouée à la recherche d'information inhérente au droit notamment Algérien en langue Arabe sur le web. Mais quelque soit la situation il faut respecter certaines règles en l'occurrence :

1. *S'assurer que la hiérarchie des classes est correcte* :

La hiérarchie des classes représente une relation de type « is-a » une classe B est une sous-classe d'une classe A si toute instance de B est également une instance de A.

Par exemple " المجالس_القضائية" est une sous classe de "القضاء_العادي" si "محكمة عنابة" est une instance de "المجالس_القضائية " alors elle est aussi une instance de " القضاء_العادي ".

Les synonymes pour le même concept ne représentent pas des classes différentes. Pour des contraintes de représentation dans protégé 2000 un concept est une métaclasse ses synonymes sont des sous métaclasses. Pour cause de lourdeur et de complexité de la hiérarchie et donc de la navigation dans l'ontologie nous avons préféré mettre les synonymes dans le champs *documentation* associée à la classe.

Enfin il faut éviter les boucles dans la hiérarchie d'une classe, nous disons qu'il y a une boucle si une classe A, a une sous classe B et qu'au même temps B est une superclasse de A, si tel est le cas les classes A et B sont considérées comme équivalentes.

Les synonymes sont associés au concept, seulement après une analyse statistique des articles sur lesquels nous avons travaillé. Cette étude consiste à détecter les synonymes les plus fortement liés au concept de la hiérarchie. Cette étude va en même temps servir à définir quelles sont les variantes morphologiques les plus représentatives du domaine juridique, qui vont aussi être associées au concept.

C'est de cette façon que nous pourrons marier les deux approches en utilisant en même temps la racine, si elle est fortement lié au concept de l'ontologie par rapport au domaine choisi, ainsi qu'une restriction des variantes morphologiques les plus représentatives du domaine, en prenant la racine (éventuellement) et les variantes du mot nous augmentons le Rappel et en restreignant ces même variantes nous allons minimiser le bruit et ainsi augmenter la Précision.

2. *Analyse des fratries dans la hiérarchie des classes* :

les fratries ou entités sœurs sont les sous-classes directes de la même superclasse, toutes les filles de la même mère doivent être au même niveau de généralité.

Par exemple :

"محكمة" و "محكمة عليا" ne doivent pas être sous−classe de la même classe parce que "محكمة" (cour) est un concept plus général que « محكمة عليا » (cour suprême).

Par contre "محكمة_عليا" ،"المجالس_القضائية"، " المحاكم_الابتدائية" sont les sous-classes d'une même classe, elles ont le même ordre de généralité. Ceci ne s'applique pas aux concepts de la racine qui peuvent être différents.

Le nombre de sous-classes directes d'une classe varie généralement entre deux et douze dans la plupart des ontologies bien structurées.

3. *Héritage multiple* : une classe peut être une sous-classe de plusieurs classes et donc elle héritera des attributs de ses superclasses.

4. *Une instance ou une classe* : si les concepts forment une hiérarchie naturelle, alors ils doivent être représentés comme des classes.

Par exemple : ( "المحكمة_العليا"، " المجالس القضائية "، " المحاكم_الابتدائية") c'est-à-dire (les tribunaux, les cours et la cour suprême ) pourraient être des instances de la classe( القضاء_ العادي) mais comme ils représentent une hiérarchie naturelle nous les avons définies comme des sous-classes de la classe « القضاء_ العادي ».

Par contre pour « المجالس القضائية », les concepts ( مجلس_ قسنطينة، مجلس_ الجزائر، مجلس_ عنابة) sont des instances pour cette classe.

Le système que nous proposons pour améliorer la recherche d'information sur le web en langue Arabe dans le domaine juridique. S'inscrit dans une architecture de moteur de recherche en langue Arabe permettant la traduction d'une requête en Français ou en Anglais pour retourner des documents écrits en Arabe, en Français ou en Anglais, il peut aussi s'intégrer en tant qu'agent intelligent qui fournit comme réponse une requête étendue, que nous pouvons soumettre à n'importe quel moteur de recherche supportant la langue Arabe.

C'est cette perspective que nous envisageons pour un premier temps et donc notre travail va se focaliser sur l'expansion de la requête en langue Arabe en utilisant l'ontologie du domaine juridique. Comme nous l'avons déjà expliqué, tous les systèmes Arabe de recherche d'information se basent soit sur la racine soit sur le stem soit sur le mot, L'originalité de notre approche est que notre recherche ne va pas se baser sur la racine ou le mot, mais nous allons tenter de nous servir des deux en même temps pour gagner autant du coté Rappel que celui de la Précision, puisque dans l'expansion nous rajoutons non seulement des termes en relation avec le mot clé spécifié dans la requête, mais aussi la racine et les variantes morphologiques fortement liées au domaine juridique.

## 3.3. La construction de l'ontologie

Pour la construction de notre ontologie nous avons essayé de suivre les étapes de construction proposées par N. Noy de Stanford dans **[Noy00]**.

### 3.3.1. Le choix de l'outil utilisé

Après comparaison des outils les plus connus pour la construction et l'édition d'une ontologie, notre choix s'est fixé sur deux outils qui nous ont semblés des plus performants, en l'occurrence Ontolingua et Protege2000, le point fort d'ontolingua c'est Chimaera, outil servant à intégrer d'autres ontologies existantes sur le site http://ontolingua.stanford.edu , le problème c'est que le code source d'ontolingua n'est pas téléchargeable sur le net, l'utilisateur a seulement accès au serveur d'Ontolingua en ligne, de plus Ontolingua, ne supporte pas la langue Arabe, du moins pour le moment, c'est ce qui justifie notre choix qui s'est porté logiquement sur protege2000.

De plus, Protégé2000 est compatible avec OKBC qui permet l'intégration d'ontologies partielles en OKBC, outre cela, il supporte l'héritage multiple permettant à une classe d'hériter de plusieurs superclasses en même temps.
Un autre avantage de Protégé, l'ontologie est facilement modifiable, ceci est très important parce qu'une ontologie est forcément « dynamique » puisqu'elle change avec les nouvelles découvertes scientifiques, ou l'introduction de nouveaux néologismes **[Das02].**

### 3.3.2. La méthode de construction

Nous savons qu'il existe trois manières de construire une ontologie, automatique, semi-automatique ou manuelle. Cette dernière est de loin la plus fiable, mais elle coûte très cher, puisqu'elle demande la collaboration entre informaticiens, linguistes et experts du domaine.
Nous pensons que, si nous commençons le noyau de notre ontologie avec la méthode automatique, il y a de forte chance qu'elle soit vouée à l'échec, Pour cette raison nous avons préféré commencer la construction manuellement du noyau de l'ontologie, même si cela devait prendre plus de temps et même si nous n'arrivons pas au grain le plus fin.
Mais nous comptons enrichir et mettre à jour notre ontologie en construisant cette fois-ci un outil de recherche automatique de concepts dans un corpus juridique Arabe en cas de travaux dans ce sens. Une construction automatique nécessiterait impérativement un corpus Arabe exhaustif et représentatif dans le domaine juridique, ce qui est encore malheureusement indisponible à ce jour. Ce corpus aurait pu servir de base pour une construction automatique de l'ontologie.
Nous pensons également à la réalisation d'un outil de classification pour affiner et mettre à jour notre ontologie. C'est ce que nous comptons faire en perspective puisque L'Académie Algérienne de la langue Arabe travaille sur un projet de réalisation d'un thesaurus de la langue Arabe **[Zem02].**
Nous avons essayé de construire une ontologie dans le domaine du droit Algérien, dont la majeure partie s'inspire du droit Français, à l'exception du statut personnel qui découle des lois de la Chariâa Islamique, et qui a donc beaucoup de points communs avec les lois d'autres pays Arabo-Musulmans, non seulement ceux du Maghreb où le rite Malékite est appliqué mais aussi avec le reste des pays Musulmans, puisque les dernières mises à jour du statut personnel se basent finalement sur les quatre rites existants aussi bien Hanafite, Hanbalite que Chaféite.

### 3.3.3. La base de l'ontologie

.

Pour la construction de l'ontologie nous nous sommes basés essentiellement sur :

- Des articles de l'ONU (www.undp.org),
- Des articles de journaux Arabes sur Internet,
- Les sites de ministères de justice Arabe, notamment Algérien (www.mjustice.org)
- L'encyclopédie judiciaire 2000, électronique Dar el Hillel
- Dz-code Berti-editions Alger.
- Lexalgeria Portail du droit Algérien(http://www.lexalgeria.net)

### 3.3.4. La stratégie de construction

Nous avons essayé de suivre une méthode descendante (top-down) dans l'élaboration de l'embryon de notre ontologie, qui est la plus simple et aussi la plus méthodique **[Jen03].**

Nous avons commencé avec le concept le plus général, c'est-à-dire « النظام القضائي الجزائري = système juridique Algérien» à chaque concept nous allons associer une liste de variantes morphologiques (celles jugées inhérentes au domaine juridique après une étude statistique).

Notons tout de même que le concept «النظام القضائي الجزائري = système juridique Algérien» est lui-même une sous classe de la classe « THING » dans Protege2000 avec sa troisième version, outil que nous avons utilisé pour l'édition de notre ontologie, puisque c'est l'un des outils les plus puissants pour la construction d'ontologies et qui supporte la langue Arabe.

En premier lieu nous avons essayé de trouver tous les hyponymes de ce mot, en se basant toujours sur les articles que nous avons pu réunir, la majorité d'entre eux, sous format numérique puis avec le concours de notre expert, nous avons essayé de corriger et valider la hiérarchie.
Tels que : (القضاء_الإداري = droit administratif) (القضاء العادي : droit commun) Les relations exprimées entre les concepts sont de types *is-a* (hyponymie ou subordination) et *a-part-of* (une partie de).

Par exemple : le concept « قانون العقوبات » (code pénal) est une sous-classe de القانون (droit).

Remarquons seulement qu'un travail complet aurait pris des années. Nous avons tenté d'attacher manuellement, quelques variantes morphologiques pour certains concepts, parce que dans notre système, ce travail devant s'effectuer automatiquement, après avoir réalisé un outil qui parcourt un corpus juridique à la recherche de mots possédant la même racine qu'un concept donné de l'ontologie. Puis effectuant une analyse statistique ne gardant que les variantes qui représentent le plus fidèlement possible le domaine juridique, et dont la fréquence d'apparition est particulièrement élevée, dans ce domaine.

Par exemple : جنحة

« "جنح" "جنحتا " "لجنحة", " لجنحه", " بجنحة", " بجنحه", " فجنحة", " فجنحه", "كجنحة ", "كجنحه", " جنحا", "جنحتي ", "جنحته ", "جنحه ", "جنحهم ", "جنحهن " » nous excluons certaines variantes telles que « "مجنحه ", "أجنحة " "جنوحه" et "جنحت ».

Nous avons tenu à respecter, autant que possible, la hiérarchie des concepts telle qu'elle est utilisée dans les palais de justices. Dans ce cadre notre expert a été très sollicité et le recours à lui et à toute son équipe a été cyclique: consulter, modifier, mettre à jour puis valider.
Nous avons commencé par un développement en largeur d'abord, mais pour certaines branches, nous avons préféré aller plus en profondeur. Ceci dépendait beaucoup plus de la disponibilité des concepts extraits des articles retrouvés sur le web, que d'un choix stratégique.

## 3.4. Aspect général du système

Notre outil s'intègre dans l'architecture générale d'un moteur de recherche, le coté indexation et catégorisation des documents ne seront pas discutées dans ce travail.
Notre système vise à réduire le bruit, qui se trouve être le problème majeur d'une recherche d'information sur Internet de nos jours, en effet une quantité impressionnante d'informations est chaque jour triée et indexée par des moteurs de recherches ou annuaires.
L'utilisateur, submergé par cette gigantesque vague diluvienne de documents retournés, se trouve perdu et désarmé: *autant chercher une aiguille dans une botte de foin* !
Améliorer le rappel ainsi que la précision de la recherche est aujourd'hui un problème crucial à résoudre, nous tentons par la réalisation de notre système d'apporter une aide appréciable pour la réduction du silence et du bruit en augmentant respectivement le rappel et la précision et ce en langue Arabe!

Le système proposé se compose de sept parties suivantes (schématisées dans la figure23).

1) L'interface utilisateur ;
2) Analyseur de requête ;

3) Traducteur de requête ;
4) L'expansion de requête avec l'ontologie Arabe ;
5) L'expansion de requête avec Wordnet ;
6) Le processus de recherche ;
7) L'affichage de résultats.

### 3.4.1. L'interface utilisateur

Permettant la saisie de la requête, formée d'une suite de mot-clés en langue Arabe.
Dans cette partie, une requête en langue Arabe est introduite. Elle peut être composée d'un ou de plusieurs mots clés.
Pour éviter d'introduire n'importe quelle suite de lettres et effectuer une recherche infructueuse, nous avons pensé à vérifier la véracité de la requête de l'utilisateur et ce en cherchant dans la liste des mots reconnus par le web, si les mots sont reconnus alors le système passe la requête à l'analyseur, dans le cas échéant le système s'informe auprès de l'utilisateur si ce dernier ne s'est pas trompé en affichant le message :

« **النظام لم يتعرف على الكلمة:** "ب ب ب ب ب " **حاول تصحيح الكلمة أو استعمل كلمة أخرى** »

Nous savons qu'en se basant uniquement sur les mots de la requête, il y a une forte probabilité de silence, certains systèmes utilisent la racine de chaque terme de la requête pour effectuer la recherche avec les différentes variantes du mot.

L'extraction de la racine, est bénéfique pour éviter le silence mais son inconvénient c'est d'augmenter considérablement le bruit, en d'autres mots elle dévie la recherche en livrant à l'utilisateur des documents non pertinents.

D'autres systèmes optent pour le stem. Le stem (le radical) s'obtient en ôtant du mot tous les préfixes et les suffixes **[AlT00]**. Bien que cette méthode augmente la probabilité de retourner des documents pertinents, ils peuvent néanmoins être cause de silence.

La troisième méthode utilisée c'est le mot. En se basant sur le mot le système risque fort de ne pas trouver dans les documents pertinents le mot exact mais des variantes de ce mot, par conséquent ces documents ne seront pas retournés, et donc c'est une grande source de silence.

Généralement à cette étape, le système effectue une extraction de la racine de chaque mot de la requête, parce que dans la majorité des cas les mots sont indexés avec leur racine. Dans notre cas, nous avons pensé effectuer une recherche avec les mots de la requête, car c'est ainsi qu'ils sont représentés dans la hiérarchie de l'ontologie.
Néanmoins après avoir pesé le pour et le contre nous avons préféré donner à l'utilisateur la possibilité de choisir entre une recherche *Standard*, qui sera effectuée par défaut et une recherche *Personnalisée*.

**La recherche Standard** *:* c'est la recherche à laquelle nous avons pensé en premier lieu, c'est-à-dire basée sur le mot.

**La recherche Personnalisée** : elle donnera à l'utilisateur la possibilité de choisir entre une recherche à ***haut Rappel*** par exemple pour un brain-storming, ou une recherche pointue à ***haute précision***.

Dans le premier cas nous privilégierions l'extraction de la racine, pour donner un taux de *Rappel* élevé.
Dans le second cas par contre, nous allons nous baser sur le mot introduit, en lui associant les variantes morphologiques fortement liées au domaine juridique, méthode permettant de retourner des résultats avec une *Précision* plus ou moins remarquable.
Déjà de cette façon, nous éliminons une grande source de bruit, d'autres part c'est vrai que nous ouvrons une petite porte au silence, mais nous verrons dans §3.4.3 comment traiter ce problème.
Toujours dans l'interface utilisateur, le système donne le choix d'opter pour une recherche monolingue ou d'effectuer une traduction automatique en Français ou en Anglais pour une recherche multilingue.
La recherche monolingue, utilisera l'ontologie juridique Arabe pour étendre la requête et retournera des documents exclusivement en Arabe.

### 3.4.2. L'analyseur de Requête

Dans cette partie, la requête de l'utilisateur est soumise, elle doit être analysée, cette analyse comporte l'élimination des mots vides et la normalisation.

- *Élimination des mots vides* :

Une fois la requête introduite, Les mots vides sont éliminés, articles, prépositions, adverbes de lieu ou de temps, tels que « ال، في، من، ك، ل، مثل..... ».

Une liste de mots vides Arabe commune, a été dégagée par TREC2001, nous allons nous baser sur cette liste pour reconnaître si un mot est signifiant ou pas et donc l'éliminer ou en tenir compte lors de la recherche.

Exemple de requête : حقوق الطفل في الجزائر

Nous gardons (حقوق ، طفل ، جزائر) la raison de cette élimination est que d'un coté ces mots ne sont pas indexés, d'un autre coté ils n'apportent aucune information utile à la recherche.

- Normalisation :

C'est une opération qui consiste à :

- ❖ Enlever la ponctuation ;
- ❖ Enlever les signes diacritiques ;
- ❖ Enlever tout ce qui n'est pas lettre ;
- ❖ Remplacer إ , آ, أ par ا ;
- ❖ Remplacer ي final par ى;
- ❖ Remplacer ة final par ه .

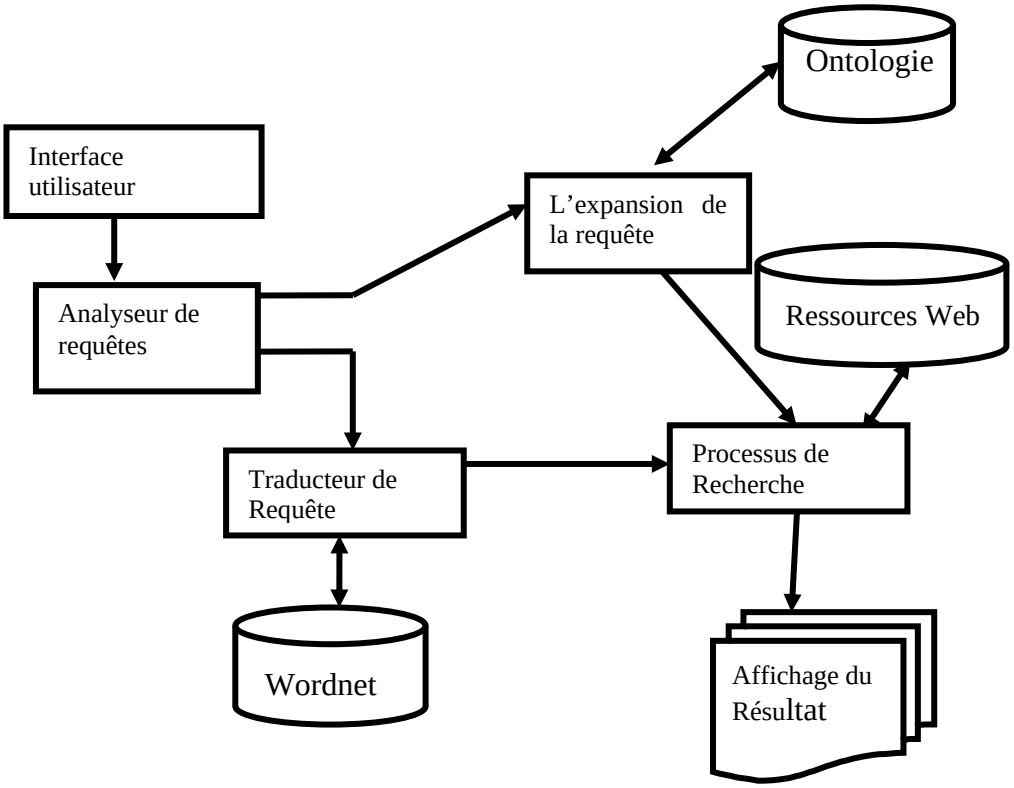



**Figure 23: Architecture générale du système**

### 3.4.3. L'expansion de la requête

L'expansion de la requête se fait en utilisant une ontologie du domaine juridique, grâce à un navigateur dans l'ontologie, le système va tenter d'associer les mots de la requêtes aux concepts de l'ontologie, retourner les synonymes et/ou des hyperonymes et hyponymes pour étendre la requête (si aucun mot de la requête n'est associé aux concepts de l'ontologie le système demande à l'utilisateur s'il ne s'est pas trompé dans la saisie des mots et peut éventuellement faire des propositions).
Le système associe éventuellement les variantes morphologiques.
Nous donnons ci-dessous un exemple de concept dans la hiérarchie de l'ontologie, en visualisant ces hyperonymes et ses hyponymes.

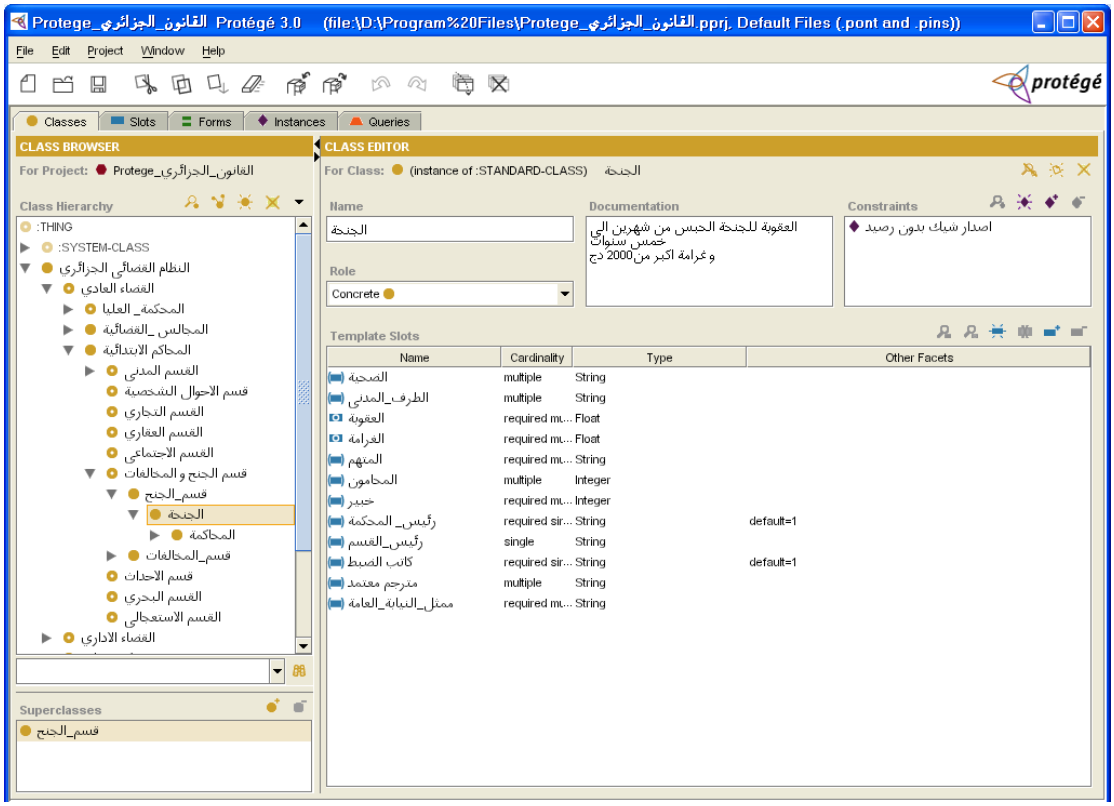



**Figure 24: le concept " جنحة" dans la hierarchie**

Si nous prenons l'exemple d'une requête qui comporte le mot clé " جنحة" le navigateur va parcourir en premier les top-types de l'ontologie, puis prendre une branche de la hiérarchie et la parcourir en profondeur, jusqu'à arriver aux feuilles.
Le système va donc passer en vue les concepts " النظام_القضائي_الجزائري", " مدونات_القانون_الجزائري", " المحاكم ", " القانون", " المجلس_الاسلامي_الاعلى".

Comme le mot " جنحة " ne se trouve pas parmi ces concepts, nous allons développer la première branche c'est-à-dire " النظام_القضائي_الجزائري " puis nous passerons au concept " القضاء_العادي " puis " المجالس_القضائية " puis " الغرفة_الجزائية " puis " قسم_الجنح_و_المخالفات " puis " قسم_الجنح " et enfin " جنحة " nous pouvons accéder aux propriétés de ce concept, qui sont :

1) les slots hérités de l'hyperonyme " قسم_الجنح "

(الضحية، المتهم، الطرف_المدني، رئيس_المحكمة، رئيس_القسم، ممثل_النيابة_العامة، كاتب_الضبط، المحامون، "خبير، مترجم_معتمد).

2) Les propres slots du concept « جنحة » et qui sont :

(العقوبة، الغرامة).

La propriété « الغرامة » a comme facette une valeur réelle d'au moins 2000 DA.

La propriété « العقوبة » a comme facette la valeur comprise entre *deux mois* et *cinq ans* d'emprisonnement.

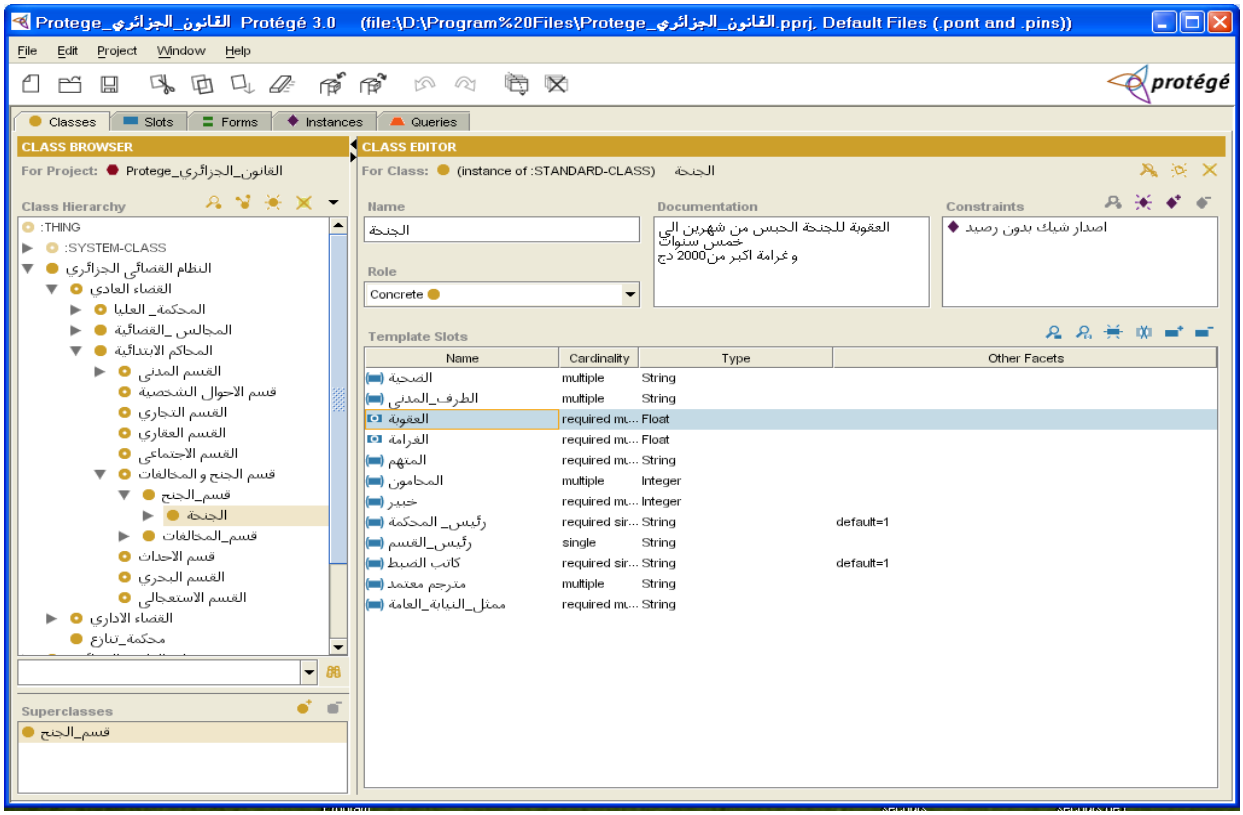



**Figure 25: les slots herités et les slots propres au concept " جنحة"**

Nous avons aussi une instance de ce concept qui est " إصدار_شيك_بدون_رصيد".

Nous avons pensé à combiner les deux méthodes de recherche celle basée sur le mot et celle basée sur la racine, pour avoir aussi bien un très bon Rappel qu'une meilleure Précision.
À la fin de la construction, nous associerons, à chaque concept non pas toutes les variantes du mot comme c'est le cas pour l'utilisation de la racine, mais nous prendrons seulement celles jugées, après

une étude statistique sur des corpus exhaustifs du domaine juridique les plus utilisées dans ce domaine.

Nous pouvons remarquer aussi que la propriété " العقوبة " est une propriété commune aux classes " جنحة" et " مخالفة", ces classes représentent le domaine de cette propriété.

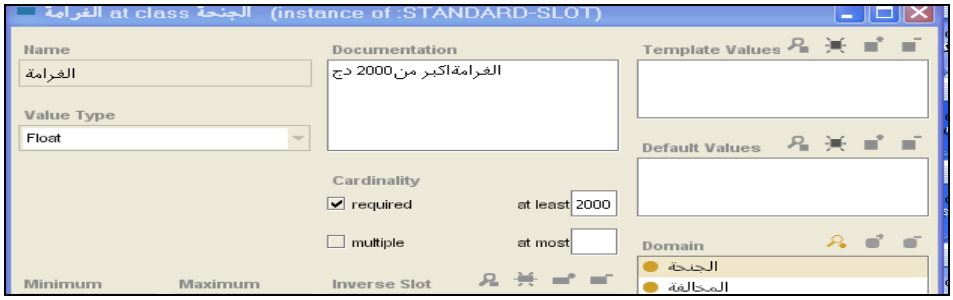



**Figure 26: Le domaine et les facettes du concept "الغرامة"**

Les variantes morphologiques associées au concept " جنحة" seront " جنحه","جنح" qui représente, sans signes diacritiques, aussi bien la racine "جَنَحَ" que (le pluriel cassé=non sain) "جُنَحٌ ", ainsi que "جنحتا", "جنحته ", "جنحتي ", "جنحا ", "كجنحه ", "كجنحة ", "فجنحه ", "فجنحة ", "بجنحه ", "بجنحة ", "لجنحه ", "لجنحة ", " أجنحة" ", "جنوحه " nous excluons certaines variantes telles " جنحت" et " جنحهن ", "جنحهم ", "جنحه مجنحه" etc. qui sont rarement utilisées dans le domaine juridique.

Par exemple pour la requête « قانون العقوبات » (le code pénal) si nous prenons le mot « عقوبات » sa racine est "عاقب" = (pénaliser) parmi ses variantes nous pouvons trouver : ،عقاب ،عقوبة، معاقِبة، معاقَبة ، العقابي، لعقاب، العقا ب ، عقوبتان، عقابكم، عقابنا، تعاقبون، يعاقبون، يعاقب... la liste est encore longue, nous pouvons deviner d'ores et déjà que si nous utilisons la racine (avec toutes ses dérivées) quelle quantité de documents allons-nous obtenir? Mais avec quel degré de pertinence ?

Ce que nous allons faire c'est classer ces termes par ordre décroissant de fréquence d'apparence dans les documents juridiques, nous garderons celles que nous rencontrons plus souvent dans les corpus juridiques et nous omettant celles qui restent, de cette manière nous préservons les avantages de la racine et nous nous débarrassons des inconvénients du mot.

Une fois les dérivées rajoutées, nous pouvons leur adjoindre soit des hyponymes soit des hyperonymes (comme "الغرفة_الجزائية", " قسم_الجنح_و_المخالفات", " etc. soit les deux en même temps. Dans les perspectives nous ferions une étude comparative entre ces méthodes d'expansion et nous donnerions une synthèse concernant les résultats obtenus.

Par exemple, nous opterons pour une expansion par hyperonymes (généralement en cas de recherche pour un taux de rappel élevé), alors que pour une recherche avec un taux de précision élevé nous préférons une expansion avec hyponymes.

### 3.4.4. Traducteur de requête

Il arrive qu'un utilisateur et pour répondre aux exigences de la mondialisation, puisse désirer obtenir des documents dans deux langues différentes voire même trois, en l'occurrence l'Arabe, le Français et l'Anglais.
Dans ce cas, le système va faire appel à un outil de traduction automatique en ligne, nous connaissons tous, les faiblesses de la traduction automatique qui, bien qu'elle ait presque commencé avec l'avènement informatique, reste encore à son étape de balbutiement. Nous savons aussi que la qualité de la traduction va influer indéniablement sur les résultats de la recherche, notre seule consolation est que nous ne traduisons pas des phrases structurées, aussi simples soient-elles, mais des mots séparés.
Nous pensons que l'ambiguïté due à la traduction ne sera que minime, relativement à la traduction en général, chose que nous avons remarqué concernant la traduction de l'Arabe vers l'Anglais, en utilisant le système *Tarjim* de *Ajeeb* de Sakhr Software**[Sak04].**
De plus nous avons pensé, en perspective, à la réalisation d'un système de traduction utilisant lui-même l'ontologie du domaine juridique, travail qui sera, peut-être, développé dans un autre cadre.
Nous avons pu remarquer combien le domaine juridique se prête assez bien à une traduction de l'Arabe vers le Français, facilité due en grande partie à notre culture bilingue et à notre histoire à travers les générations successives. Nous devons tout de même faire exception d'une partie importante de notre système judiciaire concernant les concepts utilisés dans le code du statut personnel " قانون الأحوال الشخصية" dont les mots se trouvent très liée à la Chari'a Islamique et à la culture et aux traditions Arabo-Musulmanes.

Il arrive que pour une raison ou pour une autre, l'utilisateur aie envie de trouver des documents en langue Française ou Anglaise (silence pour la requête en Arabe ou simplement pour des raisons de recherches).
La requête traduite sera ensuite étendue, grâce à la base lexicale Wordnet ou EuroWordnet.

Outre ses innombrables utilisations, Wordnet fut utilisé entre autre dans la désambiguïsation du sens des mots **[Res95]** et **[Voo93],** dans la catégorisation des textes **[Gom97],** l'extraction d'information **[Cha97],** dans le traitement du langage naturel **[Seg97]** et notamment dans l'expansion de la requête **[Voo94].**

***Phase traduction*** *:*
Pour la traduction de la requête nous avons choisi d'utiliser le système de traduction en ligne disponible gratuitement *Tarjim* de *Ajeeb* (www.sakhrsoft.com). Nous pensons que c'est un outil de traduction des plus fiables sur le web, pour le passage de l'Arabe vers l'Anglais. Par contre pour la paire Arabe-Français, il nous a été difficile de trouver un outil plus ou moins crédible, en désespoir de cause nous avons opté pour (***1-800-translate***) qui est libre d'utilisation sur le Web, mais en nombre limité de mots.

***Phase expansion :***
Pour l'expansion de la requête traduite en Anglais, nous utilisons l'ontologie générique disponible sur le web *Wordnet*, et nous pensons utiliser *Eurowordne*t pour la requête traduite en Français.
Nous n'avons pu travailler avec EuroWordnet son exploitation est sujette à une licence que nous n'avons pu obtenir gratuitement. Son utilisation paie en nombre de synsets. Par conséquent, nous n'avons pu expérimenter l'expansion d'une requête en Français avec EuroWordnet, et tous nos exemples sont étendus avec Wordnet.

Nous avons préféré utilisé dans un premier temps, seulement les définitions des mots, données par *Wordnet*, parce que nous avons remarqué un nombre, des fois impressionnant, de synsets qui n'ont dans la majorité des cas qu'un fil très ténu avec le domaine juridique.
Une fois l'expansion effectuée la requête est soumise au système de recherche pour rendre cette fois-ci des documents en Français ou en Anglais selon le choix de l'utilisateur.
Remarquons que si la recherche est effectuée avec un moteur ayant l'option de la recherche avancée, ce que nous gagnerons c'est seulement l'expansion de la requête.
Les résultats de notre expérimentation avec Wordnet sont exposés au chapitre 4.

### 3.4.5. Processus de recherche

Quelle soit traduite ou pas, la requête étendue est soumise à un outil de recherche de documents qui va avoir en sortie une liste de pages web. Dans nos expérimentations nous avons effectué notre recherche avec différents moteurs, bien que comparer les performances de certains moteurs de recherche n'ait pas été notre but. Cependant dans notre conception nous supposons le développement d'un outil conçu spécialement pour la recherche dans des bases de données sur le web, ou l'utilisation d'un outil spécialement conçu pour la recherche de documents tel le moteur Smart par exemple, qui est largement utilisé dans les conférences d'évaluation de nouvelles techniques de recherche, comme c'est le cas pour TREC. Des exemples seront discutés dans le chapitre suivant.

### 3.4.6. L'affichage des résultats

Les documents retrouvés par le système, vont subir un traitement afin de les classer dans un ordre décroissant de pertinence.
Si nous utilisons un moteur de recherche alors l'affichage des résultats se fera selon l'algorithme de pertinence du moteur utilisé.
Dans le cas du développement d'un outil de recherche, nous spécifierons selon quelle règle un document sera jugé plus pertinent que les autres, pour être classé en premier et comment sera calculée le score de pertinence de chaque mot du document vis-à-vis de la requête et juger du degré de similarité entre la requête et un document, ceci ferait l'objet d'un autre projet, si une quelconque partie donnerait suite à notre travail.

Remarquons enfin, que la hiérarchie peut être visualisée en format HTML.
Nous pouvons aussi voir la description de chaque concept, sous forme de frame, avec ses différents champs tels que ses super- classe, ses sous- classes, ses instances, ses slots et les différentes contraintes sur ces slots, comme c'est montré dans la figure suivante :

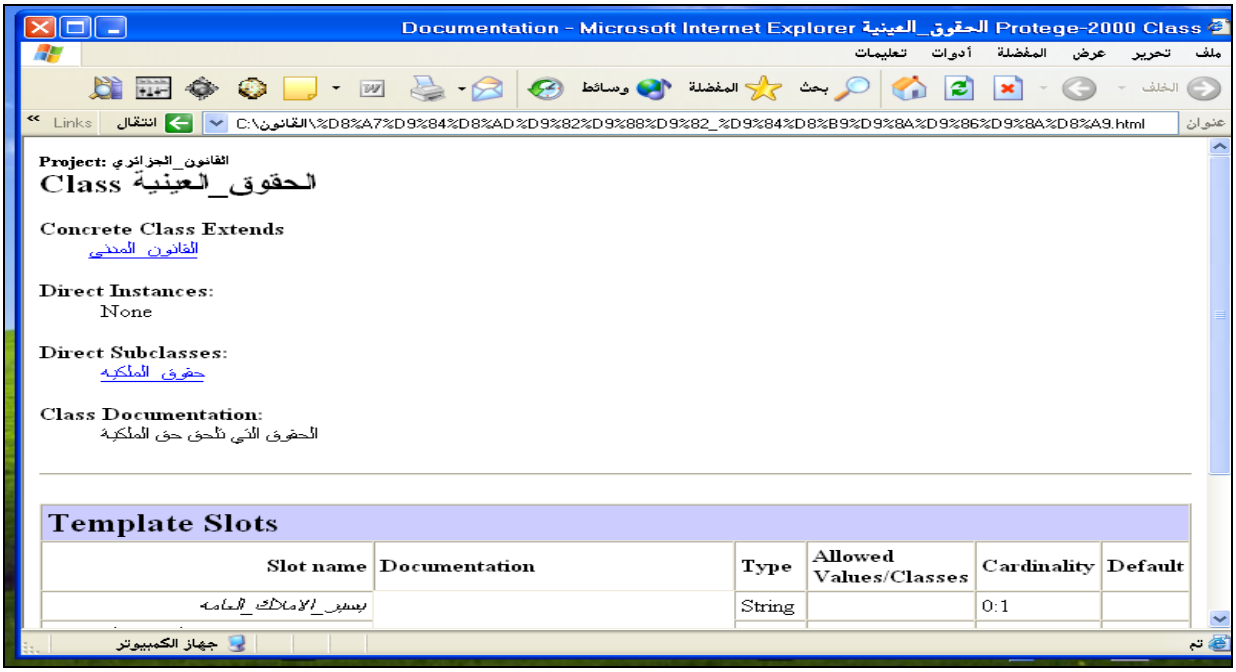



**Figure 27: les differents champs du concept " الحقوق_العينية"**

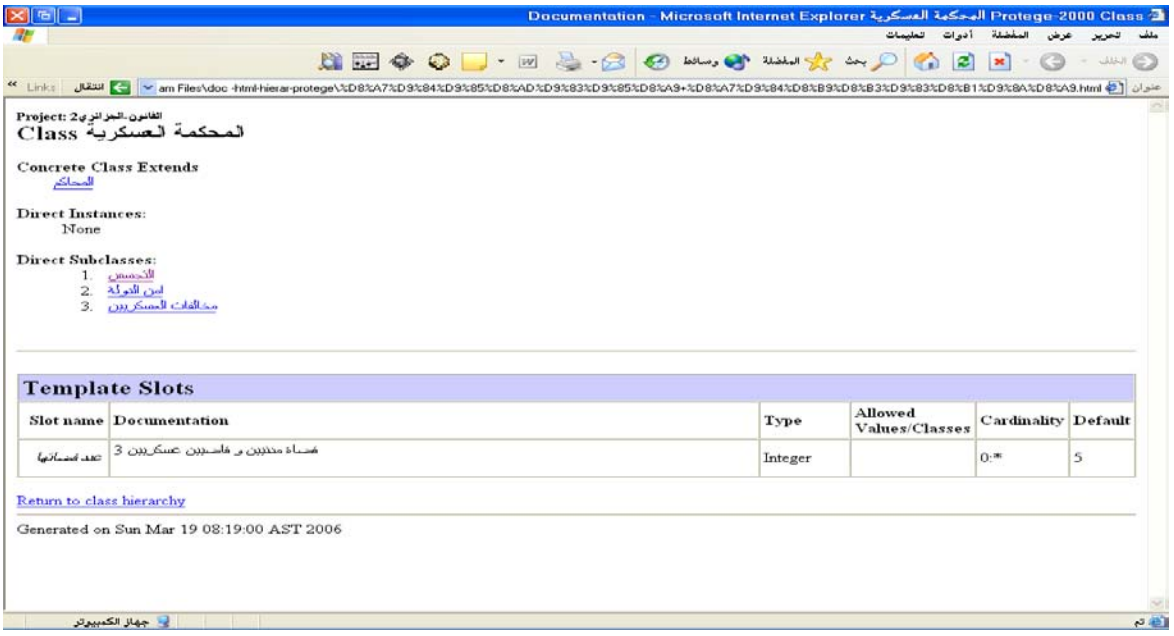



**Figure 28: Les differents champs du concept" المحكمة_العسكرية"**

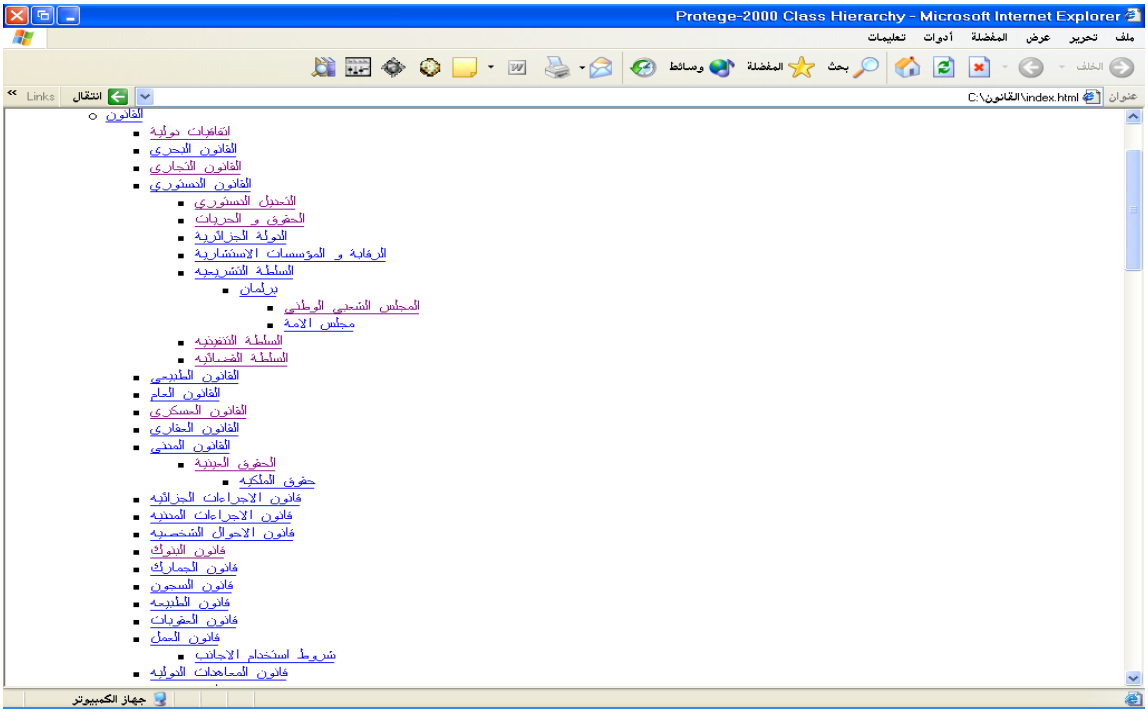



**Figure 29: La description de la hierarchie**

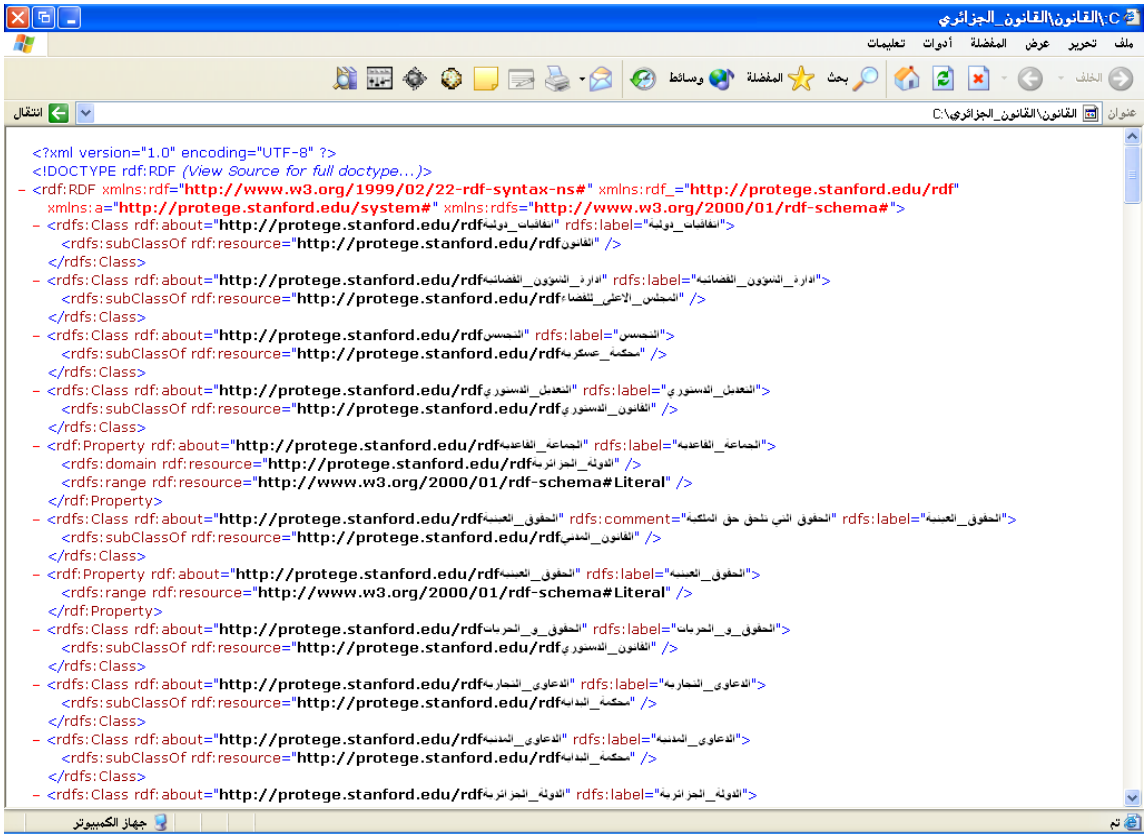



**Figure 30: La presentation des concepts sous format RDF.**

### 3.4.7. Conclusion

Nous avons essayé d'expliciter les différentes étapes de notre système, nous avons particulièrement insisté sur la construction de l'ontologie et la façon d'étendre la requête en utilisant les concepts de l'ontologie, parce que nous avons focalisé notre travail sur ces deux points. Néanmoins nous avons tenté de donner ne serait-ce qu'un petit aperçu sur les autres parties pour deux raisons, la première c'est pour les perspectives du système, la seconde étant pour pouvoir situer notre travail dans l'architecture générale d'un moteur de recherche.

Dans le chapitre suivant, nous essaierons de donner quelques exemples d'expansion de requête, nous discuterons les résultats obtenus, en utilisant différents moteurs de recherche supportant des requêtes en langue Arabe.

# 4. Résultats et discussion

## 4.1. Introduction

Dans notre travail, nous avons expérimenté au début, l'expansion par les synonymes puis hyperonymes, ne dépassant pas les deux niveaux, ainsi que par les hyponymes et les propriétés des concepts exploitant les relations existant entre aussi bien concepts-concepts, que concepts-slots. Puis nous avons essayé l'expansion hétérogène.
Nous avons un peu délaissé l'expansion par synonymie. La raison en est que, d'une part nous n'avons pu réunir les synonymes de tous les concepts, dû au manque d'ouvrages dans nos bibliothèques et de sites gratuits sur Internet. D'autre part, d'après l'expérimentation de Voorhees **[Voo93]**, l'expansion par synonymie a détérioré la performance de la recherche.
D'un autre coté, les travaux de Calcagno, Buscaldi **[Cal04]**, **[Bus05]**, ont montré que le rappel a été amélioré mais la précision s'est trouvée détériorée.

### Les outils de recherche utilisés

Dans le chapitre3, nous avons expliqué que notre système peut être intégré dans l'architecture d'un moteur de recherche ou être utilisé comme agent intelligent qui fourni à un moteur de recherche une requête étendue. Nous avons essayé de focaliser notre travail sur la construction de l'ontologie et la manière d'étendre la requête, en utilisant cette ontologie.
Après quoi, nous avons tenté de donner une petite évaluation en terme de résultats de recherche avec expansion de requête et ce en utilisant des moteurs différents pour voir la fluctuation des résultats selon l'outil utilisé.

### Les mesures

Les mesures utilisées usuellement pour comparer les systèmes de recherche d'information sur le Web, sont la précision et le rappel, définis dans les chapitres précédents. Les chercheurs sont beaucoup plus intéressés par la précision des résultats sur la première page ou les deux premières, plutôt que sur le rappel. Par conséquent la mesure usuelle est la précision calculée sur les dix ou vingt premiers documents retournés **[Haw00]**.
Outre cela, le Rappel ne peut être calculé que sur une collection statique de document, comme celle utilisée dans TREC.

Pour notre système nous donnons le nombre de documents retournés puis nous calculons la précision sur les vingt premiers documents.

### Traitement de quelques exemples

Nous allons traiter quelques exemples de requête et essayer de donner une première synthèse du travail effectué.

Nous rappelons que notre but n'est point de comparer les performances des moteurs de recherche utilisés, mais juste la comparaison entre une requête simple et une requête étendue, en utilisant l'ontologie que nous avons construite, ainsi que la manière d'étendre cette même requête, mais traduite, avec Wordnet.

**Exemple1:**

**Requête simple:** قانون العقوبات

| Moteur utilisé | Nombre de pages Arabes | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Google | 29000 | 8 | 0.4 |

**Tableau 1: Recherche simple de la requete"قانون العقوبات" avec Google**

**Requête étendue** avec synonymes : قانون العقوبات حكم تشريع نص مخالفة جناية جنحة

| **Moteur utilisé** | Nombre de pages Arabes | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Google | 137 | 12 | 0.60 |

**Tableau 2: Recherche étendue de "قانون العقوبات"**

En introduisant plus de 15 synonymes pour قانون العقوبات tapant la requête en Arabe nous avons utilisé le *OR* en Anglais, aucun document n'a été retourné. Puis nous avons utilisé le أو en Arabe cela n'a rien donné non plus, dans la documentation sur Google que nous avons consulté nous n'avons pas trouvé d'explication. En diminuant les synonymes à 8 nous avons obtenu les résultats ci-dessus.

**Exemple2:**

**Requête simple :** مجلس الأمة

| **Moteur utilisé** | Nombre de pages Arabes retournées | Nombre de documents pertinents sur les 30 premiers | Précision |
|---|---|---|---|
| Google | 63800 | 10 | 0.33 |

**Tableau 3: Recherche simple de "مجلس الامة"**

**Requête étendue avec hyperonymes : (مجلس الأمة ، البرلمان، السلطة التشريعية، القانون الدستوري)**

| **Moteur utilisé** | Nombre de pages Arabes retournées | Nombre de documents pertinents sur les 30 premiers | Précision |
|---|---|---|---|
| Google | 320 | 13 | 0.43 |

**Tableau 4: Recherche étendue de "مجلس الامة"**

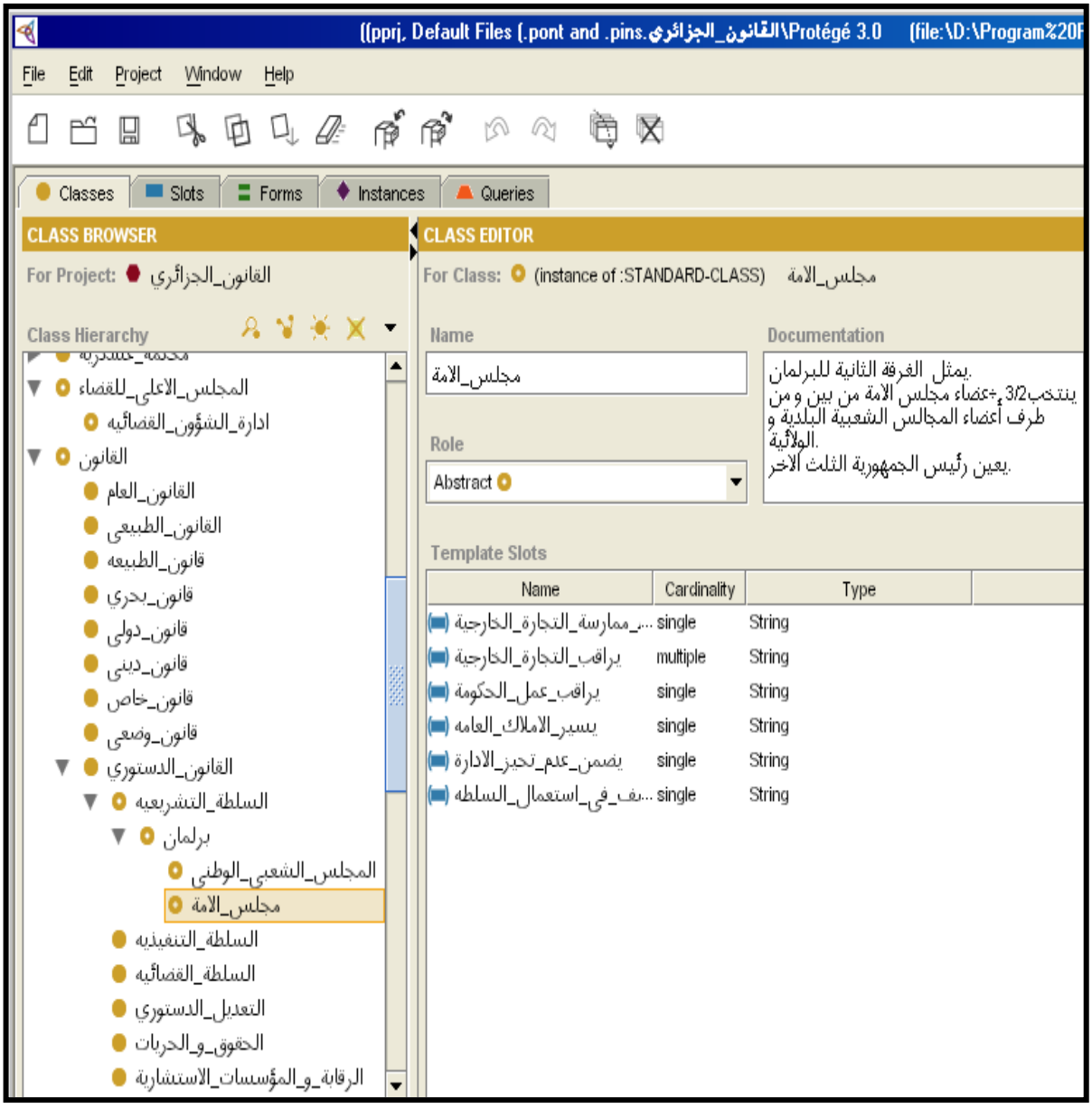



**Figure 31: les hyperonymes du concepts "مجلس الأمة"**

**Exemple3 :**

**Requête simple** : شروط استخدام الأجانب

| **Moteur utilisé** | Nombre de pages | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Hahooa | 24 | 0 | 0.00 |

**Tableau 5: Recherche simple de "شروط استخدام الاجانب"**

**Requête étendue** avec propriétés (relations entre concept) : شروط استخدام الأجانب، قانون العمل، عقد عمل، رخصة عمل

| **Moteur utilisé** | Nombre de pages Arabes | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Hahooa | 285 | 6 | 0.30 |

**Tableau 6: Recherche étendue de "شروط استخدام الاجانب"**

**Exemple 4 :**

**Requête simple** : الحقوق العينية

| **Moteur utilisé** | Nombre de pages | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Ayna | 51 | 4 | 0.20 |

**Tableau 7: Recherche simple de "الحقوق العينية"**

**Requête étendue** avec hyperonymes et hyponymes : الحقوق العينية، القانون المدني، حقوق الملكية

| **Moteur utilisé** | Nombre de pages Arabes | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Ayna | 5 | 3 | 0.60 |

**Tableau 8:Recherche étendue de "الحقوق العينية":**

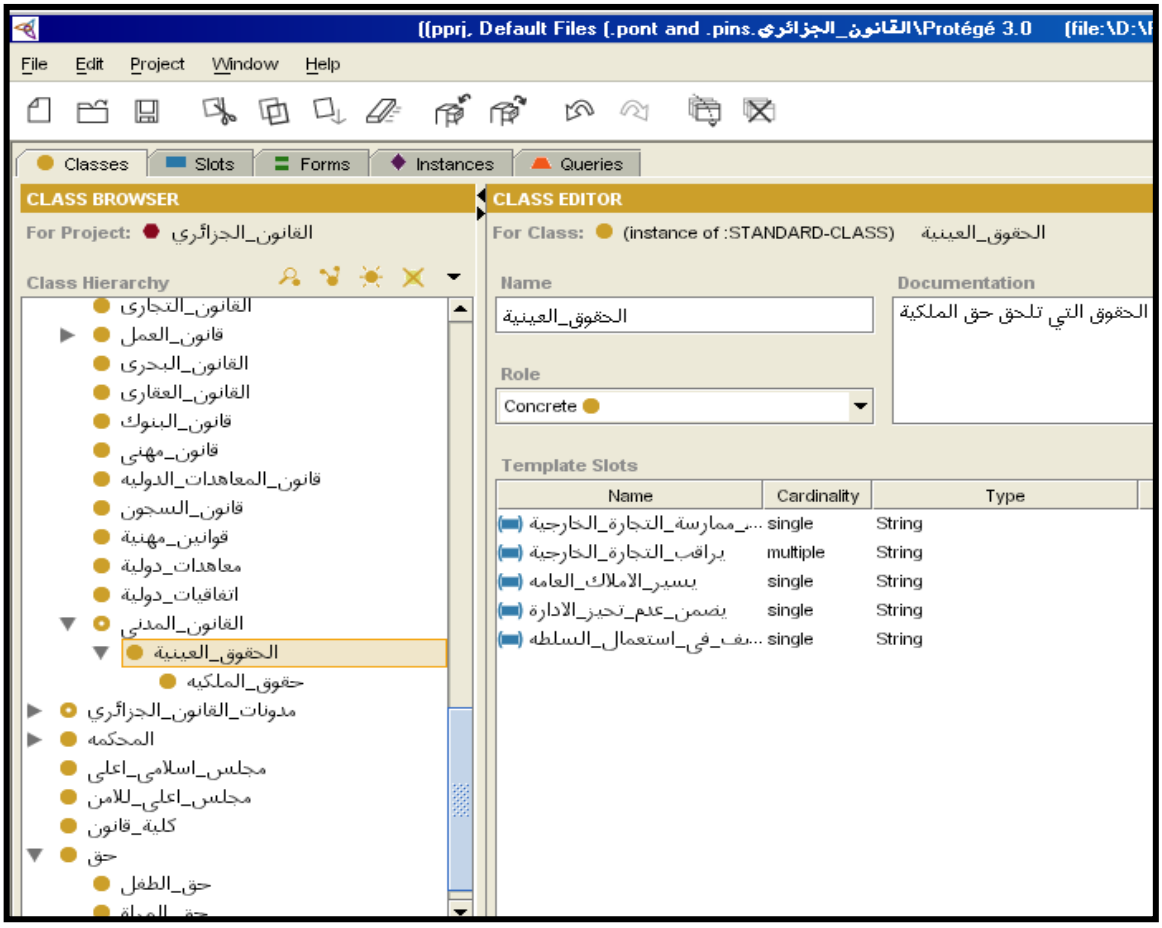



**Figure 32: hyperonymes et hyponymes du concept "الحقوق العينية"**

**Exemple5:**

**Requête simple:** حقوق الطفل في الجزائر

| **Moteur utilisé** | Nombre de pages retournées | Nombre de documents Pertinents Sur les 20 premiers | Précision |
|---|---|---|---|
| **AltaVista** | 69200 | 3 | 0.15 |

**Tableau 9: Recherche simple de "حقوق الطفل في الجزائر"**

**Requête étendue** avec propriétés: حقوق الطفل في الجزائر ، مؤسسة الرعاية الاجتماعية، جمعية الأمم المتحدة،حماية الطفل

| **Moteur utilisé** | Nombre de pages retournées | Nombre de documents Pertinents Sur les 20 premiers | Précision |
|---|---|---|---|
| **AltaVista** | 1 | 1 | 1.00 |

**Tableau 10: Recherche étendue de "حقوق الطفل في الجزائر"**

Nous allons essayer de donner maintenant l'exemple d'une requête traduite simple et une requête traduite étendue avec Wordnet. Nous ne pouvons utilisé EuroWordnet en ligne que sous licence**.**

**Exemple6:**

**Requête simple :** الخلع

| **Moteur utilisé** | Nombre de pages retournées | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Google | 102000 | 16 | 0.80 |

**Tableau 11: Recherche simple de "الخلع"**

**Requête étendue** avec définition (الخلع فك الرابطة الزوجية من طرف الزوجة)

| **Moteur utilisé** | Nombre de pages Arabes | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Google | 46 | 11 | 0.55 |

**Tableau 12: Recherhce étendue de "الخلع"**

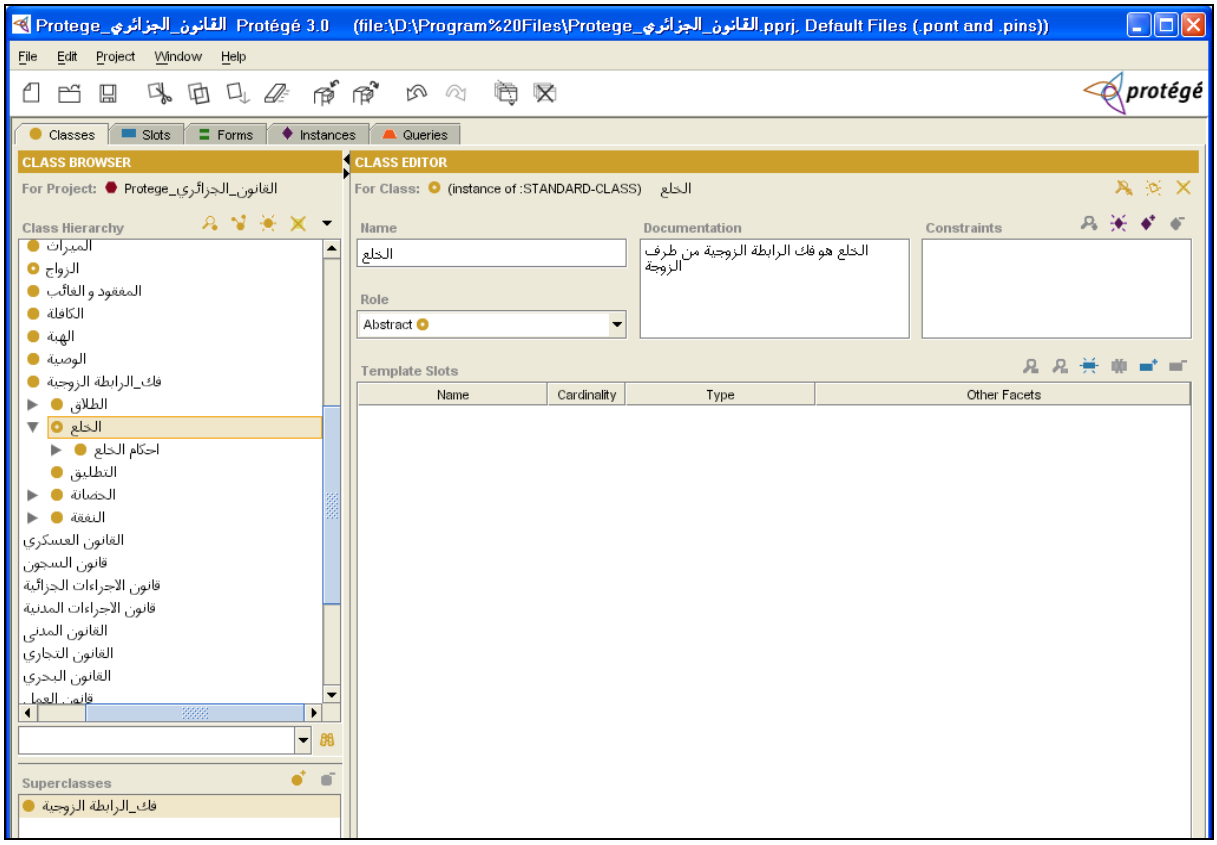



**Figure 33: " الخلع" dans la hierarchie des classes**

### Traduction de la requête

Avec Tarjim nous avons obtenu la traduction suivante en Anglais: *the removal*

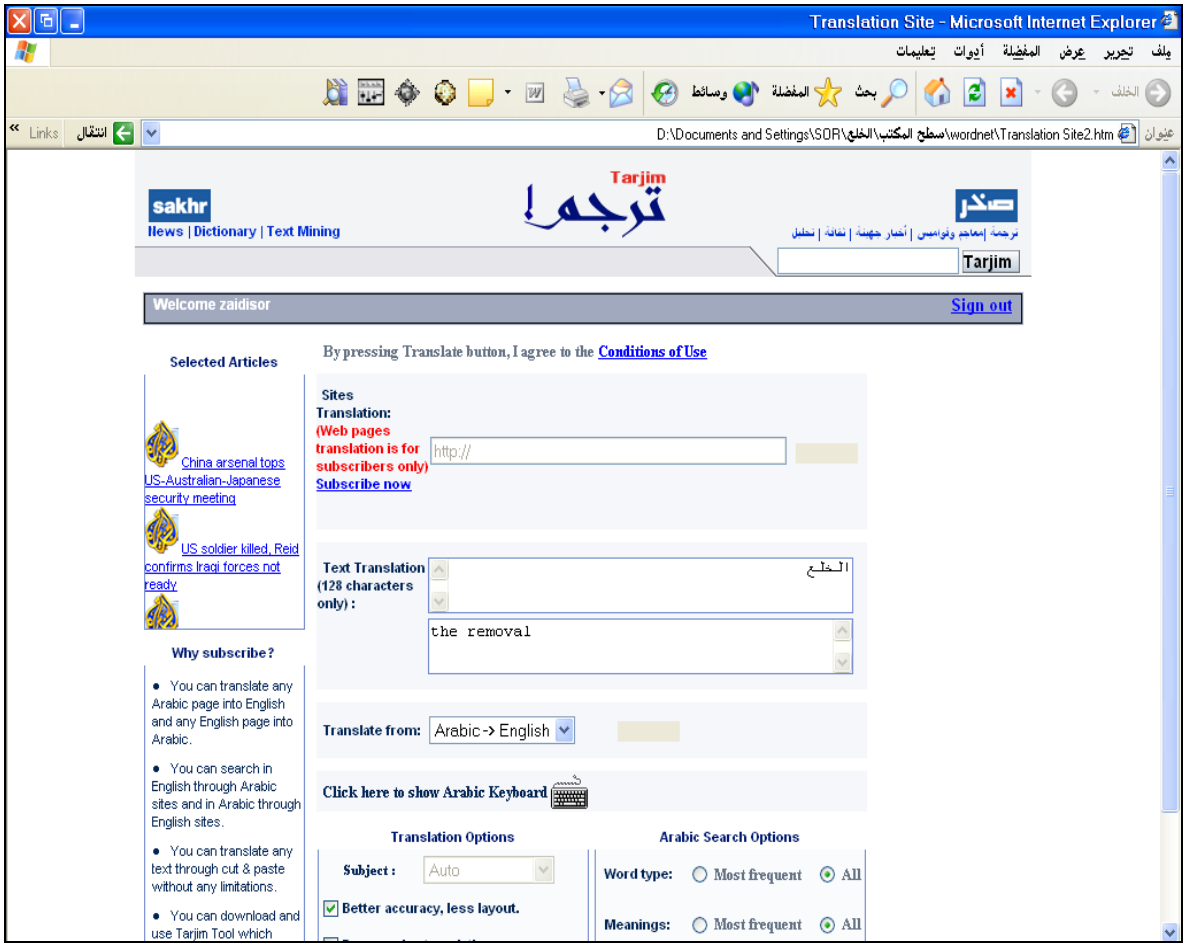



**Figure 34: Traduction de "الخلع" avec Tarjim.**

### Expansion de la requête avec Wordnet

Nous allons prendre la traduction telle quelle est et nous allons chercher les mots suggérés, à Wordnet pour rechercher leurs définitions**.**

En tapant, the *removal* dans Wordnet, qui est un terme trop général, pour designer un sens aussi pointu que celui dans la législation Algérienne, nous obtenons la fenêtre suivante :

.

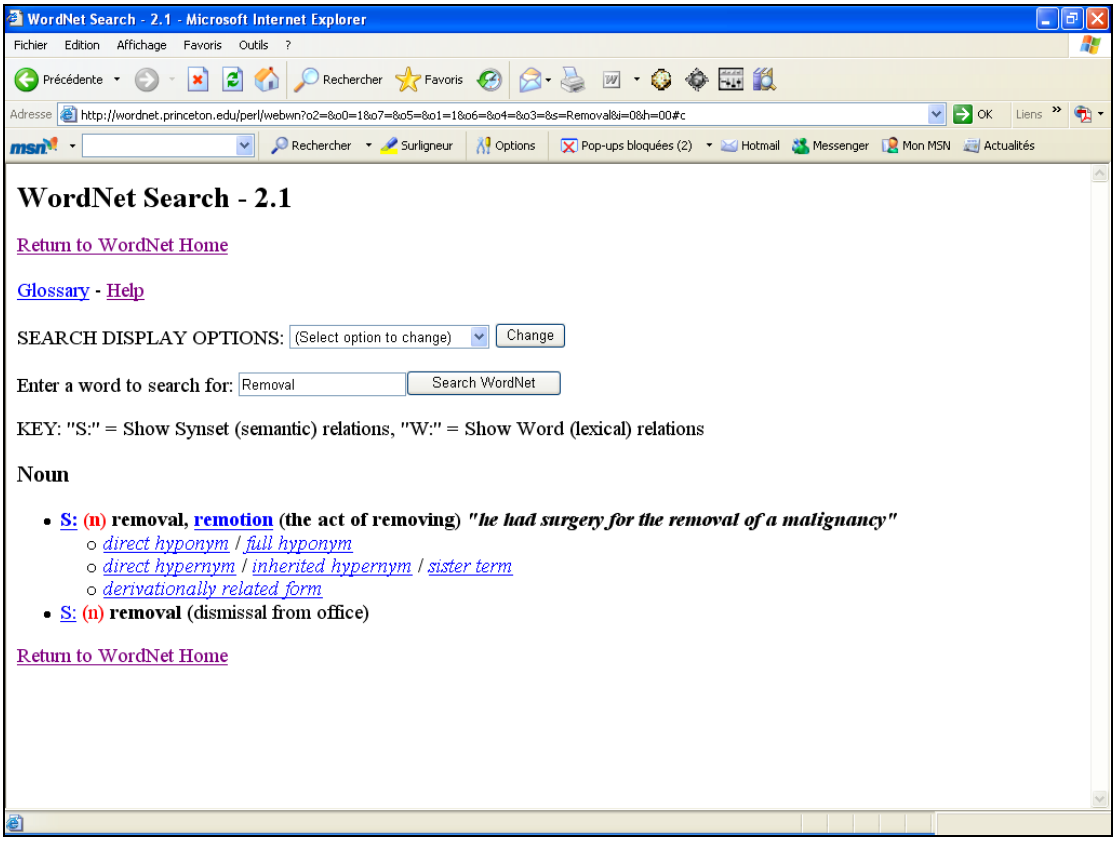



**Figure 35: Recherche de *"Removal"* dans Wordnet**

Nous avons effectué la recherche avec la requête étendue avec la définition de Wordnet, nous avons obtenu les résultats suivants:

| **Moteur utilisé** | Nombre de pages retournées | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Google | 17600000 | 00 | 0.00 |

**Tableau 13: Recherche étendue de "*removal*"**

**Exemple7:**

**Traduction de la requête :** قانون العقوبات

Après la traduction en Anglais avec Tarjim de Ajeeb (Sakhr) قانون العقوبات a donné: *Penal code*

**Expansion de la requête avec Wordnet**

En tapant *Penal code* dans Wordnet nous avons eu comme réponse :

**penal code** -- (the legal code governing crimes and their punishment)

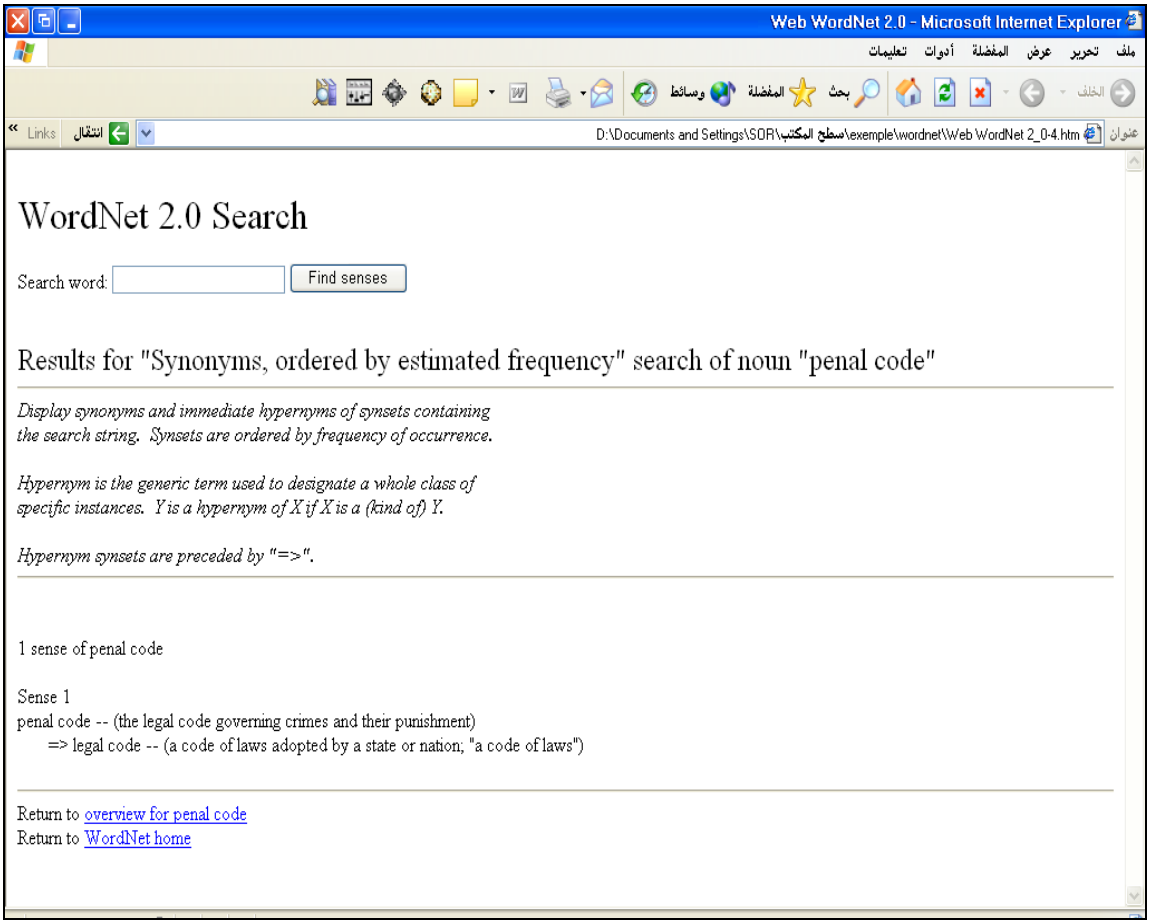



**Figure 36: recherche de "*penal code*" avec Wordnet**

Nous avons préféré étendre la requête avec la définition plutôt que les synsets, qui dans la majorité des cas nous éloigne de la requête initiale. Nous tenons à rappeler que Wordnet est une ontologie générique, et les synsets associés à chaque concept sont d'ordre général.
Nous avons soumis de nouveau la requête étendue cette fois ci au moteur et nous obtenons :

| Moteur utilisé | Nombre de pages | Nombre de documents pertinents sur les 20 premiers | Précision |
|---|---|---|---|
| Google | 185000 | 7 | 0.35 |

**Tableau 14: Recherche étendue avec Wordnet de "*penal code*"**

**Exemple8**: حقوق الطفل
La traduction avec Tarjim de Ajeeb a donné : *children Right*

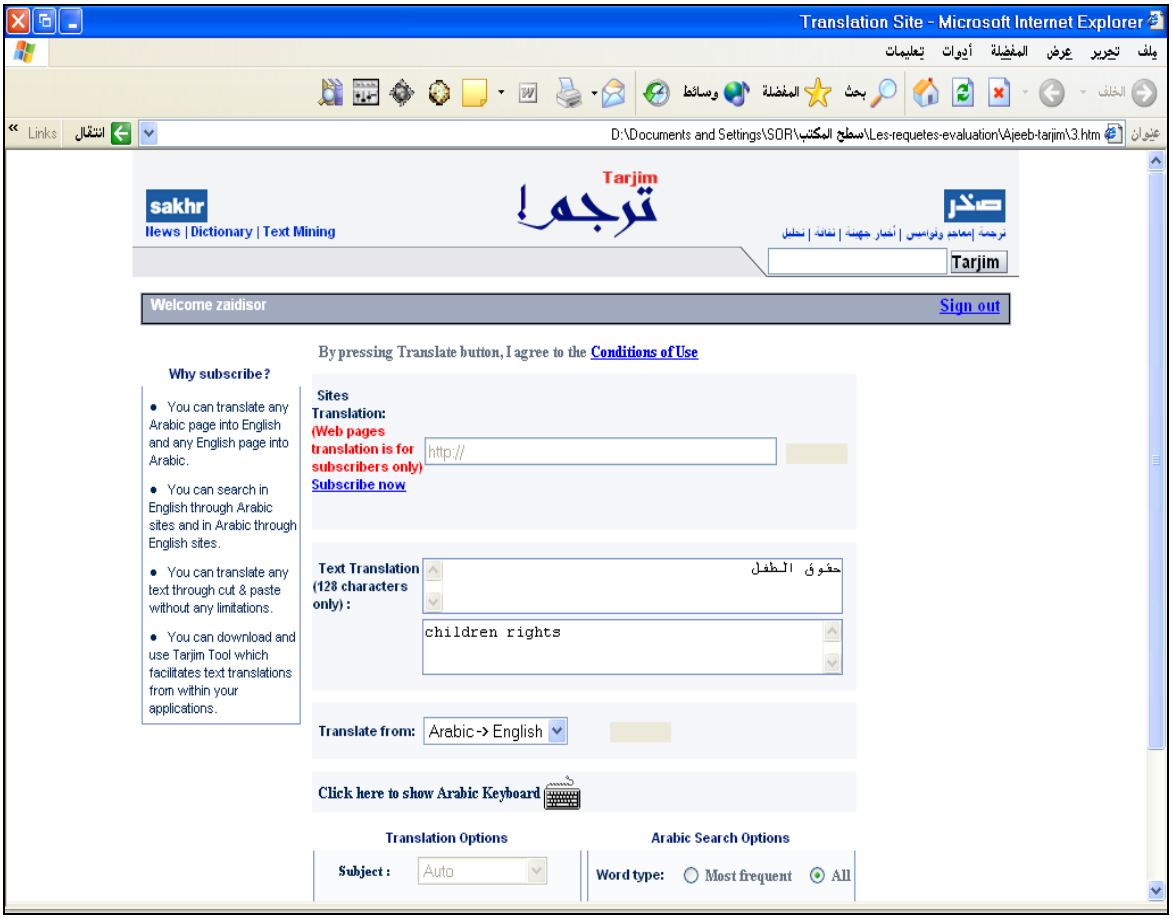



**Figure 37: traduction de la requete avec Tarjim de Ajeeb**

Pour étendre avec Wordnet, nous avons été obligé d'utiliser les définitions de chaque mot à part, parce que Wordnet n'a pu retourner une définition ou des synsets pour l'expression *children rights*.

Nous avons été obligé de chercher children puis Right séparément, nous avons ensuite réuni les deux définitions pour lancer la recherche avec la requête étendue.


**Figure 38: definition du mot *children* par wordnet**

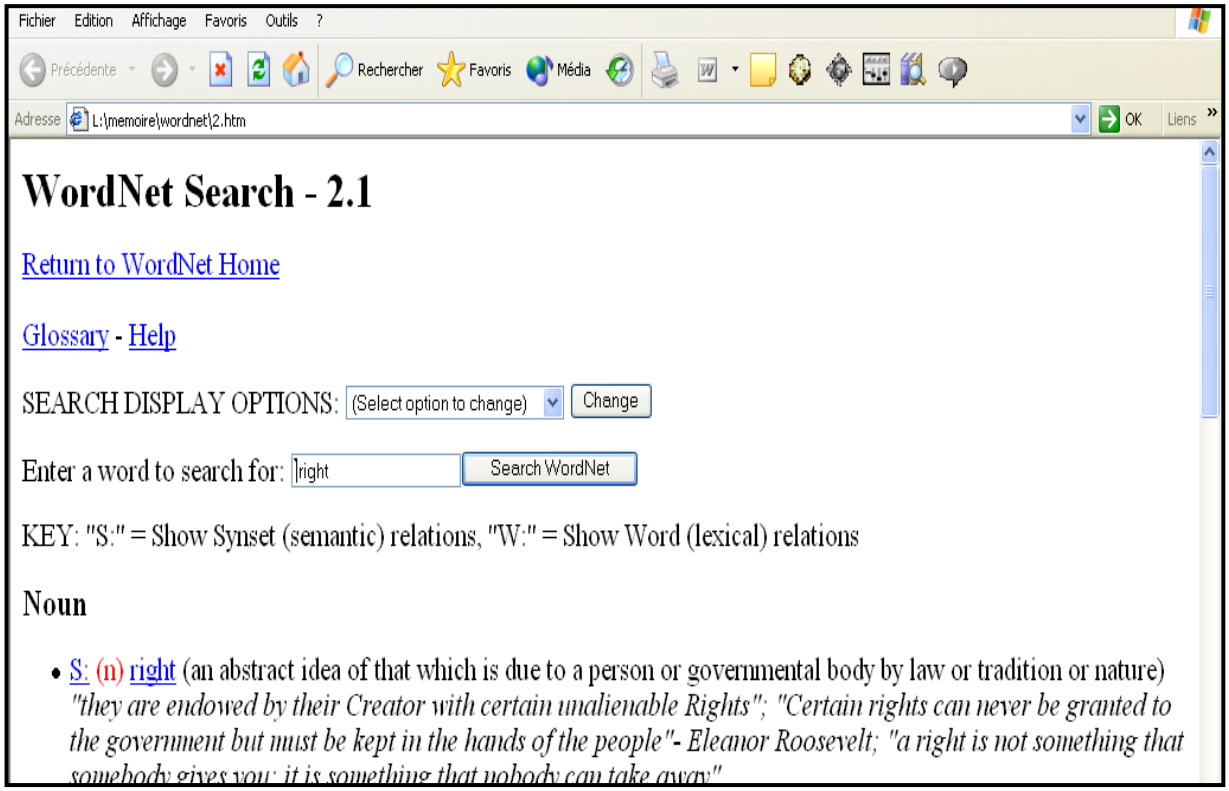



**Figure 39: définition du mot "Right" par Wordnet**

Nous utilisons les définitions données par Wordnet pour reformuler la requête initiale et nous obtenons:

| Moteur utilisé | Nombre de pages retournées | Nombre de documents pertinents sur les 30 premiers | Précision |
|---|---|---|---|
| Google | 119000 | 2 | 0.066 |

**Tableau 15: Recherche étendue de "*Children Rights*"**

## 4.5. Analyse et discussion

Nous avons décrit le processus d'expansion de la requête dans les exemples précédents nous résumons dans le tableau suivant les résultats obtenus dans les autres cas traités.

| **Type d'expansion** | **Précision moyenne**<br><br>**Avant expansion** | **Précision moyenne**<br><br>**Après expansion avec l'ontologie** | **Précision moyenne**<br><br>**Après expansion avec Wordnet** |
|---|---|---|---|
| Avec synonymes | 0.40 | 0.30 | 0.13 |
| Avec hyperonymes | 0.33 | 0.43 | |
| Avec attributs | 0.00 | 0.30 | |
| Avec définition | 0.80 | 0.55 | |
| Hétérogène | 0.20 | 0.60 | |

**Tableau 16:Tableau récapitulatif de la précision moyenne**

Pour résumer nous avons les résultats suivants:

| **Précision moyenne (synonymes exclus)** | | **Précision moyenne (synonymes inclus)** | | **Précision moyenne Après expansion avec Wordnet** |
|---|---|---|---|---|
| Avant expansion | Après expansion<br><br>Avec l'ontologie | Avant expansion<br><br>Avec l'ontologie | Après expansion<br><br>Avec l'ontologie | 0.13 |
| 0.33 | 0.47 | 0.34 | 0.37 | |

**Tableau 17: comparaison de la précision moyenne « avec » et « sans » synonymes**.

**Lecture des résultats:**
Nous avons essayé de travailler sur une vingtaine de requêtes, avec différente manière d'extension et de comparer les résultats des requêtes simples avec les requêtes étendues avec notre ontologie ainsi qu'avec Wordnet, après traduction de la requête en Anglais, avec toujours le même outil, en l'occurrence Tarjim de Sakhr Software.
Nous remarquons que :
Nous avons obtenu les meilleurs résultats avec l'expansion hétérogène, donc en ajoutant aussi bien les hyponymes, les hyperonymes que les slots.
Dans l'expansion avec synonymie en général, la précision c'est trouvée plus ou moins détériorée, ce qui confirme les résultats de Voorhees **[Voo94].**

Dans l'exemple5, le mot est très spécifique et n'est pas très polysémies, ce qui explique la précision élevée sans expansion, en lui rajoutant sa définition, qui contient des mots plus ou moins généraux la précision a diminué, d'un autre coté ce mot "الخلع" est très lié à la Chariaa ce qui explique la précision nulle, une fois étendu avec Wordnet.

Dans les autres cas de figure, la précision a connu une amélioration significative pour ne pas dire importante, exception faite pour certains mots très particuliers, remarquons que cela dépendra toujours de l'outil utilisé. Nous donnons cependant, l'exemple d'Alta Vista où nous pensons que le résultat est un peu aléatoire, d'après le travail que nous avons effectué nous pensons que AltaVista est moins bien adapté à la recherche en langue Arabe, nous avons obtenu des résultats trop aléatoires pour pouvoir conclure d'une façon objective.

Pour les requêtes étendues avec Wordnet, nous pensons que la faible précision face au taux élevé de pages retournées est due à l'ambiguïté de la traduction, à la non spécificité de la base et aux carences de Wordnet lui-même qui n'a pu retrouver la phrase children rights, en tant qu'expression.

Dans l'exemple2 la requête ne présente pas une grande ambiguïté. L'association des hyperonymes a permis de retourner des articles contenant les mots: مجلس الأعيان، مجلس الشيوخ، مجلس الشورى Qui sont d'autres noms de مجلس الأمة dans d'autres pays Arabes comme l'Arabie Saoudite la Jordanie etc..

Sur une vingtaine de requêtes nous avons constaté que le nombre de pages augmente dans l'Expansion par attribut et diminue nettement dans l'expansion par hyponymie due à la spécificité. Dans le cas général il y a une amélioration de la précision.

Concernant l'expansion avec Wordnet, Les résultats obtenus sont au deçà de nos attentes, ceci est essentiellement dû à la qualité de la traduction automatique et à l'expansion elle-même.
Dans Wordnet par exemple en traduisant محكمة عليا on obtient *suprem court*, Wordnet ne retrouve pas de synonymes, ce sont deux mots qui cooccurrent, et donc la recherche n'a pu aboutir.

L'inconvénient de la recherche booléenne est bien la conjonction et la disjonction des connecteurs qui rendent fastidieuse la prise en charge d'une requête avec une distributivité entre des *AND* et des OR, et donc nous pouvons avoir des cas où nous n'obtenons aucun résultats.

## Conclusion

Nous n'avons pas insisté sur l'expansion par synonymie. Les travaux précédents ont montré que dans le cas général, nous n'obtenons pas d'amélioration concernant la précision, les résultats obtenus au début ont confirmé cette hypothèse.
Enfin, nous avons effectué nos tests directement sur le Web, ceci nous rapproche beaucoup de la réalité, mais reste une tache périlleuse! Cependant nous aurions aimé confirmer nos résultats en travaillant sur une collection reconnue sur le plan international, ce qui demandera un travail énorme pour le choix des documents pour constituer une collection exhaustive et représentative, des outils standard de recherche pour pouvoir avancer des résultats sans aucune équivoque !

# 5. Conclusion et Perspectives

## Conclusion

Dans ce travail, nous avons abordé la problématique de la recherche d'information sur le Web en général, puis nous avons spécifié la recherche en Langue Arabe. Nous avons parlé des difficultés liées à cette recherche où nous avons insisté sur celles liées à la langue Arabe.
Dans l'état de l'art, nous avons cité un grand nombre de techniques proposées ou expérimentées pour pallier aux inconvénients d'une recherche simple, nous avons essayé dans la mesure du possible de donner les avantages et les faiblesses de chaque méthode proposée. Nous avons tiré celle que nous avons jugée meilleure, en l'occurrence l'expansion d'une requête à l'aide d'une ontologie, pour l'appliquer à la langue Arabe avec de menues modifications et nous avons essayé de l'expérimenter sur la recherche dans le domaine juridique.
Nous avons donc, commencer par l'étude des ontologies, ce qu'est une ontologie, à quoi sert-elle, comment la construire, par où commencer et avec quels outils?
Après seulement, nous avons entamé la construction de notre ontologie dans le domaine juridique, tache très épineuse et exténuante: domaine nouveau, termes étrangers à notre jargon informatique, nous avons été contraints de confronter rigueur de la technologie et subtilité de la nature humaine, en d'autres termes hémisphère gauche et hémisphère droit de notre cerveau, puis ce fut la confrontation avec les juristes experts, il fallait apprendre, comprendre, proposer puis discuter.

Nous avons donné dans la partie système proposé, la conception des différentes étapes suivies dans la recherche d'une information, en langue Arabe ainsi que l'éventualité de traduire une requête en Anglais ou en Français, mondialisation oblige!

Notre ontologie se compose de plus de 500 concepts entre classes et propriétés, elle peut être considérée comme un noyau pour toute autre ontologie touchant de près ou de loin le domaine juridique, elle peut même s'intégrer dans une ontologie générique et y constituer une partie importante.
Notre système, une fois réalisé avec toutes ses parties, peut aussi bien constituer l'élément principal dans l'architecture d'un moteur de recherche en langue Arabe, comme il peut être considéré comme un agent intelligent fournissant à un moteur de recherche une requête Arabe étendue.

La partie la plus importante de l'ontologie en l'occurrence le système judiciaire, qui représente la pierre angulaire de notre ontologie, a été validée par des spécialistes dans le domaine, traînant derrière eux des années d'expérience, pratiquant chaque jour sur le terrain. Maintes fois nous avons été obligés de remettre en question la structure de la hiérarchie des classes, ainsi que leurs différentes propriétés. Il nous reste d'automatiser la recherche de concepts dans l'ontologie, pour l'expansion de la requête, chose qui ne doit pas présenter de difficultés puisque nous pouvons générer notre ontologie sous format HTML ou sous format texte, de plus avec Protege2000 nous pouvons générer une représentation sous format RDF, donc à base de XML.
La partie présente dans la conception, mais qui n'a pas été réalisée et qui peut être l'objet d'un autre projet de recherche, c'est l'analyse statistique pour la détermination des variantes les plus représentatives du domaine juridique et leur association aux concepts de l'ontologie. Cette étape nécessite le concours d'experts linguistes et elle est très importante ! Même si nous ne sommes pas entrés dans les rouages de son fonctionnement interne, car elle représente un moyen efficace pour la

réduction du bruit et l'amélioration de la précision, en plus de l'adjonction aux termes de la requête des hyperonymes, des hyponymes ou des attributs.
Les résultats obtenus prouvent qu'il y a une différence plus ou moins signifiante entre une requête simple et une requête étendue, malgré les insuffisances de l'ontologie et du système en entier, un travail complémentaire donnerait sûrement d'autres éléments intéressants à considérer et à étudier.

Pour conclure, ce travail peut être exploité dans différents travaux concernant le domaine juridique aussi bien la recherche sur le Web que la traduction automatique ou la compréhension du langage naturel ou tout autre travail touchant la sémantique des mots et ce dans le domaine légal.

## Perspectives

En perspective, nous comptons valider le reste de l'ontologie, de la compléter avec l'adjonction des définitions, des synonymes, des variantes morphologiques ainsi que de nouveaux concepts si nécessaire.
Nous pensons à la mise au point d'un outil de recherche des variantes les plus utilisées dans le domaine juridique. Cet outil se basera sur un corpus le plus exhaustif et le plus représentatif possible du domaine cité, il cherchera pour chaque concept de l'ontologie les variantes les plus utilisées, et les attachera au dit concept. Quand une requête est soumise, les mots clés sont recherchés dans l'ontologie, des que nous trouvons un concept nous lui rajoutant les variantes préalablement attachées, en plus des hyponymes et hyperonymes. Si la recherche est à haute précision, nous privilégierions l'adjonction des hyponymes, si par contre elle est à haut rappel, nous prendrons les hyperonymes.
Dans le cas où l'outil de navigation dans l'ontologie ne trouve pas le concept recherché, nous pensons à développer une interface qui guidera l'utilisateur pour la correction de sa requête si elle est erronée, de donner des suggestions ou recevoir des propositions de la part d'un utilisateur potentiel, qui peut, aussi bien et selon certaine conditions, contribuer à l'enrichissement de l'ontologie avec de nouveaux concepts.
Pour un travail futur nous pensons à faire le mapping entre les langues puisque l'ontologie est en langue Arabe, il serait intéressant de considérer une table d'index qui, pour chaque concept en Arabe donnera son équivalent en Français ou en Anglais et en déduire des ontologies parallèles.
Nous pouvons aussi utiliser l'ontologie pour la traduction automatique de la requête, cela peut présenter aussi le travail d'un autre projet.
La même ontologie, peut aussi être utilisée dans l'indexation des documents, ou leur catégorisation pour faciliter la recherche d'information.
Enfin nous espérons qu'une quelconque partie donnerait une suite positive à notre travail, pour la contribution dans le développement de ce que nous appellerons désormais la e-justice dans le monde Arabe.

# Références et bibliographie


**1 [Abd04]** A. Abdelali, J. Cowie, H.S. Soliman *Arabic information retrieval perspectives* JEP-TALN 2004, Arabic language processing, Fez, 19-22 april 2004 .

**2 [Abu92]** H. Abu Salem *A microcomputer based Arabic bibliographic information retrieval system with relation thesaurus (Arabic-IRS),* Ph.D.Thesis Chicago Illinois Institute of Technology 1992.

**3 [AlJ01]** M. Aljlayl, O. Frieder. *Effective Arabic-English Cross-Language Information Retrieval via Machine-Readable Dictionaries and Machine Translation CIKWOI,* November S-10,2001, Atlanta, Georgia, USA. Pages 295—302.

**4 [AlK91]** I. Al Kharashi. *A microcomputer-based Arabic information retrieval system comparing words, stems, and root as index terms*, Ph.D.Thesis.Chicago Illinois Institute of Technology 1991.

**5 [AlN89]** F. Al Naim *text analysis and automatic indexing for Arabic based automated information retrieval system*. MSc Thesis, Chicago California State University 1989.

**6 [AlT98]** M. S. Al Tayyar, K. Bechkoum. *The Effectiveness of Morphological Analysis for Text Retrieval in Arabic,* 6th International Conference on Multi-lingual Computing, Cambridge, UK, 17-18 April 1998.

**7 [AlT00]** M. S. Al Tayyar. *Arabic information retrieval system based on morphological analysis (AIRSMA) a comparative study of word, stem, root and morpho-semantic methods*. Ph. D. Thesis DeMonfort University 2000.

**8 [Aus02]** N. Aussennac-Gilles *GT documents Multimédias*, 08 02 2002.

**9 [Bae99]** R. Baeza-Yates, B. Ribeiro-Neto, « *Modern information retrieval* », ACM Press books, Addison-Wesley, 1999, ISBN-0-201-39829-X.

**10 [Bat90]** J. A. Batemane et al. *A general organization of knowledge for natural language processing: the PENMAN upper model*, Technical report, USC/Information Sciences Institute, Marina del Rey, California.1990.

**11 [Ben99]** M. Ben Henda. *L'indexation par éléments méta dans le processus de référencement du texte arabe entre HTML4, Unicode et Dublin Core*.- in : colloque ISKO, Lyon (France), 21 - 22 octobre 1999.- pp. 105 - 111.

**12 [BLA04]** W. J. Black, S. El-Kateb: *A Prototype English-Arabic dictionary based onWordNet* In Proceedings of 2nd Global WordNet Conference, GWC2004, Czech Republic: 67-74 (www.globalwordnet.org/AWN/meetings/GWA).

**12 [Bla03]** G.Blain, *Ontologie et droit* (www-etud.iro.umontreal.ca/ ~blaingu/travaux/ontology/index.html)

**13 [Bou02]** C. Boudry *Typologie et mode de fonctionnement des outils de recherche d'information sur internet en biologie/médecine* parue dans MEDECINE/SCIENCES 2002 ; 18 : 616-22

**14 [Bus45]** V. Bush As We Think *Atlantic Monthly, 176:101-108, July 1945.*

**15 [Buc95]** C. Buckley, J. Allan, G. Salton, A. Singhal. *Automatic query expansion using SMART*: TREC 3. In Proceedings of the Third Text REtrieval Conference (TREC-3), pages 69–80. NIST Special Publication 500-225, April 1995.

**16 [Bus05]** D. Buscaldi et al. *A WordNet-based Query Expansion method for Geographical Information Retrieva*l, in Proceedings of GeoCLEF 2005.

**17 [Cal04]** L. Calcagno et al. Comparison of Indexing Techniques based on Stems, Synsets, Lemmas and Term Frequency. In: Workshop "Red Tem´atica en Tecnolog´ıa del Habla, Valencia, Spain 2004, pp. 171-176.

**18 [CAR04]** Cariboo. CIRCA : la technologie d'Applied Semantics au coeur des Adwords et des Adsense de Google [2] septembre2004 (http://www.Google.com).

**19 [Cha97]** J.Y. Chai, A. Biermann. *The use of lexical semantics in information extraction*. In Proceedings of the Workshop in Automatic Information Extraction and Building of LezicalSemantic Resources, 1997 pages 61-70.

**20 [Cla03]** V.Claveau. *Acquisition automatique de lexiques sémantiques pour la recherche d'information*. Thèse de doctorat, Université de Rennes 1, France. 2003.

**21 [Cla04]** V. Claveau, P. Sebillot. *Extension de requêtes par lien sémantique nom-verbe acquis sur corpus*. In proceedings of TALN 2004, Fès Maroc 2004.

**22 CLEF –** Cross language Evaluation Forum (http: //www.clef-campaign.org/)

**23 Computer Guide** . IRSAD 1993. pp1, 39 et 45

**24 [Cle67**] C. W. Cleverdon. *The Cranfield tests on index language devices*. Aslib Proceedings, 19:173–192, 1967

**25 [Das02]** S. Das, K. Shuster, C. 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, July 2002.

**26 [Deb04]** E. Debanne Query *by Example for Symbolic Still Image Retrieval*. in Première Conférence en Recherche d'Information et Applications - CORIA'04, Toulouse, pp363-376, 10-12 Mars, 2004.

**27 [Dui99]** Duineveld A., Studer R., Weiden M., Kenepa B., Benjamins R. *WonderTools? A comparative study of Ontological engineering tools.* In Proceedings of KAW99. Banff, Canada.

**28 [Dun93]** Dunning, T. and Davis, M. Multi-lingual information retrieval. Technical Report MCCS-93-252. Computing Research Laboratory, New Mexico State University. 1993.

**29 Dz**-code Berti-editions Alger 2000

**30 [EDC01]** Egyptian Demographic Center.http://www.ficu.eun.eg/wwwihomepage/cdc/cdc.htm 2001.

**31** [**ElK 04]** M. El Kourdi, A. Bensaid, T. Rachedi. *Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm* 20th International Conference on Computational Linguistics August 28th Geneva 2004**.**

**32 Encyclopedie** juridique 2000 Dar El Hillel ( الموسوعة القضائية 2000 دار الهلال للخدمات الإعلامية)

**33 [Fag89]** J. L. Fagan. *The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieva*l. Journal of the American Society for Information Science, 40(2):115–139, 1989.

**34 [Far96]** A. Farquhar R. Fikes J. Rice *The Ontolingua Server: a Tool for Collaborative Ontology Construction* Knowledge Systems Laboratory Stanford University.

**35 [Fel98]** C. Fellbaum.. *WordNet:An Electronic Lexical Database.* The MIT Press, Cambridge, MA 1998.

**36 [FGDC]** Federal Geographic Data Committee, *Content standard for digital geospatial metadata.* (http://www.fgdc.gov/metadata/metadata.html).

**37 [Fur01]** F. Furst *Opérationalisation d'une ontologie de la géométrie projective* Rapport de stage de recherche DEA Informatique 2000-2001 , Université de Nantes.

**38 [Gal91]** W. A. Gale, K.Church. *Identifying word correspondences in parallel texts* Proceedings of the 4th DARPA Speech and Natural Language Workshop, (1991). P.152-157.

**39 [Gan99]** B. Ganter, R. Wille *Formal Concept Analysis: mathematical foundation.* Springer, Berlin-Heidelberg 1999.

**40 [Gat00]** M. Gatius i Vila. ArchiWordNet*: Integrating WordNet with Domain-Specific Knowledge :Using an Ontology for Guiding Natural Language Interaction with Knowledge Based Systems*(Barcelona, septembre 2000).

**41 [Gom97]** J.M. Gomez-Hidalgo, M. Rodriguez. *Integrating a lexical database and a training collection for text categorization.* In Proceedings of the Workshop in Automatic Information Extraction and Building of Lexical Semantic Resources, 1997.pages 39-44.

**42 [Gru94**] T. Gruber *what is an ontology* ? (http://ksl- web.stanford.edu/people/gruber)

**43 [Gru95]** T. Gruber. T*oward Principles for Design of Ontologies Used for KnowledgeSharing*, International Journal of Human and Computer Studies, 43 (5/6): 907-928

**44 [Gua96] N.** Guarino *Understanding, Building, And Using Ontologies 5 Oct 1996* LADSEB-CNR, National Research Council Corso Stati Uniti 4, I-35127 Padova, Italy.

**45 [Gua98]** N. Guarino. *Formal Ontology and Information Systems*, In Proceedings of the 1st International Conference, Trento,Italy, June 1998, IOS Press, Amsterdam, pp 3-15

**46 [Gua99]** N. Guarino, C. Masolo, G Vetere. *OntoSeek: Content-Based Access to the Web*, IEEE Intelligent Systems, May/June 1999, pp 70-80

**47 [Haa00]** H-M. Haav, J. F. Nilsson. *Approaches to Concept Based Exploration of Information Resources*, W. Abramowicz, J. Zurada (Eds), Knowledge Discovery for Business Information Systems, Kluwer Academic Publishers, 2000, ch 4, pp 89-111

**48 [Haa01]** H. Haav, T. LUBI. *A Survey of Concept-based Information Retrieval Tools on the Web.* (www.science.mii.lt/ADBIS/local2/haav*)*

**49 Hahooa** (http://www.hahooa.com/nav.php?ver=ar)

**50 [Har93]** D. K. Harman. *Overview of the first Text REtrieval Conference* (TREC-1). In Proceedings of the First Text REtrieval Conference (TREC-1), pages 1–20. NIST Special Publication 500-207, March 1993.

**51 [Has96]** A. Hasnah. *Full Text Processing and Retrieval: Weight Ranking, Text Structuring, and Passage Retrieval for Arabic Documents.* C.S. Ph.D. Dissertation, Illinois Institute of Technology, 1996.

**52 [ Hme95]** I-I. Hmeidi *Design and implementation of automatic word and phrase indexing for information retrieval with Arabic documents.* Ph.D. Thesis. Chicago Illinois Institute of technology.

**53 [Haw00]** D. Hawking. *Measuring the quality of public search engines*, 2000. pastime˙anu˙edu˙au/TAR/Search Engines Conf/.

**54 [Haw03]** D. Hawking *Very Large Scale Information Retrieval* S. Renals, G. Grefenstette (Eds.): Text- and Speech-Triggered Info. Access, LNAI 2705, pp. 106-144, 2003. Springer-Verlag Berlin Heidelberg 2003

**55 [Hul96]** D. Hull, G.Grefenstette. *Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval*. In proceedings of the 19 annual international ACM SIGIR 1996, Zurich, Switzerland, p 49-57.

**56 [ISO03]** *The dublin core metadata element set.*- ISO TC46/ SC 4 N 515.-26-02-2003. (http://www.niso.org/international/SC4/ )

**57 [Jes00]** B. Jesen. et al. *Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web.* Information Processing and Management, 36(2), 207–227. 2000

**58 [Jen03]** O. Jenhani Ontologies *pour le WEB: relations, construction d'ontologies et méthodes de raisonnement pour la génération de langue naturelle*. Rapport Réalisé par Olfa Jenhani INRIA- ARC GeNI Mai 2003.

**59 [kas88]** N. Kassem "خصوصيات الاسماء و الصفات العربية و أثرها في خزن المعلومات و استرجاعها" = particularities of Arabic nouns and adjectives and their effect in information storage and retrieval, A'adab Almostanseeriah Journal, 16, pages :705-737

**60 [Kow97]** G. Kowalski *Information Retrieval Systems - Theory and implementation* Kluwer Academic Publishers, 1997, ISBN-0-7923-9926-9.

**61 Le code penal** (2002 الديوان الوطني للأشغال التربوية وزارة العدل : قانون العقوبات ) 2002.

**62 Lexalgeria**: Le Portail du droit algérien (http://www.lexalgeria.net/)

**62 [Luh57]** H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1957.

**63 [Man97]** R. Mandala, T. akenobu, T. Hozumi *The Use of WordNet in Information Retrieval* In COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, 1997.

**64 [Mar60]** M. E. Maron, J. L. Kuhns. *On relevance, probabilistic indexing and information retrieval. Journal of the ACM*, 7:216–244, 1960.

**65 [Mic99]** A. Michard. *XML langage et applications.* Paris : Eyrolles, 1999.- 361p.

**66 [MIL90]** G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller. *Introduction to WordNet: an on-line lexical databas*e. In International Journal of Lexicography 3 (4), 1990, Revised August 1993 - accessible at (ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps)

**67 [Morf91]** A. H. Morfeq *Bayan: a text management system for Arabic engineering documents.* Ph.D. Thesis. Colorado: Colorado University 1991.

**68 [Morr91]** J. Morris, G. Hirst. *Lexical cohesion computed by thesaural relations as an indicator of the structure of text.* In *Proceedings of A CL Conference,* 1991. pages 21--45.

**69 [Noy00a]** N.F. Noy, M.A. Musen. *PROMPT: Algorithm and tool for Automated Ontology Merging and Alignement.* In seventeeth National Conference on Artificial Intelligence(AAAI-2000). Austin, TX, 2000.

**70 [Noy00b]** N. Noy, D. McGuinness **A** *Guide to Creating Your First Ontology* (http://protege.stanford.edu/publications/ontology Development/ ontology101.html**)**

**71 NTCIR:** NACSIS *Test Collection for Information Retrieval* (http://www.rd.nacsis.ac.jp)

**72 [Oar98]** D. Oard *A Comparative Study of Query and Document Translation for Cross-language Information Retrieval.* In Machine Translation and the Information Soup. 3rd Assoc. for Machine Transl. in the Americas Conf., 472-83, 1998.

**73 [Oue03]** T. Ouerfelli. *La description des documents électroniques diffusés sur le web : pour une recherche pertinente* Colloque International Francophone en Sciences de l'Information et de la Communication (Bucarest 28 juin - 2 juillet 2003)

**74 [Pea91**] H. Peat, P.Willett. *The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems*. Journal of the American Society for Information Science, 42(5), (1991). p378–383.

**75 73 [Pet01]** C.Peters, P. Sheridan. *Multilingual Information Access*. In M. Agosti, F. Crestani, Lectures on G. Pasi (eds.).Information Retrieval, Lecture Notes in Computer Science 1980, pp51-80, Springer Verlag, 2001.

**76 [Pir98]** A. Pirkola. *The Effects of Query Structure and Dictionary Setups in a Dictionary-based Cross-Language Information Retrieval.* SIGIR 1998, Melbourne, Australia.

**77 POGAR** ( Programme On Governance in Arab Region) UNDP(united nation development programme) (http://www.undp.org)

**78 Pro**tege2000 ( version 3.0 build 141) http://protege.stanford.edu.

**79 [Qiu93]** Y. Qiu, H. P. Frei. *Concept based query expansion. Research and Development in Information Retrieval*, ACM-SIGIR, (1993). p160-169.

**80 [Qiu95]** Y.Qiu, H.-P. Frei *Improving the Retrieval Effectiveness by a Similarity Thesaurus*. Rapport interne 225, ETH Zurich, Department of Computer Science, Zurich, Suisse1995.

**81 [Rac03]** T. Rachidi, et al. *Barq : "distributed multilingual internet search engine with focus on Arabic language"* In proc of IEEE conf. On sys., Man and Cyber., Washington DC, Oct.5-8, 2003.

**82 [Ras04**] F. Rastier. *Ontologies* C.N.R.S. (Article paru dans la *Revue des sciences et technologies de l'information*, série : *Revue d'Intelligence artificielle*, 2004, vol.18, n°1, p15-40.

**83 [Res95]** P.Resnik. *Disambiguating noun grouping with respect to wordnet senses*. In *Proceedings of 3rd Workshop on Very Large Corpora* 1995.

**84 [Ric95]** R. Richardson, A.F. Smeaton. *Using wordnet in a knowledge-based approach to information retrieval.* Technical Report CA-0395, School of Computer Applications, Dublin City University. 1995.

**85 [Rob77]** S. E. Robertson. *The probabilistic ranking principle in IR*. Journal of Documentation, 33:294–304, 1977.

**86 [Roc71]** J. J. Rocchio. *Relevance feedback in information retrieval*. In Gerard Salton, editor, *The SMART Retrieval System—Experiments in Automatic Document Processing*, pages 313–323, Englewood Cliffs, NJ, 1971.Prentice Hall, Inc.

**87 [Roc03]** C. Roche *La construction d'ontologies: Quel constat?* EGC 2003 Lyon - 22-23-24 janvier 2003.

**88 [Roui00]** J. Rouillard *Les enjeux d'un dialogue Homme-Machine sur Internet - L'Hyperdialogue.* Bulletin d'informatique approfondie et applications, revue de l'université de Provence, 52, France, 3-20. 2000.

**88 [Sad01]** F.Sadat, A.Maeda, M.Yoshikawa, S.Uemura. *Cross-Language Information Retrieval via Dictionary-based and Statistical-based Methods* Proceedings of the 2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM'01), 2001.

**89 [Sad02**] F. Sadat. *Cross-Language Information Retrieval via Hybrid Combination of Query Expansion Techniques.* In Proceedings of the Association for Computational Linguistics ACL-02 Student Research Workshop, Philadelphia, USA. 2002.

**90 [Sakhr 97]** Sakhr software compagny . Integrated Information Management system. Cairo.

**91 [Sakhr04**] Sakhr software compagny (www.sakhrsoft.com) 2004.

**92 [Sal71]** G. Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall inc. Englewood Cliffs, NJ, 1971.

**93 [Sal83]** G. Salton, « *Introduction to Modern Information Retrieval* », McGraw-Hill, 1983.

**94 [Seg97]** F. Segond, A. Schiller, G. Grefenstette, J. Chanod. *An experiment in semantic tagging using hidden markov model tagging.* In Proceedings of the Workshop in Automatic Information Extraction and Building of Lexical Semantic Resources, 1997 p78-81.

**95 [She96]** P.Sheridan, J.P. Ballerini. *Experiments in Multilingual Information Retrieval using the SPIDER System.* The 19th Annual International ACM SIGIR 1996, 58-65.

**96 [She97]** P.Sheridan, M. Braschler, P. Schäuble. *Cross-Language Information Retrieval in a Multilingual Legal Domain.* In ECDL'97 Proceedings, Pisa, Italy, p253–268, 1997

**97 [Sme95]** A.F. Smeaton, C. Berrut. *Running tree-4 experiments: A chronological report of query expansion experiments carried out as part of tree-4.* Technical Report CA-2095, School of Comp. Science, Dublin City University 1995.

**98 [Sta97]** M.A. Stairmand. *Textual context analysis for information retrieval*. In roceedings of the 2Oth ACM-SIGIR Conference, 1997 p140-147.

**99 [Str97]** T. Strzalkowski, L. Guthrie, J. Karlgren, J. Leistensnider, F. Lin, J. Perez-Carballo, T. Straszheim, J.Wang, J. Wilding. *Natural language information retrieval*: TREC-5 report. In Proceedings of the Fifth Text REtrieval Conference (TREC-5), 1997.

**100 [Tai04]** O.M. Tailliez *les ontologies* cours de DEA I3 module de Fouille de texte 4/02/2004**.**

**101 [Tay90]** M.Tayli, A. Al-Salamah. *Building Bilingual Microcomputer Systems* In Communications of the ACM, Vol. 33, No.5, 1990 p495-505.

**102 [Ten90]** C. Tenopir, J. Soon Ro. *Full text databases*. New York. GreenWood Press.

**103 The Verity** K2 Discovery Tier *The Importance of Advanced, Effective Search Tools* Verity, Inc. 894 Ross Drive, Sunnyvale 2003. (www.verity.com)

**104 TREC** - *Text Retrieval Conference Series* (http://trec.nist.gov/)(sheridan72)

**105 [Van79]** C.J. Van Rijsbergen. *Information Retrieval* Buttherwords, London, 1979.

**106 [Voo93]** E.M. Voorhees. *Using wordnet to disarnbiguate word senses for text retrieval*. In Proceedings of the 16th A CM-SIGIR Conference, 1993. p171-180.

**107 [Voo94]** E.M. Voorhees. *Query expansion using lexical-semantic relations*. In Proceedings of the 17th ACM-SIGIR Conference, 1994. p61-69**.**

**108 [Vos98]** P. Vossen, EuroWordNet: A Multilingual Database with Lexical Semantic Networks (University of Amsterdam) [Reprinted from *Computers and the Humanities*, 32(2-3), 1998]Dordrecht: Kluwer Academic Publishers, 1998, 179 pp; hardbound, *Reviewed by Graeme Hirst University of Toronto* WordNet, the on-line English thesaurus.

**109 [Vos99]** P.Vossen *Building a multilingual database with wordnets for several European languages*. 1999 (http://www.hum.uva.nl/~ewn).

**110 [Wan01**] B. Wang, H. Xu, Z. Yang, Y. Liu, X. Cheng, D. Bu, S. Bai *Filtering, Web and QA* TREC-10 Experiments at CAS-ICT Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China 2001.

**111 [Win87]** E.M.Winston. *Taxonomy of Part-Whole Relations,* COGNITIVE SCIENCES 11, 417-444. 1987.

**112 [Zai05a]** S. Zaidi, M.T. Laskri K. Bechkoum. *A Cross-language Information Retrieval Based on an Arabic Ontology in the Legal Domain* In Proceedings of the international conference on signal-image technology & internet–based systems, IEEE SITIS.Yaoundé November 27th - December 1st 2005. p86-91**.**

**113 [Zai05b]** S.Zaidi, M.T. Laskri *A Cross-language Information Retrieval Based on an Arabic Ontology in the Legal Domain* , In Proceedings of The International Arab Conference on Information Technology (ACIT) December 6th- 8th, 2005 Jordan.

**114 [Zai05c]** S.Zaidi, M.T. Laskri**.** *Expansion de requête à l'aide d'une ontologie en Arabe***,** In Proceedings of International Workshop on « Text, Image & Speech Recognition » TISR'05 12-13 December 2005 Annaba. P186-192.

**115 [Zai06a]** S.Zaidi, M.T. Laskri**.** *Expansion de requête basée sur une ontologie Arabe pour une recherche interlingue dans le domaine juridique* journées scientifiques du groupe de recherche en Intelligence Artificielle, JGRIA avril 2006 Annaba.

**116 [Zai06b]** S.Zaidi, M.T. Laskri**.**Une recherche interlingue sur le web basée sur une ontologie Arabe du domaine juridique**,** Ds Proceedings du Séminaire National en Informatique SNIB 2-4 mai 2006 Biskra

**117 [Zai06c]** S.Zaidi, M.T. Laskri**.**
*"أداة ما بين اللغات للبحث على الانترنت بالستعمال أنطولوجيا باللغة العربية في مجال القانون"*
الملتقى الوطني الرابع حول الذخيرة العربية 9، 10 ماي. عنابة 2006 .

**117 [Zem02]** M. Zemouli Etajerib Erahina haoula haoussabet Enoussous Elati Taâtamid Elougha ElArabia (التجارب الراهنة حول حوسبة النصوص التي تعتمد اللغة العربية)= expérimentations actuelles sur l'informatisation des textes écrits en langue Arabe paru dans (magazine de la langue Arabe N° 7 2002 par le haut conseil de la langue Arabe) (www.Csla.dz).

# Annexes

Annexe A : Le questionnaire proposé aux étudiants, enseignants, juristes…

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Badji Mokhtar Annaba

Faculté des sciences de l'ingénieur

Département d'informatique

## Analyse de l'Existant

Fonction :……..

1) Utilisez-vous Internet ? Régulièrement ☐ Occasionnellement ☐

Rarement ☐ Jamais ☐

2) Généralement vous recherchez quoi ?

Informations générales ☐

Informations liées à votre profession ☐

Autres (images, musique…) ☐

3) Vous utilisez quel Moteur de recherche ?……

4) Combien de mots clés ? …….. .

5) Etes-vous satisfait des résultats retournés ?……

6) A votre avis quel est le problème ?……

7) Pensez-vous que l'aide apportée par Internet est considérable pour vous ?……

8) Avez-vous des suggestions?……