République Algérienne Démocratique et Populaire

Ministère de l'enseignement Supérieur
Université d'Alger
Faculté des Lettres et Langues
Département des Sciences de la Langue

DEVELOPPEMENT D'UNE ONTOLOGIE DE DOMAINE POUR LE JOURNAL OFFICIEL EN LANGUE ARABE

Mémoire pour l'obtention d'un diplôme de Magister

Spécialité **T**raitement **A**utomatique de la **L**angue

Présenté par :
BACHIR BOUIADJRA Soumya

Dirigé et proposé par :
Dr. ALIANE Hassina
Pr. ALIMAZIGHI Zaïa

Soutenue le : / /

Devant le jury composé de :

Dr. SALMI Abdelamadjid,	Université Alger2,	Président ;
Dr. BOUKHALFA Kamal,	USTHB,	Examinateur;
Dr. ALIANE Hassina,	USTHB, CERIST,	Encadreur;
Pr. ALIMAZIGHI Zaïa,	USTHB,	Encadreur.

2012

Remerciements

Je tiens à adresser en premier lieu mes plus chaleureux remerciements à Madame la professeure Hassina ALIANE et Madame la professeure Zaïa ALIMAZIGHI de m'avoir proposé ce sujet et de m'avoir encadré ainsi que pour leur compréhension.

Mes sincères remerciements s'adressent à Monsieur le professeur Abdelmadjid SALMI et à Monsieur le professeur Kamal BOUKHALFA membres du jury pour l'honneur qu'ils me font en acceptant de juger ce travail.

Je tiens à exprimer ma profonde gratitude à Mr Ahmed HEMRI et Mr Khaled BAAZI pour l'intérêt et la disponibilité qu'ils ont manifestés tout au long de ce travail.

Je remercie vivement Mr Abdelhamid HACHANI pour son soutien et sa patience.

Dédicaces

À mes très chers parents, mon mari et mes enfants

Khalil, Zakaria et Meriem

Que Dieu les garde.

À toute ma famille.

À mon cher pays.

Je dédie ce modeste travail.

RESUME :

Le Journal Officiel est une publication officielle qui diffuse par principe obligatoirement les textes juridiques (lois, décrets,…), ainsi que d'autres informations juridiques officielles ; l'informatisation de ces données quant à elle a connu plusieurs étapes et plusieurs réalisations logicielles.

Les ontologies informatiques, phénomène relativement récent. Les possibilités que paraît ouvrir la création d'une ontologie sont nombreuses et tentantes : L'amélioration de la recherche documentaire, l'automatisation de l'indexation ou encore l'accessibilité et la visibilité du produit sur le Web.

Le présent travail consiste alors à proposer une approche ontologique pour la modélisation sémantique, l'indexation et l'interrogation des différents textes du Journal Officiel. Cet objectif est motivé par le besoin d'avoir un modèle qui facilite la réutilisabilité et la recherche sémantique dans ces textes.

De ce fait, il n'existe aucune ontologie en relation avec ces différents textes officiels ; et pour cela nous avons choisi d'exposer les différentes étapes de construction de l'ontologie en relation avec le contenu du Journal Officiel ;

Il s'agira donc dans ce modeste travail de suivre pas à pas la conception d'une telle ontologie, le choix a été fait de procéder en deux temps, le premier théorique où nous découvrirons ce qu'est une ontologie et comment elle se construit ; et le second pratique dans lequel nous mettrons la théorie en application à travers la réalisation de la mission : développement d'une ontologie de domaine pour le Journal Officiel en langue arabe. Où nous allons procéder à étudier le Journal Officiel et ses approches informatiques. Cela révèle le besoin d'un modèle sémantique qui peut être réalisé à l'aide d'une approche ontologique.

Mots clés
Le Journal Officiel ; Ontologie de domaine ; Web sémantique ; W3C ; Ingénierie des connaissances ; Editeur ; Protégé ; OWL.

Tables des matières

Liste des figures et tableaux

1. Figures :

2. Tableaux:

Introduction

Le Journal Officiel est, sous sa forme actuelle, un service spécialisé du Secrétariat Général du Gouvernement, dont la mission est d'assurer la publication et la diffusion des textes législatifs et réglementaires pris par les autorités compétentes conformément à la Constitution, la publication et la diffusion des actes de procédure, des actes de sociétés, d'associations et de protêts, des partis politiques, des dessins et modèles industriels, des marques de fabrique, de commerce et de service ainsi que tout autre acte visé par la Loi, la mise à jour et la coordination des textes législatifs et réglementaires.

Il constitue la référence de base pour toutes les sciences Juridiques.

Le Journal Officiel en Algérie paraît une ou plusieurs fois par mois. Les publications du Journal Officiel sont destinées au public, aux animateurs des institutions, aux magistrats, aux avocats, aux autres praticiens du droit, aux étudiants et aux chercheurs. Les publications du Journal Officiel sont portées à la connaissance du public par le support papier et par la mise en ligne ou sur cd-rom sous format PDF.

Le Journal Officiel est édité en langue arabe et il est traduit en français.

Le modèle JORADP développé et publié par la Direction du Journal Officiel, du Secrétariat Général du Gouvernement, ne permet pas une recherche par concept c'est-à-dire recherche sémantique, par exemple par synonyme dans le contenu du Journal Officiel, pour cela nous nous intéresserons à l'élaboration d'un modèle sémantique des textes juridiques qui facilitera la recherche d'une manière sémantique dans le contenu du Journal Officiel.

Nous avons besoin alors d'un modèle sémantique qui permet de modéliser l'aspect sémantique des différents textes du Journal Officiel, qui facilite la recherche par concept, l'indexation, la communication et la réutilisation de ces textes entre différentes applications informatiques autour du Journal Officiel.

Pour réaliser un tel modèle sémantique, un ensemble d'interrogations est soulevé : existe-t-il déjà des modèles sémantiques pour le Journal Officiel en Algérie ? Quelle est la technologie à

utiliser pour une telle réalisation ? Comment modéliser des index pour servir la recherche par concept ?

La technologie utilisée repose sur la construction d'une ontologie pour le Journal Officiel et l'établissement des liens entre ses concepts. Cette ontologie est nécessaire aussi pour les sciences juridiques, car elle nous permet d'élaborer une description plus contrôlée et plus cohérente du domaine en général et de ses sous-domaines aussi. Ceci va faciliter l'apprentissage de cette science et permettre de contrôler l'usage des relations entre les concepts du même domaine. De même, cette description pourra établir les équivalents entre les termes des différentes langues, élaborer les définitions de façon systématique, cohérente et précise, accéder plus facilement à l'information et créer de termes nouveaux mieux adaptés aux autres dénominations du même domaine.

Dans un système TAL, une ontologie est nécessaire pour le chercheur pour la compréhension des concepts et leur représentation dans un langage formel. L'intérêt de cette application dans le système de la traduction automatique permet également de produire à partir de ces représentations, des concepts ou des textes dans une ou plusieurs langues.

Notre travail est composé de trois parties, La première tente de dégager une connaissance générale sur les ontologies, le web sémantique ainsi que l'ingénierie des connaissances.

Dans la deuxième partie de ce travail, Nous envisagerons étudier comment une ontologie peut être conçue, construite et utilisée dans le domaine de recherche d'information.

La troisième partie sera consacrée à la mise en place de l'ontologie de domaine dédiée au Journal Officiel que nous allons étudier avec ses approches informatiques, Nous finirons par étendre et enrichir le modèle obtenu pour permettre l'indexation et l'interrogation sémantique des textes juridiques.

Nous Conclurons ce travail en montrant l'importance du traitement de la connaissance dans tous les domaines technologiques et scientifiques. On ne peut obtenir de traduction automatique de qualité, de réponses à des questions ou de recherches d'informations si l'on ne dispose pas de connaissances stables, structurées, analysées sur le domaine concerné par l'application.

Chapitre 1 :

Etat de l’art

Chapitre 1 : Etat de l'art :

Nées des besoins de représentation des connaissances, les ontologies sont à l'heure actuelle au cœur des travaux menés en Ingénierie des Connaissances (IC). Visant à établir des représentations à travers lesquelles les machines puissent manipuler la sémantique des informations, la construction des ontologies demande à la fois une étude des connaissances humaines et la définition de langages de représentation, ainsi que la réalisation de systèmes pour les manipuler. Les ontologies participent donc pleinement aux dimensions scientifiques et techniques de l'Intelligence Artificielle (IA).

Nous allons commencer par ce chapitre où nous allons présenter les principales notions et concepts des ontologies ainsi que du Web sémantique et de l'intelligence artificielle.

1. Qu'est-ce qu'une ontologie de domaine ?

Introduit en Intelligence Artificielle (IA) il y a 25 ans, le terme d'ontologie est cependant usité en philosophie depuis le XIXème siècle. Dans ce domaine, l'Ontologie désigne l'étude de ce qui existe, c'est à dire l'ensemble des connaissances que l'on a sur le monde [WELTY 2001]. En IA, de façon moins ambitieuse, on ne considère que des ontologies, relatives aux différents domaines de connaissances.

1.1. Définitions :

La définition donnée aux ontologies a évoluée au court du temps au sein de la communauté de l'Intelligence Artificielle. La plus communément admise est celle donnée par Tom Gruber en 1997 qui la définie comme étant une spécification formelle et explicite d'une conceptualisation, "*An explicit specification of a conceptualization*" [Gruber 1993a].

En 2007, T. Gruber précise [Gruber 2007] que dans le contexte de l'informatique et des sciences de l'information, une ontologie définit un ensemble de primitives de représentation pour modéliser un domaine de connaissances. Les primitives de représentation sont généralement des classes (ou des ensembles), des attributs (ou des propriétés), et des relations (ou des liens qui relient des éléments de classe). Les définitions des primitives de représentation incluent des informations sur leurs significations et des contraintes sur leurs applications, qui doivent être logiquement cohérentes. Dans le contexte des systèmes de base de données, l'ontologie peut être considérée comme un niveau d'abstraction des modèles de données, analogue aux modèles hiérarchiques et relationnels, mais destinée à la modélisation des connaissances sur les individus, leurs attributs et leurs relations avec d'autres individus.

Les ontologies sont généralement décrites dans les langages qui permettent l'abstraction indépendamment des structures de données et des stratégies de mise en œuvre. En pratique, les langages de description des ontologies ont une puissance expressive plus proche de la logique du premier ordre que celle des langages utilisés pour les modèles de bases de données. Pour cette raison, on dit des ontologies qu'elles sont de niveau "sémantique", tandis que les schémas de bases de données sont des modèles de données de niveau "logique" ou "physique".

Du fait de leur indépendance par rapport aux modèles de données de niveau inférieur, les ontologies sont utilisées pour l'intégration de bases de données hétérogènes, permettant une interopérabilité entre des systèmes disparates, et la spécification d'interfaces de services indépendants de la connaissance. Dans le stack technologique des standards du Web sémantique, les ontologies représentent explicitement une couche [Gruber 1993a] (Figure 2).

1.2. Caractéristiques des ontologies :

Les ontologies possèdent des caractéristiques fondamentales [Kaveh 2004].

1.2.1. Les ontologies sont formelles :

Ceci signifie qu'elles sont exprimées dans une langue qui a une syntaxe clairement définie et base mathématique pour leur signification. Comme les concepts sont exprimés formellement, ils peuvent être traités par des programmes informatiques. Les « concepts » ou les « objets » qui existent dans des techniques de modélisation traditionnelles (schéma relationnel et UML, par exemple) sont seulement semi formels. Elles ne peuvent donc pas être manipulées automatiquement par des logiciels sans un effort considérable (et coûteux) de programmation de manière à faire ressortir leurs significations.

1.2.2. Les ontologies sont lisibles par les humains :

Ceci signifie qu'elles peuvent être développées, partagées, et comprises non seulement par des programmes informatiques, mais aussi par les communautés d'experts de domaine ainsi que des utilisateurs potentiels [Cimiano 2005].

1.2.3. Les ontologies sont vastes :

Elles sont conçues avec le but d'inclure toute la signification appropriée des concepts liés à un domaine ; pas simplement celles requise pour une application particulière. Cela veut dire que si toute la signification des concepts est capturée par une ontologie, elle peut être comprise, modifiée, et contrôlée par n'importe quel expert de domaine.

1.2.4. Les ontologies sont partageables :

Elles sont construites sur la base de bibliothèques communes de concepts fondamentaux et sont utilisables à travers de multiples domaines d'application. Ceci facilite la combinaison des ontologies développées séparément pour permettre la communication entre les systèmes d'information qui doivent partager des informations basées sur des concepts communs.

1.3. Classification des ontologies (top-level ontologies):

Les ontologies peuvent être classifiées en fonction de deux dimensions : leur niveau de détail et leur niveau de dépendance par rapport à une tâche particulière, un point de vue. Plus précisément, Guarino [Guarino 1998] propose une classification des ontologies selon leurs niveaux de généralité.

1.3.1. Les ontologies de haut niveau :

Décrivent les concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc…, qui sont indépendants d'un problème ou d'un domaine d'application particulier [Guarino 1998].

1.3.2. Les ontologies de domaine et les ontologies de tâche :

Décrivent, respectivement, le vocabulaire lié à un domaine générique (comme la médecine, ou les automobiles) ou une tâche ou une activité générique (comme le diagnostic ou la vente), en spécialisant les concepts présentés dans les ontologies de hauts niveaux [Falquet 2003]. Elles donnent une représentation formelle des concepts du domaine étudié ainsi que des différentes relations qui lient ces derniers ; elle ne contient pas les concepts pédagogiques, narratifs et structurels.

1.3.3. Les ontologies d'application (application ontologies) :

Décrivent des concepts qui dépendant à la fois d'un domaine et d'une tâche particulière, qui sont souvent des spécialisations des deux ontologies relatives. Ces concepts correspondent souvent aux rôles joués par des entités de domaine tout en exécutant une certaine activité, comme l'unité remplaçable ou le composant disponible [Lri-annaba] (figure 1).

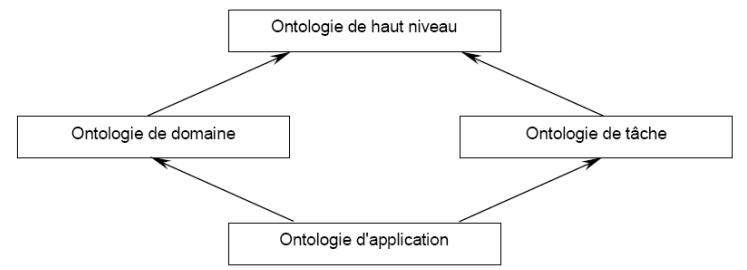


Figure 1: Différents types d'ontologie selon leur degré de dépendance vis-à-vis d'une tâche particulière ou d'un point de vue(Les flèches représentent des relations de spécialisation) [Teimzit 2010].

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

1.4. Rôle des ontologies :

1.4.1. Modularité et réutilisabilité des connaissances :

Les ontologies sont surtout utilisées pour la représentation des connaissances et l'application de raisonnements sur ces connaissances. Cependant une ontologie possède des caractéristiques qui, au-delà de cette représentation, favorisent la réutilisation et le partage de données. Déjà en 1991, Gruber insistait sur le rôle que pouvaient tenir les ontologies pour favoriser la modularité et la réutilisabilité dans les systèmes informatiques [Gruber 1993b]. Gruber souligne les difficultés techniques occasionnées par la conception d'ontologies communes. Ces idées ont été beaucoup approfondies et développées dans [Gruber 1993a]. Pour lui les systèmes à base de connaissance mettent en place des techniques d'interopérabilité basées sur la communication et les opérations à partir de représentations formelles de la connaissance. Ils peuvent souvent être comparés à des agents qui négocient et échangent des connaissances. Trois niveaux de convention sont alors nécessaires :

- Le format de représentation du langage ;
- Le protocole de communication des agents ;
- La spécification du contenu du vocabulaire partagé. C'est surtout sur ce dernier point que les ontologies peuvent jouer un rôle intéressant.

Le partage et l'échange de données entre agents exigent le respect de certaines propriétés [Guarino 1996]. Le rôle clef d'une ontologie en extraction d'information est d'établir l'accord entre le descripteur recherché et les données.

Une ontologie permet de définir les mots d'un langage naturel, les prédicats utilisés dans les calculs de prédicats, les types de concepts et de relations des graphes conceptuels, les classes d'un langage orienté objet ou les champs des tables d'une base de données relationnelle. Or la plupart de ces méthodologies sont connues et utilisées parce qu'elles favorisent l'échange et la réutilisation de connaissances.

1.4.2. Communication :

Il existe trois types de communications dans un projet : communication homme-homme, homme-système ou entre les différents modules du système. Ces trois types possèdent tous des caractéristiques particulières qui engendrent certains problèmes auxquels les ontologies peuvent apporter des solutions.

La communication entre humain pose surtout des problèmes quand les acteurs de cette communication ne sont pas du même domaine et ne parlent donc pas forcément le même langage. La réutilisation, le partage de connaissance et d'ontologies, suppose que plusieurs utilisateurs soient d'accord sur les ontologies partagées. Il a été proposé d'aider les spécialistes de l'ingénierie de la connaissance en utilisant la terminologie définie dans WordNet comme base de la communication, car c'est un standard [Martin 1995].

Une fois que les acteurs humains d'un projet sont d'accord sur une ontologie, la communication avec le système se fait naturellement, en utilisant cette ontologie. De plus l'adaptation des ontologies à la description de textes en langage naturel, semi-structurés [Klein 2000] améliore la communication dans le sens machine-homme.

Les ontologies peuvent également être utilisées pour harmoniser la communication entre différentes applications ou entre différents agents. Cette idée, également présente dans les publications de Gruber [Gruber 1993a], repose souvent sur une ontologie du domaine. Pourtant d'autres chercheurs veulent aller plus loin en dotant les agents d'une connaissance sur une ontologie de tâche indépendante du domaine.
Pour synthétiser, on peut dire que si le rôle principal d'une ontologie est de favoriser le partage et la réutilisation de la connaissance, il faut cependant distinguer plusieurs types d'utilisation qui entraînent des besoins différents :

- Une ontologie peut être utilisée comme un répertoire dans lequel on stocke et organise des connaissances et des informations. Elle peut concerner des données simples, standardisées dans un domaine particulier ou bien des données distribuées ;
- En acquisition de connaissance, les ontologies rassemblent les définitions des termes d'un domaine ce qui permet à plusieurs acteurs de communiquer sans ambiguïté ;
- L'ontologie doit également contenir certaines définitions qui permettent d'assurer la consistance de la base de connaissance et son utilisation correcte ;
- Les ontologies se justifient souvent par la volonté de réutiliser la connaissance pour la construction de nouvelles applications ;
- Enfin, une ontologie peut être utilisée comme la base d'un langage de représentation des connaissances.

2. Le Web sémantique :

L'expression Web sémantique, attribuée à Tim Berners-Lee [Berners-Lee 2001] au sein du W3C, fait d'abord référence à la vision du Web de demain comme un vaste espace d'échange de ressources entre êtres humains et machines permettant une exploitation, qualitativement supérieure, de grands volumes d'informations et de services variés. Espace virtuel, ou les utilisateurs déchargés d'une bonne partie de leurs tâches de recherche, de construction et de combinaison des résultats, grâce aux capacités accrues des machines à accéder aux contenus des ressources et à effectuer des raisonnements sur ceux-ci.

Le Web sémantique, concrètement, est d'abord une infrastructure pour permettre l'utilisation de connaissances formalisées. Cette infrastructure doit permettre d'abord de localiser, d'identifier et de transformer des ressources de manière robuste et saine tout en renforçant

l'esprit d'ouverture du Web avec sa diversité d'utilisateurs. Elle doit s'appuyer sur un certain niveau de consensus portant, par exemple, sur les langages de représentation ou sur les ontologies utilisées. Elle doit contribuer à assurer, le plus automatiquement possible, l'interopérabilité et les transformations entre les différents formalismes et les différentes ontologies. Elle doit faciliter la mise en œuvre de calculs et de raisonnements complexes tout en offrant des garanties supérieures sur leur validité. Elle doit offrir des mécanismes de protection (droits d'accès, d'utilisation et de reproduction), ainsi que des mécanismes permettant de qualifier les connaissances afin d'augmenter le niveau de confiance des utilisateurs. Mais restreindre le Web sémantique à cette infrastructure serait trop limitatif. Ce sont les applications développées sur celle-ci qui font et feront vivre cette vision et qui seront, d'une certaine manière, la preuve du concept. Bien sûr, de manière duale, le développement des outils, intégrant les standards du Web sémantique, doit permettre de réaliser plus facilement et à moindre coût des applications ou des services développés aujourd'hui de manière souvent ad-hoc.

2.1. Objectifs du Web sémantique :

Un des principaux objectifs du Web sémantique est de permettre aux utilisateurs d'utiliser la totalité du potentiel du Web: ainsi, ils pourront trouver, partager et combiner des informations plus facilement. Aujourd'hui tout le monde est capable d'utiliser des forums, d'utiliser des réseaux sociaux, de chatter, de faire des recherches ou même d'acheter différents produits. Néanmoins, il serait mieux que la machine fasse tout ceci à la place de l'homme, car actuellement, les machines ont besoin de l'homme pour effectuer ces tâches. La raison principale est que les pages Web actuelles sont conçues pour être lisibles par des êtres humains et non par des machines. Le Web sémantique a donc comme principal objectif que ces mêmes machines puissent réaliser seules toutes les tâches fastidieuses comme la recherche ou l'association d'informations et d'agir sur le Web lui-même.

2.2 Le Web sémantique : Approche par Couches [PAPINI 2010] :

Couche XML :
- base syntaxique

Couche RDF :
- RDF : modèle de données basique pour les faits
- RDF Schéma : langage pour les ontologies

Couche Ontologie :
- langage plus expressif que RDF Schéma
- standard courant pour le web : OWL

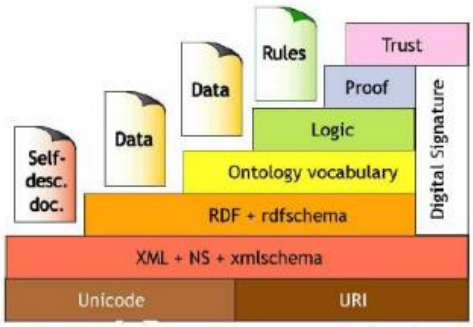


Figure 2: Les couches du Web Sémantique [Laublet 2003]

2.3 Le Web sémantique et la langue arabe :

2.3.1 Importance de la langue arabe :

On doit bien reconnaitre la grande importance de cette langue; L'arabe est la langue maternelle pour plus de 250 million d'arabes, la plupart d'eux se trouvent entre le Maroc à l'Ouest et l'Iraq

à l'Est. Les musulmans considèrent la langue arabe classique comme une langue sacrée du fait que c'est la langue du Coran qui a gardé sa pureté aux yeux des musulmans.

La langue arabe a changé un peu à travers les années, pour cela le Coran est resté une source de la langue.
La langue arabe, comme toutes les langues vivantes, s'est développée avec le temps pour comporter des nouveaux mots et des expressions comme les nouvelles terminologies techniques. La traduction des ouvrages et des sciences accélère ce développement. Cela crée des espaces de recherche qui contestent, dans certains cas, l'adoption des mots étrangers à la langue ainsi que la formation des nouveaux mots qui ont un caractère arabe pur pour couvrir le nouveau vocabulaire. Il y a des autres changements qui comprennent la perte de beaucoup des terminologies anciennes surtout celles qui concernent la vie quotidienne.

La langue arabe est connue par sa richesse et sa complexité morphologique ; sa morphologie était un défi pour les spécialistes de traitement automatique des langues naturelles et un terrain tenace d'essai pour les différentes technologies et approches d'analyse automatique [Mesfar 2008].

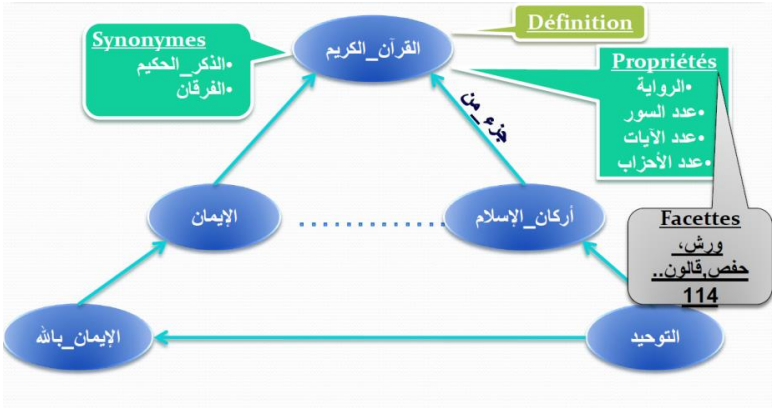


Figure 3: Construction d'ontologie à partir de textes arabes [Zaidi 2008]

2.3.2 Difficultés de travail

La langue arabe est une langue difficile qui risque d'entraver le développement des outils du Web sémantique. La langue arabe a de nombreuses particularités comme voyelles courtes, l'absence de lettres majuscules et complexes morphologie. La langue arabe est composée de noms, des verbes. Ces entités représentent des morphèmes dérivés à partir d'un ensemble fermé de près de 10.000 racines. L'arabe est aussi hautement flexionnelle et dérivationnelle, ce qui rend l'analyse morphologique d'une tâche très complexe. Il n y a pas de capitalisation en arabe, ce qui rend difficile d'identifier noms propres, acronymes et abréviations.

2.3.3 La langue arabe et le Web sémantique :

Il y a diverses études [Beseiso 2010] menées sur la langue arabe dans le Web sémantique. Pour améliorer l'extraction d'informations en arabe sur le Web par un moteur de recherche arabe soutenant la traduction de requêtes arabes en anglais ou en français. L'objectif était de remettre des documents écrits en Arabe, en français ou en anglais. Vossen, Pease et Fellbaum travaillé sur la Parole de l'arabe net (AWN) sur la base des méthodes développées pour EuroWordNet (ROE) et depuis appliquée à des dizaines de langues à travers le monde. Le EuroWordNet est une approche qui maximise la compatibilité entre les Wordnets et met l'accent sur le codage manuel c'est l'un des plus importants et complexe. Les critères de base pour AWN sont la connectivité, la pertinence, et la généralité, de l'anglais vers l'Arabe et de l'arabe vers l'anglais. Des enquêtes sur l'efficacité de la récupération de l'amélioration des moteurs de recherche pour

mettre les accents sur les documents en arabe, en construisant un arabe Anglais de système de RI basé sur une approche de traduction automatique [Hammo 2009]. AbdulJaleel et Larkey [Abduljaleel 2003], ont proposé une approche statistique pour la translittération arabe-anglais IR.

Grefenstette et al. [Yan 2005], décrivent les changements nécessaires pour modifier le système de RI dans leur langue, qui a été conçu pour les langues européennes à intégrer la langue arabe. Abdelali et al. [Grefenstette 2005], ont décrit comment la précision peut être améliorée dans l'expansion de requête en utilisant LSI. Enfin, Semmar et Fluhr [Semmar 2007] ont présenté une nouvelle approche pour aligner Arabe-français phrases extraites d'un corpus parallèle basé sur un système de RI de contre-langue. Cette approche est essentiellement basé sur la construction d'une base de données de phrases du texte cible et considérant chaque phrase du texte source en tant que requête pour la base de données [Hammo 2009].

Guo et Ren [Guo 2009] ont souligné que l'utilisation de la technologie du traitement du langage naturel (PNL) comme un élément important dans le Web Tools sémantique. La PNL est une branche de la linguistique, qui utilise la technologie informatique pour réaliser le traitement humain sur le langage de manière efficace. Son objectif ultime est de comprendre automatiquement le langage humain avec le soutien de la technologie d'intelligence artificielle. Il est aussi appelé compréhension du langage naturel et est parfois utilisé pour transformer l'information à des données du Web sémantique. La recherche traditionnelle d'information peut également être transformée en connaissance découverte. Al-Khalifa, Al-Yahya, Bahanshal et Al-Odah [Al-Khalifa 2009] ont proposé un cadre pour représenter une opposition sémantique dans le Saint Coran à l'aide des technologies du Web sémantique. Des recherches antérieures dans le domaine de l'Informatique et du Saint Coran peuvent être classées en six catégories, à savoir : recherche d'information, Reconnaissance de discours, la reconnaissance optique des caractères, l'analyse morphologique, contrôle sémantique et applications éducatives.

Très peu de travaux ont été faits dans le domaine des technologies du web sémantique pour servir la sémantique lexicale du Saint-Coran.

Hammo, Abou-Salem et Lytinen [Hammo 2005] ont développé un système QARAB dont l'objectif principal est d'identifier les passages de texte qui répondent à une question en langage naturel. Les tâches de QARAB peuvent être résumées comme suit: Etant donné un ensemble de questions exprimé en arabe, trouver des réponses aux questions sous les hypothèses suivantes:

- La réponse existe dans une collection de textes de journaux en arabe extrait du journal Al-Raya publié au Qatar ;
- La réponse ne couvre pas dans les documents (c.-à-toutes les informations de support pour la réponse réside dans un seul document) ;
- La réponse est un court passage [Hammo 2005].

Ce ne sont que quelques études menées directement ou indirectement dans le web sémantique en langue arabe. Basé sur les informations recueillies, on peut conclure que le travail en langue arabe pour le Web sémantique est encore dans l'enfance.

En raison de cela, il est possible d'aller plus loin, outre les actuelles disponibles sémantiques arabes comme ceux qui sont utilisés dans le Coran.

2.4. Ontologie pour la langue arabe :

L'Ontologie arabe est le fondement de la création du Web Sémantique et la catégorisation de base des terminologies et des significations dans un domaine en langue arabe. L'interrelation entre un mot aux autres qui correspondent à son sens peut également entraîner des tiges et des branches de la sémantique. L'objectif d'une ontologie d'apprentissage est d'extraire automatiquement les concepts pertinents et les relations du corpus donné ou d'autres types d'ensembles de données pour former l'ontologie [Guo 2009]. Le cycle de vie dans le développement de l'ontologie peut être subdivisé dans les catégories suivantes :

L'extraction de termes, Elaboration des synonymes, l'obtention de concepts, l'extraction de hiérarchies de concepts, la définition les relations entre les concepts, déduction des règles et des axiomes.

Ces procédés sont utilisés afin de rendre l'ontologie possible et de rendre aussi les branches connexes des sujets disponibles pour tout utilisateur.

Alkhalil : une ontologie OWL pour la Linguistique arabe :
Al-Khalil [Aliane 2010] est une ontologie OWL en cours de réalisation ; Son développement s'étale sur deux étapes :

- Démarrer manuellement l'ontologie par le choix des concepts linguistiques de la linguistique arabe et à les relier aux concepts de GOLD ;
- Utilisation d'un algorithme d'extraction automatique pour extraire de nouveaux concepts à partir de textes linguistiques pour enrichir l'ontologie. L'algorithme est basé sur les méthodes de calcul des segments répétées. L'architecture générale du Système est représentée sur la figure 4.

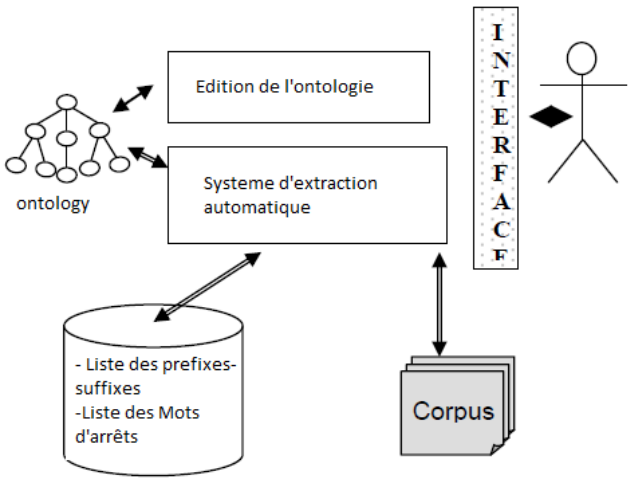


Figure 4: L'architecture du prototype [Aliane 2010]

3. Ingénierie des connaissances :

3.1. L'ingénierie des connaissances :

L'ingénierie des connaissances correspond à l'étude des modèles symboliques formels plongés dans des systèmes d'usage [JFIC 2009] : c'est l'ingénierie informatique et logique des modèles en fonction des usages qu'ils rendent possibles et des appropriations qu'ils permettent.

Ingénierie évoque un ensemble de techniques et de méthodes appliquées pour la résolution des problèmes complexes. Ingénierie de la connaissance, équivalent anglais de Knowledge engineering, serait synonyme de Génie cognitif :

- Intégration des techniques d'intelligence artificielle et du génie logiciel en vue de concevoir et de construire des systèmes experts ;
- Discipline étudiant l'extraction et la formalisation de connaissances provenant d'un expert humain en vue de leur intégration dans des systèmes experts.

3.2. Système expert :

Un système expert est un ensemble de logiciels modélisant, dans un domaine précis (généralement très circonscrit), les compétences et les modes de raisonnement d'un ou de plusieurs experts. Évolutif, le système expert évite d'avoir à écrire de nouveaux programmes pour réinjecter de l'information : grâce à son module d'acquisition, on peut incorporer une donnée nouvelle en cours d'utilisation. À l'inverse, la machine rendue interactive peut "pointer du doigt" une erreur commise en cours de tâche par la personne qui l'utilise.
En fait, le système expert est conçu pour aider un utilisateur dans un domaine particulier à trouver la solution adaptée à son questionnement, et ce, bien évidemment, dans l'état actuel des connaissances spécialisées.

L'ingénieur cogniticien utilise, pour élaborer son système expert, une méthode essentiellement "clinique" (qui procède par étude de cas individuels). D'une manière générale, le concepteur vise à dégager trois niveaux au sein de la masse des connaissances. En premier, le niveau structurant concerne les procédures déductives utilisées dans le domaine considéré pour atteindre la certitude. C'est là que se niche le fameux "sens commun", impossible à globaliser : il n'est étudiable qu'au coup par coup. Quand le cogniticien maîtrise ce niveau, il est à même de représenter la connaissance au niveau conceptuel, où figureront les concepts dont le spécialiste fait un usage courant. Enfin, le niveau cognitif contiendra une quantité maximale de connaissances brutes relatives au domaine en question.

Le système expert comprend : la base de faits, qui contient les connaissances intangibles nécessaires à la pratique et les informations déduites par le système à l'issue de solutionnements successifs ; le moteur d'inférence, logiciel fabriquant des raisonnements en se fondant sur la base ; les interfaces, programmes permettant le dialogue avec le système, en langage naturel pour le non-expert. Le moteur d'inférence et les interfaces forment le système essentiel, ainsi nommé parce qu'on peut le coupler à diverses bases de faits pour créer des systèmes experts distincts.

4. Conclusion :

Nous avons présenté dans ce chapitre les notions de base des ontologies ; Car il est nécessaire de les définir, les classifier et de présenter leurs caractéristiques ainsi que leurs rôles avant de passer à la phase de construction. Aussi il est important de présenter les concepts de base du Web sémantique et de l'Intelligence Artificielle et de citer quelques études menées sur la langue arabe dans le domaine du Web sémantique et des ontologies.

Chapitre 2 :
Comment construire une ontologie ?

Chapitre 2 : Comment construire une ontologie ?

De manière générale, l'utilisation de connaissances en informatique a pour but de ne plus faire manipuler en aveugle des informations à la machine mais de permettre un dialogue, une coopération entre le système et les utilisateurs. Pour cela, le système doit avoir accès non seulement aux termes utilisés par l'être humain mais également à la sémantique qui leur est associée, afin qu'une communication efficace soit possible. Les ontologies visent à représenter cette connaissance en étant à la fois interprétables par l'homme et par la machine.

1. Modélisation du contexte d'une recherche :

L'Ingénierie des Connaissances (IC) est une branche de l'intelligence artificielle axée sur la connaissance. Les principales préoccupations de ce domaine sont l'acquisition, la modélisation, le stockage et la consultation de connaissances, le raisonnement automatique sur les connaissances stockées et la modification des connaissances stockées. Lorsque les connaissances à construire sont issues de documents, l'IC s'appuie sur des méthodologies développées dans le domaine de la linguistique et du traitement automatique des langues pour assurer une compréhension des contenus des documents considérés.

Parallèlement, la Recherche d'Information (RI) est une activité dont la finalité est de mettre en regard des informations et un utilisateur. C'est une activité par laquelle un utilisateur accède à un granule d'information à partir d'un besoin qu'il spécifie. Les systèmes de RI (SRI) développés depuis le début des années 50 reposent essentiellement sur des approches statistiques et des approches linguistiques de bas niveau. Ces approches prennent uniquement en compte le niveau lexical, parfois le niveau syntaxique, du contenu textuel des granules afin d'identifier les mots permettant de retrouver les granules répondant aux besoins de l'utilisateur. Un enjeu actuel de la RI, comme du Web avec le Web Sémantique, est de s'appuyer sur des connaissances pour enrichir les systèmes en apportant une couche sémantique.

Dans le cadre de la RI, la problématique posée est de doter le SRI de connaissances lui permettant d'être un intermédiaire entre l'utilisateur et les granules d'information. Ce rôle d'intermédiaire se joue entre un utilisateur dont le système doit être capable d'interpréter les besoins et des granules d'information dont le système doit pouvoir interpréter le contenu. Aussi, l'utilisation de connaissances par un SRI doit lui permettre de connaître le contexte de la recherche. D'un côté, le système doit être capable d'inférer les intentions de l'utilisateur en fonction de la tâche visée et, de l'autre, d'expliciter le contenu d'un granule d'information [Hernandez 2005].

1.1. Connaissances sur le contexte d'une recherche d'information :

L'exploitation du contexte en RI repose sur deux fondements [Johnson 2003]. Dans le domaine du langage et de la communication, le contexte est utilisé pour appréhender le sens des mots. Il permet de désambiguïser le sens des unités sémantiques en fonction du domaine, de la discipline et du contexte linguistique (document, phrase …) auxquels elles appartiennent. De plus, le contexte permet de déterminer l'action sociale dans laquelle se place un individu. Il est en effet plus facile de modéliser le comportement des utilisateurs appartenant à un groupe restreint et identifié que de déterminer les caractéristiques propres à l'ensemble de la population. Comme le souligne [Freund 2005], ces deux aspects sont fondamentaux en RI.

Plus généralement, la notion de contexte recouvre en RI différents aspects. Dans un modèle analytique de la RI est présenté sur la figure 5.

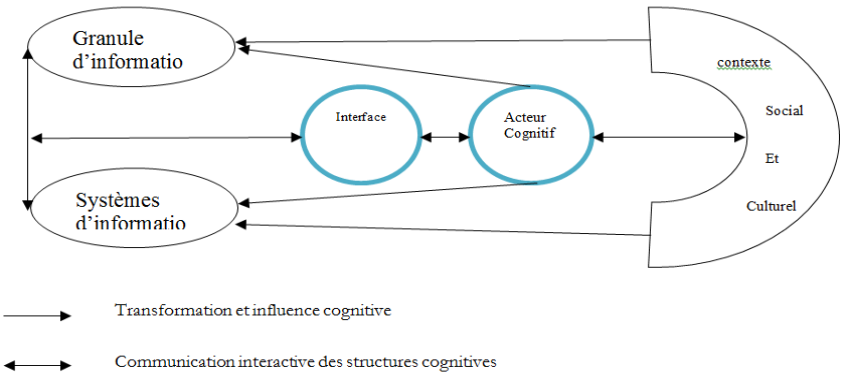


Figure 5: Modèle cognitif de la recherche d'information adaptée [Jarvelin 2004]

Chacun des acteurs est représenté (nœud du graphe) ainsi que leurs liens. Les acteurs cognitifs (comme les utilisateurs recherchant une information) sont entourés de différents acteurs de contexte (granules d'information, systèmes d'information, interface, contexte social, culturel et organisationnel). Tous ces acteurs sont en interaction. Les granules d'information correspondant aux éléments gérés par le système d'information peuvent être très variés (collections de documents, documents, parties de documents, phrases, données). Le système d'information lui-même repose sur différents modèles et méthodes (modèle de représentation de l'information, méthode de mise en correspondance), en lien avec les objets d'information gérés et l'interface de manipulation. L'acteur cognitif de son côté est influencé dans sa recherche par le contexte culturel et social dont il dépend ; les objets auxquels il s'intéresse en dépendent. Le contexte du système dépend à la fois de l'information gérée et des spécificités de l'utilisateur. L'interface, qui permet de faire la correspondance entre l'utilisateur et la collection d'objets traités et restitués par le système, doit s'adapter en fonction des contextes (types d'information manipulée, utilisateur, technologie de RI utilisée).

Les systèmes de RI actuels visent à satisfaire la majorité des utilisateurs dans la plupart des cas. La variabilité des contextes qui viennent d'être présentés les rend difficiles à modéliser. Pour cette raison, bien que le contexte soit omniprésent, les systèmes de RI n'en intègrent que certains aspects. Les travaux menés dans le cadre du workshop IRiX (Information Retrieval in conteXt) en 2005 ont conclu que quatre dimensions principales du contexte pouvaient être retenues : les granules d'information, la tâche, l'utilisateur, le système.

1.1.1. Contexte et granules d'information :

1.1.1.1. Représentation du contenu des granules d'information :

Les index jouent un rôle primordial en RI en définissant les descripteurs (mots ou groupements de mots) à partir desquels l'information contenue dans les granules est représentée. Le langage d'indexation est donc un langage artificiel, c'est-à-dire construit à l'aide d'un ensemble de règles données, servant à la représentation abrégée du contenu d'un document [Rivier 1990]. Dès lors, l'indexation consiste à détecter les termes les plus représentatifs du contenu du document.

Deux idées très simples ont rendu possible l'indexation automatique du contenu textuel de documents : considérer les mots qui apparaissent dans les granules et compter leurs occurrences [Zipf 1949]. Les recherches dans le domaine ont montré qu'il est essentiel pour une indexation efficace et donc pour une recherche efficace de prendre en compte la distribution de ces mots

dans les documents car les fréquences relatives auxquelles les mots apparaissent ou co-occurrent permettent de déterminer les thématiques et leur signification dans les textes.
D'ailleurs, les modèles de RI se sont attachés à rendre la meilleure représentation possible du contenu de la collection de documents (indexation) pour leur mise en correspondance avec la requête. A l'inverse des premiers systèmes automatiques, il s'agit de ne plus considérer que les termes issus des documents soient indépendants. Il s'agit là d'une forme de connaissance sur les termes et leur utilisation. Plusieurs méthodes ont été introduites.

La représentation des documents par des radicaux, plutôt que par les termes tels qu'ils apparaissent dans les documents, est une première démarche permettant une meilleure représentation des contenus et prenant en compte la dépendance entre termes. Cette radicalisation peut s'appuyer sur différentes stratégies : troncatures simples [Denjean 1989], suppression des suffixes [Porter 1980], utilisation de connaissances linguistiques [Savoy 1993]. Il s'agit ici de considérer les différentes formes lexicales d'un terme comme équivalentes. La majorité des SRI actuels reposent sur une représentation par termes radicalisés.

Cependant les mots retenus par l'indexation peuvent être ambigus. Les descripteurs peuvent en fait se rapporter à des termes ayant plusieurs sens et donc ne pas indiquer clairement la thématique abordée dans le document. Différents descripteurs peuvent également se rapporter à une même notion dans le cas où les mots choisis sont synonymes. L'index est alors surchargé par des éléments représentant la même information. D'autre part, la recherche peut échouer si les termes de la requête n'y apparaissent pas. L'utilisation de connaissances dans le but d'aider le SRI à interpréter le contenu des documents permet d'accéder à la sémantique associée aux mots issus du contenu. Cette sémantique repose sur l'identification des concepts et des relations entre ces concepts de manière à établir clairement les notions, les termes et les objets associés aux mots issus des textes des granules. Une représentation de la connaissance abordée dans le granule permet au système de prendre en compte la sémantique sous-jacente aux mots qui composent le contenu des granules. De façon générale, une analyse statistique permet l'extraction de descripteurs des granules d'information mais pas leur compréhension. Ceci signifie que l'utilisateur ou le système possède un ensemble de connaissances préalables et que la « compréhension du texte » lui permet de modifier cette connaissance en ajoutant, supprimant ou modifiant la connaissance qu'il avait déjà.

De la même façon, l'enjeu des activités documentaires est décrit dans le projet ASSTICCOT comme visant à permettre que des connaissances produites par un auteur engendrent des connaissances "nouvelles" c'est-à-dire différentes, pour les utilisateurs [ASSTICCOT 2003]. Typiquement, les connaissances diffusées dans un brevet d'invention vont permettre de produire d'autres connaissances pour les utilisateurs (connaissances se traduisant par un positionnement stratégique par exemple dans le domaine de la veille). Représenter des documents d'un domaine à partir de la connaissance issue de ce domaine, peut permettre de mettre à jour cette connaissance par le traitement des documents. L'opération effectuée par le système se rapproche alors de la compréhension des documents. Fournir à l'utilisateur la ressource de connaissances permettant le traitement par le système peut, de plus, l'aider dans sa compréhension des documents.

Certains auteurs affirment que l'indexation peut être considérée comme une forme d'acquisition de connaissances sur le contenu documentaire [Dachlet 1990]. La connaissance est une information active et interprétée par l'homme. Dans le cas des index, l'information extraite a pour but de représenter l'information du granule et de permettre d'établir une correspondance entre son contenu et le besoin de l'utilisateur. Cette information n'est pas active dans le sens où le seul processus dans lequel elle s'inscrit est la mise en correspondance. Elle ne permet ni

d'inférer de nouvelles connaissances, ni d'être interprétée par l'homme pour qu'il mette à jour ou ne complète ses connaissances.

Le traitement des documents doit prendre en compte non seulement leur contenu textuel mais aussi les métadonnées qui peuvent être associées aux documents.

1.1.1.2 Métadonnées associées aux granules d'information :

Les métadonnées sont des données factuelles qui contiennent de l'information sur l'information des granules. Plus précisément, c'est un ensemble structuré d'informations décrivant une ressource. Elles sont associées aux ressources sans ambiguïté comme, par exemple, le nom des auteurs, la date de publication, les mots clés choisis pour indexer le document… Les ressources étant généralement partagées, plusieurs standards ont été définis pour permettre leur description à l'aide des métadonnées.

Le Dublin Core [dublincore] définit un ensemble de 15 métadonnées relatives :

- au *Contenu* : Titre, Description, Sujet, Source, Couverture, Type, Relation ;
- à la *Propriété intellectuelle* : Créateur, Contributeur, Editeur, Droits ;
- à la *Version* : Date, Format, Identifiant, Langage.

Les métadonnées associées par le Dublin Core sont considérées comme descriptives car elles sont externes aux contenus même des documents et elles indiquent comment le granule a été créé [Baeza-Yates 1999]. D'autres types de métadonnées comme les métadonnées associées par le système Medline (http://medline.cos.com/) sont relatives aux contenus même des granules. Des métadonnées à propos de maladies, symptômes ou anatomies sont associées par ce système à des articles de médecine.

RDF (Resource Description Framework) [Lassila 19991] est un moyen d'encoder, d'échanger et de réutiliser des métadonnées structurées. Les métadonnées peuvent aussi bien être descriptives que relatives aux contenus des granules.

Pour être interprété par un système, il est nécessaire qu'une sémantique soit associée à ces métadonnées. Le système doit être capable d'interpréter le rôle de la métadonnée dans la représentation du document. De plus, il doit être capable d'interpréter les liens entre différentes métadonnées associées aux documents. Une ontologie permet de spécifier la connaissance nécessaire au système pour interpréter le rôle sémantique des métadonnées.
Peu de systèmes intègrent à la fois des descripteurs liés aux contenus des granules et aux métadonnées.

1.1.1.3. Représentation de la requête et reformulation :

A l'autre bout de la chaîne de RI, les requêtes sont considérées. La représentation de la requête se limite à l'ensemble des termes issus de la formulation de la requête par l'utilisateur. Cependant, les mécanismes de reformulation de requêtes permettent d'améliorer cette représentation à partir de connaissances extraites des contenus des granules ou de ressources externes.

Un premier type d'approches repose sur l'analyse globale de la collection de documents considérée. L'extraction des liens de cooccurrences entre termes des documents, calculés de façon statistique [Harper 1978] ou faisant intervenir des connaissances linguistiques [Grefenstette 1992], l'extraction de groupes de mots [Pohlmann 1997] et celle des liens contextuels entre termes [Bruandet 1983] [Véronis 1989], font partie de cette catégorie. Les informations ainsi extraites sont généralement utilisées pour reformuler automatiquement une requête par ajout des termes liés aux termes initialement présents dans la requête. Les termes

ainsi ajoutés sont issus des documents et permettent donc une meilleure adéquation entre le besoin d'information et la collection.

L'ajout de termes issus de ressources terminologiques est une autre méthode de reformulation de requête. Dans ce cas, les termes liés aux termes initiaux de la requête sont extraits de thésaurus [Jarvelin 1996] ou de ressources telles que WordNet [Voorhees 1993] [Mandala 1999].

Enfin, le principe de réinjection de pertinence [Rocchio 1971] [Harman 1992] vise également à reformuler une requête initiale pour qu'elle corresponde mieux au contenu de la collection. Le principe est le suivant :

L'utilisateur soumet sa requête initiale, le système restitue un premier ensemble de documents que l'utilisateur doit juger (pertinent, non pertinent). La connaissance de la pertinence des documents initialement restitués est utilisée pour sélectionner des termes à ajouter à la requête initiale (les termes qui sont caractéristiques de la pertinence en quelque sorte), voire décider des termes à éliminer de la requête [Rocchio 1971] [Salton 1990]. Cette méthode repose donc sur l'hypothèse que les documents pertinents se ressemblent. Des améliorations de 20 à 30% de la précision moyenne ont été mesurées [Harman 1992]. L'ambition du « tout automatique » a fait dire que cette méthode impliquait un dialogue système / utilisateur trop lourd. Ainsi, cette méthode a été automatisée. Pour éviter la lourdeur du mécanisme de jugement de pertinence des documents initialement restitués, la réinjection de pertinence aveugle prend en compte non pas la pertinence utilisateur, mais la pertinence système. Dans cette méthode, les premiers documents restitués par rapport à la requête initiale sont considérés comme pertinents. Seule une pertinence positive est alors considérée. Des études ont montré que cette approche permettait d'améliorer les résultats par rapport à une recherche simple. Parallèlement, l'utilisation du contexte local des termes de la requête dans les documents a également été étudiée [Xu 2000].

1.1.1.4. Domaine :

Les portails thématiques ou les librairies électroniques, en se focalisant sur un thème ou un usage, considèrent l'aspect culturel et social du contexte. Le contexte est donc partie intégrante de la collection, soit par son aspect thématique, soit par le type de documents qu'elle contient. L'utilisateur peut être plus confiant sur l'adéquation des documents qu'il va retrouver.

La notion de domaine peut être également abordée à travers la représentation des documents par un langage contrôlé. Par exemple, l'utilisation de la hiérarchie thématique MeSH (Medical Headings) pour représenter les documents de MedLine impose de fait le vocabulaire utile pour la recherche (celui qui permettra de retrouver des documents s'il est utilisé dans une requête) [Hearst 1997]. Le même type d'approche est utilisé dans le système IRAIA [Englmeier 2003] gérant des documents économiques. L'approche multi-facette retenue ici permet de représenter les documents selon différentes hiérarchies de concepts, chacune correspondant à un aspect du domaine (type d'entreprise, pays, indicateurs économiques).

L'indexation de collections à partir d'un vocabulaire de domaine présente de plus les avantages suivants :

- Aider l'utilisateur à formuler sa requête. En présentant le vocabulaire du domaine à l'utilisateur, il est possible de le guider dans le choix des termes de sa requête. Il a été montré qu'un thésaurus permet à l'utilisateur de construire une conceptualisation de ce qu'il est en train de chercher et peut aider à la formulation des requêtes dans le cas de la RI ad-hoc [Baeza-Yates 1999]. Cependant, de nombreuses limites ont été trouvées à l'utilisation de thésaurus [Baeza-

Yates 1999] car, d'une part, leur construction est orientée terminologie et ne capture que les termes d'un domaine et, d'autre part, les relations entre termes restent limitées sémantiquement. L'utilisation de ressources conceptuelles permet de combler ces lacunes,

- Faciliter la RI au sein de collections hétérogènes en indexant tous types de documents à partir des mêmes termes.

La connaissance associée à un domaine peut être représentée de façon plus formelle au travers d'une ontologie. Pour un utilisateur, accéder par une ontologie à la connaissance à partir de laquelle l'information d'un corpus a été indexée peut lui permettre de spécifier son besoin et les lacunes de sa connaissance par rapport à l'information qui lui est disponible. D'autre part, la représentation des granules d'information à partir d'une ontologie peut définir un vocabulaire contrôlé (termes et concepts) à partir duquel l'utilisateur spécifiera son besoin. La description du besoin correspond, dans ce cas-là, aux caractéristiques des granules car elles ont été indexées à partir des mêmes ressources.

1.1.2. Contexte et utilisateur :

Différents facteurs ont été proposés pour représenter le contexte de l'utilisateur [Belkin 2004]:

- la familiarité de l'utilisateur avec le domaine relatif à sa recherche ;
- l'expérience de l'utilisateur dans l'utilisation du ou des systèmes de RI ;
- les documents déjà connus de l'utilisateur ;
- le type des documents recherchés [Rauber 2001] [Freund 2005];
- l'objet de la recherche ;
- la tâche dans laquelle s'inscrit la recherche ;
- les autres activités de l'utilisateur pendant sa recherche. L'utilisateur s'engage dans une recherche d'information parce qu'il a un manque d'information. Cependant, l'utilisateur a une idée plus ou moins définie des lacunes de ses connaissances et donc de son besoin en information. La première difficulté à laquelle doit faire face un SRI est que le besoin en information est interne à l'utilisateur. L'utilisateur juge en effet les éléments qui lui sont retournés par rapport à l'interprétation de son besoin et non pas par rapport à l'ensemble des granules à sa disposition et susceptibles de l'intéresser dans la collection [Turtle 1991].

Ainsi, le projet Profildoc permet de filtrer les documents retrouvés par rapport au profil de l'utilisateur. Ce profil utilisateur contient des informations telles que le niveau éducationnel, le champ disciplinaire (sciences de l'information, agronomie…), le type de recherche (recherche généraliste ou pointue)… Ce profil est utilisé afin d'identifier au sein des documents retrouvés ceux qui correspondent au profil de l'utilisateur et ainsi réduire le nombre de documents en éliminant ceux qui ne seraient pas pertinents pour lui.

De la même façon, la tâche « Hard » de TREC (trec.nist.gov) s'intéresse à l'étude de l'impact de certains de des facteurs liés à l'utilisateur sur les performances d'un système de RI [Allan 2003b]. Plus particulièrement, en 2003, les besoins d'informations comprenaient, en plus des champs traditionnels des « topics » TREC (titre, description, narration), les métadonnées :

- Familiarité avec le thème ;
- Type de documents souhaités ;
- Objet de la recherche ;
- Spécification géographique sur les documents recherchés.

En 2004, les métadonnées ont été simplifiées pour ne retenir que :

- La connaissance sur le thème ;
- Le type de documents recherchés ;
- La spécification géographique (le document est relatif aux Etats-Unis ou non). Dans cette tâche, les participants doivent d'abord fournir les documents retrouvés sans prendre en compte les métadonnées associées à la connaissance de l'utilisateur ; dans un deuxième temps, les participants fournissent les résultats qu'ils obtiennent lorsque leurs systèmes intègrent ces informations sur le contexte. L'influence du contexte est donc mesurée en comparant les premiers résultats avec les seconds.

Enfin, la recherche d'information collaborative correspond à un autre type de processus mis en œuvre pour considérer l'utilisateur. Cette technique vise à permettre à des utilisateurs de bénéficier de jugements de pertinence émis par d'autres utilisateurs supposés partager le même profil. Ces approches reposent essentiellement sur le contenu des informations recherchées. De façon similaire, les systèmes de recommandation visent à optimiser la recherche d'information en proposant automatiquement à l'utilisateur de nouveaux documents au regard de ses besoins exprimés ou de ses actions. Une étude de ces systèmes de recommandation appliqués au contexte d'Internet peut être trouvée chez Montaner et lopez [Montaner 2003]. A la frontière entre les outils de recommandation et de recherche d'information collaborative, Chevalier [Chevalier 2002] propose d'utiliser les utilisateurs comme source d'information pour la recommandation, et ce, au travers des documents qu'ils visitent sur le Web. Par ailleurs, ce système propose un type de filtrage collaboratif au travers de jugements de pertinence déduits de l'organisation des documents (signets) que chaque individu possède.

1.1.3. Contexte et tâche :

Une tâche est définie comme « une activité réalisée pour atteindre un but » [Vakkari 2003]. Une tâche de recherche intervient quand l'utilisateur est en manque d'information. Pour étudier les besoins d'information, Dervin [Dervin 1992] propose, une méthode qui emploie la métaphore « *situation-gap-use* » selon laquelle tous les besoins d'information viennent d'une lacune dans les connaissances d'un individu ; cette lacune entraîne une situation spécifique à laquelle l'individu peut remédier par différentes tactiques. Le but de la recherche détermine le type d'information que l'utilisateur sollicite et l'utilisation qu'il souhaite faire de cette information.

La RI ad-hoc vise à restituer (tous) les documents pertinents (et seulement ceux-là) par rapport à un besoin d'information formulé sous forme de requête par un utilisateur. La plupart des SRI fonctionnent avec une interface qui permet à l'utilisateur de formuler son besoin en information à partir d'une requête. Le système présent ensuite à l'utilisateur le résultat de la recherche sous forme d'une liste de références vers les documents retrouvés. S'il s'agit de la tâche la plus connue, d'autres tâches de RI existent. L'utilisateur peut souhaiter ne consulter que les documents ou granules nouveaux en rapport avec son besoin d'information [Soboroff 2003] ou filtrer les documents par rapport à un profil de recherche [Roberston 2002]. Il peut vouloir une réponse à une question précise [Voorhees 2004] (systèmes questions-réponses) ou en rapport avec un cadre spécifique comme la génomique [Hersh 2004]. La recherche sur le Web [Hawking 1999] correspond à une tâche spécifique dans la mesure où la présence de liens hypertextes peut modifier l'idée de l'utilisateur sur son besoin d'information. La recherche multilingue ou crosslingue [Jones 2000] correspond à un autre type de tâche. Alternativement, l'utilisateur peut vouloir explorer une collection de documents pour les classifier [Sebastiani 2006] ou pour découvrir des informations non implicitement présentes dans les documents comme dans une activité de veille [Chrisment 2006]. La RI est également une activité qui est

intégrée dans de nombreuses tâches comme l'apprentissage pédagogique (e-learning), la gestion de la mémoire d'entreprise, etc.

La nature des différentes tâches de RI est diverse et implique des traitements de l'information adaptés à chacun des objectifs qu'elles doivent atteindre.

1.2. Qu'est-ce que la connaissance ?

Définir la connaissance en soi relève de la philosophie. Le propos de cette section n'est pas de répondre à une telle question mais de caractériser la connaissance, notre cadre de réflexion étant l'informatique. Nous entendons ici par informatique, non pas « la science des ordinateurs », mais « la science du traitement de l'information ».

Nous situerons tout d'abord la notion de connaissance par rapport aux différentes notions auxquelles elle est associée dans le domaine de l'informatique. Nous définirons ensuite ces différentes caractéristiques et nous expliciterons l'intérêt d'en réaliser une représentation pour permettre sa manipulation.

1.2.1. De l'information à la connaissance :

Il convient tout d'abord de caractériser la connaissance par rapport à plusieurs termes auxquels elle est abusivement assimilée. Même s'il n'existe pas de frontières clairement établies entre les notions de donnée, information, processus et connaissance, chacune de ces notions joue un rôle propre en fonction de son niveau d'entrée dans un processus d'action d'un système informatique [Charlet 2002].
La donnée est le moins porteur de sens de tous ces termes. Tout instrument informatique et technologique crée de l'accumulation de données. Les données ne sont ni vraies, ni fausses, ni significatives à moins d'être récupérées, représentées et réinterprétées. Elles sont transmises à un système ou un programme qui les traite, les modifie et les fait évoluer.

Toute information est issue de données qui sont structurées pour constituer une information. L'information fait référence aux « messages » qui peuvent être restitués par le système et à l'usage des données. Les données deviennent informations quand elles prennent un sens soit pour le système soit pour l'utilisateur.

L'information, constituée des données, devient connaissance à partir du moment où elle sert de fondement à une inférence, au déclenchement d'un processus [Lame 2002]. Une inférence est définie par Kasyer comme *« une façon générique de désigner l'ensemble des mécanismes par lesquels des entrées (perceptives ou non) sont combinées à des connaissances préalables afin d'obtenir des comportements élaborés* ».

1.2.2. Caractéristiques de la connaissance :

Tâchons maintenant de caractériser la connaissance à partir de définitions et de travaux issus de la littérature dans le domaine de l'IC.

«Une connaissance est la capacité d'exercer une action pour atteindre un but.» [Bachimont 2004]

«Il n'y a présomption de connaissance que si la faculté d'utiliser des informations à bon escient est attestée.»… «Tandis que les informations sont exploitées par des processus sans pouvoir modifier leur déroulement, les connaissances sont des données qui influencent le déroulement de processus.» [Kayser 1997].

«La connaissance est l'information organisée qui est applicable à la résolution de problèmes.». [Woolf 1990]

«La connaissance inclut des restrictions implicites et explicites entre objets ainsi que des opérations et des relations, qui permettent de définir des heuristiques générales et spécifiques comme les procédés d'inférences liés à la situation à modéliser» [Sowa 1984].

1.2.2.1. Information active :

Les connaissances sont des informations actives, dans la mesure où elles peuvent influencer le déroulement d'un processus, produire de nouvelles informations ou permettre de prendre des décisions [Furst 2004]. La connaissance est définie dans un cadre bien précis et prend sa signification dans le contexte de son utilisation. On ne peut pas parler de connaissance a priori [Charlet 2002].

1.2.2.2. Interprétée par l'homme :

On ne peut parler de connaissance qu'à partir du moment où l'information manipulée par le système prend un sens pour l'utilisateur, c'est-à-dire qu'il peut établir un lien avec cette information et celle qu'il possède déjà [Charlet 2002].

1.2.2.3. Outil informatique : support de sa genèse et de sa mémorisation :

L'informatique permet la mémorisation et la genèse de la connaissance [Charlet 2002]. En effet, les outils et les supports de stockage informatiques permettent à l'homme de constituer des connaissances, de les accumuler et de les faire évoluer.

1.2.2.4. Théorique ou pratique :

Bachimont distingue la connaissance par rapport à son caractère théorique et son caractère pratique [Bachimont 2004]. La connaissance pratique se réfère à des actions associées à une activité dans le monde matériel. Elle permet la modification physique et matérielle du monde. Elle renvoie au savoir-faire. La connaissance théorique quant à elle correspond à une activité non pas dans le monde mais dans notre représentation du monde. Elle fournit une explication dans un code de représentation.

1.2.2.5. Accessibilité de la connaissance : tacite, explicite et implicite :

Les auteurs Nonaka et Takeuchi [Nonaka 1995] proposent de diviser la connaissance en deux catégories: la connaissance tacite et la connaissance explicite. La connaissance tacite correspond à la connaissance obtenue à travers l'expérience, à la connaissance simultanée (liée à la situation immédiate) et à la connaissance analogue (aptitude physique). La connaissance explicite correspond à la connaissance rationnelle, à la connaissance séquentielle (réaction par rapport à la situation immédiate) et à la connaissance codifiée (production électronique).

Cette catégorisation est étendue chez Liebowitz [Liebowitz 1998] par la proposition d'un troisième niveau de connaissance qui est la connaissance implicite. On accède à la connaissance tacite dans l'esprit humain et dans les organisations à travers un processus d'extraction de connaissance et d'observation de comportements. La connaissance implicite est accessible à partir de consultations et de discussions. Finalement la connaissance explicite se trouve dans les documents et les systèmes informatiques par l'intermédiaire de formalisation de la connaissance.

Ces différentes typologies de la connaissance permettent d'établir les caractéristiques que celle-ci doit avoir dans le domaine de la RI. Il est primordial qu'elle soit interprétable à la fois par le système et par l'utilisateur du SRI. La connaissance vise à mettre en place un dialogue entre ces deux acteurs. La connaissance doit être active dans le processus de recherche en permettant d'une part au système de sélectionner les granules qu'il restitue et d'autre part à l'utilisateur de

situer le contexte et les raisons de cette restitution. De plus, le système peut être enrichi par la connaissance tacite de l'utilisateur comme par exemple ses compétences dans la tâche de recherche qu'il effectue. La spécification de la connaissance implicite présente dans les documents que le SRI manipule peut également l'aider dans la restitution des documents.

Afin que cette connaissance soit intégrée au SRI, elle doit être acquise. La section suivante décrit cette problématique.

1.2.3. De l'acquisition à l'ingénierie :

La connaissance qui peut être fournie à un système informatique a tout d'abord besoin d'être capturée et modélisée. Le domaine de l'Ingénierie des Connaissances (IC) a une finalité applicative reposant sur cette problématique. L'IC est définie dans Charlet [Charlet 2000] comme «*l'étude des concepts, méthodes et techniques permettant de modéliser et/ou acquérir les connaissances pour des systèmes réalisant ou aidant des humains à réaliser des tâches ne se formalisant a priori pas ou peu* ».

L'IC a pris la place du domaine de l'Acquisition des Connaissances à partir des années 80. L'évolution des procédés liés à l'acquisition de connaissances peut s'analyser à travers l'évolution de ces deux domaines de recherche.

Acquérir des connaissances est une tâche difficile, dont l'objectif est d'expliciter et de capturer des connaissances explicites sans introduire de biais. Avant la naissance du domaine de l'IC, la connaissance utilisée par les Systèmes Experts était celle d'un expert qui la codait directement dans un langage de représentation. Cette démarche a été remise en cause pour laisser place à de nouveaux systèmes reposant sur une connaissance construite coopérativement avec un ou plusieurs experts à partir d'un modèle de la connaissance. Dans les années 1990, le modèle était un modèle réel du monde tel qu'il était observé à partir des connaissances des experts du domaine ou d'autres sources. Après 1990, un modèle n'a plus uniquement pour objectif de représenter une observation du monde mais une interprétation liée à l'opération que l'on souhaite faire avec la connaissance. Un modèle est alors une abstraction qui permet de réduire la complexité de la modélisation en se focalisant sur certains aspects en fonction du but à atteindre. Le modèle conceptuel est alors défini pour manipuler des objets du monde ainsi que l'interprétation des résultats de la manipulation. Un modèle conceptuel en IC repose sur trois niveaux de connaissance [Shadbolt 1993]. Il exprime tout d'abord comment une tâche va être effectuée. Il utilise également la connaissance d'un domaine qui définit les concepts à manipuler et leurs relations. Finalement, le modèle explicite la manière dont le système résout le problème à partir de la connaissance qu'il utilise.

1.2.4. Représentation de la connaissance :

Le processus d'ingénierie des connaissances définit des étapes pour organiser les connaissances au sein de représentations formelles. Un modèle conceptuel de la connaissance est ensuite traduit en une représentation qui pourra être manipulée par les systèmes informatiques.

Représenter la connaissance a pour objectif de modéliser la connaissance en omettant certains détails non significatifs pour en permettre une meilleure manipulation [Kayser 1997]. Cette question est au cœur des travaux en Intelligence Artificielle. La représentation ne correspond pas à l'entité dans son intégralité. Prenons par exemple une carte routière, son intérêt est de représenter une région afin de pouvoir prévoir un déplacement. Une carte à taille réelle n'aurait aucun intérêt.

Une représentation est une structure composée de symboles construite à partir d'un ensemble de règles de formation [Kayser 1997]. L'ensemble des règles de formation est défini par le langage de représentation choisi. Un ordinateur gère des symboles, il est médiateur de la connaissance, comme l'est un livre. L'utilisateur de l'ordinateur accède, lui, à la sémantique associée à la représentation [Bachimont 1999]. Pour la RI, cette représentation doit intégrer les termes permettant de détecter la connaissance dans les documents.

La représentation des connaissances utilisée dans les Systèmes Experts reposait sur des règles logiques. Le domaine de l'IC a dépassé la problématique des Systèmes Experts pour proposer de nouveaux formalismes pouvant représenter la richesse sémantique de la connaissance en amont de sa représentation formelle opérationnelle. La représentation de la connaissance s'appuie alors sur des représentations au niveau conceptuel pouvant modéliser la « structure cognitive » d'un domaine [Guarino 1994]. Par niveau conceptuel, on entend ici une formalisation sur la description des connaissances avant de se préoccuper de la manière dont un système inférentiel pourra les traiter. Les langages à base de Frame [Minsky 1975], les logiques de description [Brachman 1985] et les graphes conceptuels [Sowa 1984] sont des langages permettant ces représentations. Ces langages seront décrits plus tard. Ils ont en commun de donner priorité au pouvoir d'expression par rapport à la capacité de raisonnement logique. Ils permettent de représenter pour un domaine de connaissance donné, les concepts, les relations entre les concepts, ainsi que la sémantique de ces relations.

1.3. Représentation de la connaissance et ontologie :

Une ontologie fournit une base solide pour la communication entre les machines mais aussi entre humains et machines en définissant le sens des objets tout d'abord à travers les symboles (mots ou expressions) qui les désignent et les caractérisent et ensuite à travers une représentation structurée ou formelle de leur rôle dans le domaine [Aussenac-Gilles 2004].

Les ontologies sont utilisées dans de nombreux domaines. Les domaines recensés en 1998 par Guarino [Guarino 1998] sont l'ingénierie des connaissances, la modélisation qualitative, l'ingénierie des langages, la conception de bases de données, la recherche d'information, l'extraction d'information, la gestion et l'organisation de connaissances. Depuis, grâce à l'essor du Web, elles sont utilisées dans le domaine de l'e-commerce et sont au cœur du Web Sémantique [Berners-Lee 2001], future version du Web actuel. Un des plus grands projets reposant sur l'utilisation des ontologies consiste à ajouter au Web une véritable couche de connaissance permettant des recherches d'information au niveau sémantique et non plus au simple niveau lexical et/ou syntaxique. A terme, il est prévu que des applications déployées sur l'Internet pourront mener des raisonnements utilisant les connaissances stockées sur la toile.

Derrière l'utilisation d'ontologies dans ces différents domaines, se cachent en fin de compte plusieurs représentations de connaissances. Ces représentations peuvent être distinguées suivant deux axes : la nature de la connaissance représentée dans l'ontologie et le degré d'engagement sémantique qui a motivé la formalisation de l'ontologie. Le premier axe fait en particulier référence au type de connaissances représentées (génériques, de domaines ou liées à la tâche). Le second axe fait en particulier référence au niveau sémantique des connaissances que l'ontologie représente (ressource terminologique versus ressource conceptuelle). Nous présentons ces deux aspects : nature des connaissances et engagement sémantique dans ce qui suit.

1.3.1. Nature des connaissances :

La première distinction à faire sur les représentations de connaissance associées à la notion d'ontologie repose sur la nature des connaissances représentées dans l'ontologie. La nature de ces connaissances peut varier soit par rapport à la structure de la connaissance soit par rapport au contenu de la connaissance.

1.3.1.1. Différentes structures de la connaissance :

La connaissance contenue dans l'ontologie peut représenter plusieurs structures. La classification décrite chez Heijst [Heijst 1997] distingue trois types d'ontologies suivant ce critère.

- Les ontologies terminologiques ou linguistiques spécifient les termes utilisés pour représenter la connaissance d'un domaine. Un exemple de ce type d'ontologie est le réseau sémantique UMLS (Unified Medical Language System) [Lindberg 1993].

- Les ontologies de l'information spécifient la structure des enregistrements d'une base de données. Les schémas de base de données en sont un exemple. Elles proposent un cadre de représentation de la connaissance stockée mais ne spécifient pas de détails sur la sémantique des champs.

- Les ontologies pour la modélisation de la connaissance spécifient la conceptualisation de la connaissance. Ces ontologies ont une structure beaucoup plus riche que celle des deux autres types. Elles sont généralement conçues en fonction de l'utilisation prévue de la connaissance qu'elles contiennent.

Pour la RI, la structure de connaissances utile se situe entre celle des ontologies terminologiques et celle des ontologies pour la modélisation de la connaissance. Elles ont pour but de définir les termes liés à la connaissance pour que celle-ci soit décelable dans les documents. Mais elles doivent également permettre d'interpréter la connaissance à partir d'un niveau conceptuel afin que des mécanismes élaborés puissent être intégrés au SRI.

1.3.1.2. Différents contenus :

Un autre critère pour la classification des ontologies est le contenu de la connaissance qu'elles représentent, c'est-à-dire le sujet de la conceptualisation [Guarino 1998].

1.3.1.2.1. Les ontologies génériques :

Définissent des concepts considérés comme génériques à plusieurs domaines. WordNet [Miller 1988] par exemple est une ontologie dont le but est de représenter la langue naturelle anglaise. WordNet est un système de références lexicales dont la conception a été inspirée par les théories de la mémoire linguistique humaine. Elle est composée d'ensembles de synonymes appelés *synsets*, où chaque terme est regroupé en classes d'équivalence sémantique. Chaque ensemble de synonymes représente un concept particulier. Chaque terme appartient de plus à une catégorie lexicale donnée (nom, verbe, adverbe, adjectif). Un terme peut appartenir à plusieurs synsets et à plusieurs catégories lexicales. Les ensembles de synonymes sont associés par des relations sémantiques : généricité/spécificité, antonymie (relation entre ensembles de mots qui, par leur sens, s'opposent). WordNet couvre le domaine de la langue générale en intégrant le sens des mots dans différents domaines. Par exemple, la figure 6 présente l'ensemble des différents sens retrouvés pour le mot **dispersion**

1. **dispersion**, scattering -- (spreading widely or driving off)

2. distribution, **dispersion** -- (the spatial property of being scattered about over an area or volume)

3. **dispersion**, dispersal, dissemination, diffusion -- (the act of dispersing or diffusing something; "the dispersion of the troops"; "the diffusion of knowledge")

Les sens répertoriés renvoient au sens dans le langage courant (sens 1 et 3) ainsi qu'au sens du mot dans le domaine scientifique, plus précisément le domaine de la physique (sens 2).

Figure 6: Ensemble des différents sens du mot dispersion dans WordNet [Hernandez 2005]

D'après Charlet [Charlet 2000], la limite de ces ontologies générales est leur difficile réutilisation car elles ont pour objectif de recouvrir tous les sens des mots et ne normalisent pas leur sens.

La normalisation sémantique consiste à organiser au sein d'un modèle conceptuel des connaissances, à partir de la compréhension du domaine et de l'application visée. Cela revient à associer aux termes une signification qui fait abstraction des variations de sens liées à d'autres domaines. Cette abstraction du contexte conduit à construire des concepts, considérés en tant que «signifiés non contextuels», normés au sens où ils sont décrits selon un certain point de vue (celui de la tâche, qui fixe un contexte de référence).

1.3.1.2.2. Les ontologies de domaine :

Sont des conceptualisations spécifiques à un domaine particulier. Les méthodes actuelles d'acquisition de la connaissance font la distinction explicite entre connaissance du domaine et ontologie du domaine. La connaissance du domaine décrit des situations factuelles du domaine alors que l'ontologie pose des contraintes sur la structure et le contenu de la connaissance du domaine. Comparées aux ontologies génériques, les ontologies de domaine ont pour avantage de permettre une normalisation des concepts dans le cadre du domaine considéré et donc de permettre une meilleure représentation de la connaissance. L'ontologie Ménélas [Zweigenbaum 1993] est un exemple d'ontologie de domaine, celui des maladies coronariennes, rassemblant des concepts et leurs relations structurés à partir de la relation «sorte de». Ménélas comprend également des lexiques sémantiques et morphosyntaxiques des mots simples et composés. Cette ontologie est dédiée à l'analyse automatique de comptes rendus d'hospitalisation.

1.3.1.2.3. Les ontologies d'application :

Contiennent toutes les définitions qui sont nécessaires pour modéliser la connaissance propre à l'élaboration d'une tâche particulière. Généralement, les ontologies d'application combinent des éléments d'ontologies de domaine et d'ontologies génériques choisies en fonction des méthodes spécifiques pour réaliser la tâche visée. Elles sont rarement réutilisables pour une autre application.

1.3.1.2.4. Les ontologies de représentation de la connaissance :

Permettent d'expliquer la conceptualisation sous-jacente aux formalismes de représentation [Davis 1993]. Elles proposent un cadre de représentation sans émettre d'hypothèse sur le monde. On les désigne également comme ontologies abstraites ou de haut niveau parce qu'elles permettent de définir des concepts abstraits et peuvent être réutilisées pour définir des concepts spécifiques. Un exemple d'ontologie de ce type est la *Frame Ontology* utilisée dans Ontolingua

[Gruber 1993a]. Ces ontologies permettent de normaliser la connaissance manipulée par le système par rapport à la connaissance qui lui est utile.

1.3.2. Engagement sémantique :

La deuxième distinction à faire sur les représentations de connaissance induites par la notion d'ontologie repose sur le degré d'engagement sémantique de celles-ci. Le degré d'engagement sémantique correspond au niveau de spécification formelle permettant de restreindre l'interprétation de chaque concept et ainsi d'en donner la sémantique [Bachimont 2000].

Afin d'expliciter ce niveau d'engagement puis de décrire chacune de ces représentations, il est important de les distinguer en définissant plusieurs notions impliquées dans leur degré de formalisation. Elles sont distinguées suivant deux principes : les ressources faisant intervenir des termes et les ressources composées de concepts.

1.3.2.1. Notions sous-jacentes :

Afin d'expliciter les engagements sémantiques dans la formalisation d'ontologies, il convient de préciser les notions de concept, relation, subsomption et axiome.

1.3.2.1.1. Concept :

Un concept se définit par Bachimont à trois niveaux [Bachimont 2004]. Un concept est une signification. Sa place dans un système de significations permet de le comprendre, de le distinguer et de le différencier par rapport à d'autres concepts. Un concept est une construction. Comprendre un concept revient à construire l'objet dont il est le concept. Un concept est une prescription. On le comprend en exécutant l'action qu'il entreprend.
Uschold [Uschold 1995] partage ce point de vue et définit de façon plus pragmatique la notion de concept. Un concept représente pour un objet matériel, une notion ou une idée. Il est composé de trois parties : un ou plusieurs **termes**, une **notion** et un **ensemble d'objets**. La notion correspond à la sémantique du concept, elle est définie à travers ses propriétés et ses attributs. La notion est appelée **intention** du concept. L'ensemble d'objets correspond aux objets définis par le concept, il est appelé **extension** du concept ; les objets sont les **instances** du concept. Le ou les **termes** permettent de désigner le concept. Ces termes sont aussi appelés **labels** de concept. Par exemple, le terme « lapin » renvoie à un animal possédant de longues oreilles et quatre pattes et à l'ensemble des objets ayant cette description. Afin que les concepts soient reconnus de façon non ambiguë par la machine, il est souhaitable qu'un concept soit identifié à partir de plusieurs termes, ce qui permet de gérer la synonymie et de les désambiguïser les uns par rapport aux autres [Gómez-Pérez 1996].

Un concept est défini à partir d'une sémantique référentielle (due à son extension) et une sémantique différentielle (due à son intention). Un concept peut avoir une extension vide, c'est le cas des concepts génériques ou abstraits comme par exemple « la vérité ». Deux concepts peuvent avoir la même extension et des intentions différentes. C'est le cas par exemple des « lapins » considérés comme « animaux de compagnie » ou bien comme « ressource culinaire ». Il est considéré par certains auteurs que l'intention d'un concept permet à elle seule de définir le sens d'un concept [Guarino 1994]; d'autres auteurs considèrent que le sens dépend de l'intention et de l'extension du concept [Kassel 1999].

1.3.2.1.2. Relation sémantique :

Une relation sémantique R représente un type d'interaction entre les concepts d'un domaine c_1, c_2, .., c_n. Elle se définit formellement à partir d'un produit de n concepts : $R : c_1 \times c_2 \times \ldots \times c_n$; « subsume », « est un phénomène lié à » sont des exemples de relations binaires.

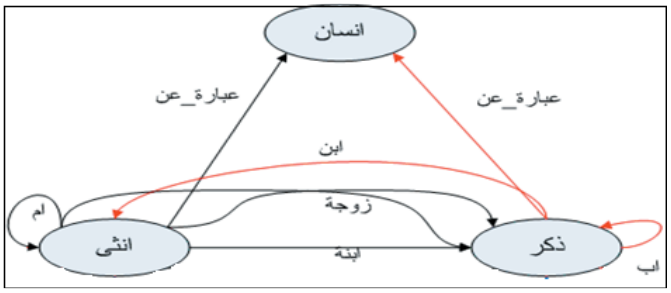


Figure 7: Les Relations [Zaidi 2008]

Les relations les plus courantes dans la littérature sont les relations d'équivalence, taxonomiques, patronymiques, de dépendance, topologique, causale, fonctionnelle, chronologique [Gómez-Pérez 2001].

1.3.2.1.2.1. Relation taxonomique (ou subsomption) :

La notion de subsomption (aussi appelée relation « est un », relation taxonomique ou relation de spécificité/généricité) est une relation binaire particulière qui implique l'engagement sémantique suivant [Guarino 2001]:

Un concept c1 subsume un concept c2 si toute relation sémantique de c1 est aussi relation sémantique de c2, en d'autres termes si le concept c2 est plus spécifique que le concept c1. Les instances se rapportant au concept c2 seront des instances de c1, par contre une partie seulement des instances de c1 seront des instances de c2. La notion abordée par le concept c2 (intention du concept) sera plus précise que celle abordée par c1. La relation de subsomption permet d'organiser hiérarchiquement un ensemble de concepts.

La relation de subsomption est une **relation d'ordre partiel** définie à partir des propriétés suivantes:

1.3.2.1.2.1.1. L'asymétrie :

Cette propriété signifie que l'inclusion d'une classe d'individus X dans une classe d'individus Y implique que Y n'est pas incluse dans X.

Formellement, cette propriété garantit : X subsume Y, si et seulement si non (Y subsume X),

1.3.2.1.2.1.2. La transitivité :

Soit une classe d'individus X qui subsume une classe Y, qui elle-même subsume Z, alors X subsume Z.

Formellement : (X subsume Y) et (Y subsume Z) ⇒ (X subsume Z)

1.3.2.1.2.1.3. La non réflexivité :

Cette propriété implique qu'un fait décrit par la relation « est un » ne peut pas s'écrire de plusieurs façons.

Formellement : non (X subsume X)

L'**héritage multiple :** est une propriété qui peut être définie sur la relation de subsomption : un concept d'une ontologie peut avoir plusieurs pères par la relation de subsomption. L'héritage multiple implique que le concept hérite des propriétés de tous ses pères.

1.3.2.1.2.2. Relation associative :

Les relations « associatives » sont des relations d'interaction entre deux concepts qui ne sont pas la relation de subsomption. La désignation « relation associative » est empruntée aux domaines de la bio-informatique [Zhang 2004], ce domaine ayant une utilisation équivalente

des ontologies par l'indexation de publications et de comptes rendus biologiques. Elles correspondent à la notion de rôle en Logique de Description et permettent de typer les concepts reliés. Ces relations sont soit à des propriétés entre concepts soit à des propriétés d'attribut dans le cas où elles associent un concept à un type de données. La sémantique qui leur est associé est référencée par un label. Elle peut également être précisée à partir de propriétés logiques associées à la relation (la transitivité, la symétrie, la fonctionnalité…).

1.3.2.1.3. Axiome :

Les axiomes ont pour but de définir dans un langage logique la description des concepts et des relations permettant de représenter leur sémantique. Ils représentent les intentions des concepts et des relations du domaine et, de manière générale, les connaissances n'ayant pas un caractère strictement terminologique [Staab 2000]. Les axiomes sont des expressions qui sont toujours vraies.

Leur inclusion dans une ontologie peut avoir plusieurs objectifs: définir la signification des composants, définir des restrictions sur la valeur des attributs, définir les arguments d'une relation, vérifier la validité des informations spécifiées ou en déduire de nouvelles.

La figure 8 présente des exemples d'axiomes formalisés à partir du langage OWL-Lite.

```
<owl:Class rdf:ID="peugeotThings">
<owl:equivalentClass>
<owl:Restriction>
<owl:onProperty rdf:resource="#locatedIn" />
<owl:someValuesFrom rdf:resource="#peugeotvehicule" />
</owl:Restriction>
</owl:equivalentClass>
</owl:Class>
```

Dans ce premier exemple, le concept peugeot Things est restreint aux vehicules de marque peugeot par l'axiome en gras.

```
<owl:AllDifferent>
<owl:distinctMembers rdf:parseType="Collection">
<voiture: voitureCouleur rdf:about="#Rouge" />
< voiture: voitureCouleur rdf:about="#Blanc" />
< voiture: voitureCouleur rdf:about="#Rose" />
</owl:distinctMembers>
</owl:AllDifferent>
```

Dans cet exemple, l'axiome permet de définir une classe composée des différentes couleurs de voiture.

Figure 8: Exemples d'axiomes formalisés à partir de OWL-Lite[owl-guide]

1.3.2.2. Ressources terminologiques :

Les ressources terminologiques ne définissent pas des concepts mais des termes. Contrairement à leur utilisation dans la définition des concepts, dans ces ressources, les termes ne font pas référence à des notions (intention du concept) et des objets (extension du concept) mais définissent uniquement le vocabulaire lié à la connaissance représentée. La notion de terme est intentionnellement choisie par rapport à la notion de mot. La notion de mot désigne une unité

textuelle, alors que la notion de terme fait référence à l'ensemble des variantes lexicales d'un mot ou d'un groupe de mots.

1.3.2.2.1. Vocabulaire contrôlé :

Soit T un ensemble de termes d'un domaine. Un vocabulaire VC contrôlé est défini par VC={t1,t2,..,tn / ti∈T}.

Un vocabulaire contrôlé est un ensemble de termes définis par un groupe de personnes ou une communauté. La signification des termes n'est pas forcément définie et il n'y a pas d'organisation logique entre les termes [Lassila 2001]. Ce vocabulaire peut être utilisé afin de labelliser des contenus documentaires. Les catalogues sont des exemples de vocabulaires contrôlés.

1.3.2.2.2 Glossaires :

Soit T un ensemble de termes d'un domaine, soit D un ensemble de définitions en langage naturel. Un glossaire G est défini par le couple (T,D).

Un glossaire est un ensemble de termes avec leur signification. La définition de chaque terme est donnée en langage naturel. Cette représentation apporte plus d'informations car une personne peut lire la définition, cependant elle n'est pas interprétable par l'ordinateur [Lassila 2001].

1.3.2.2.3 Hiérarchie informelle :

Soit T un ensemble de termes et R une relation de T x T où R(t1,t2) signifie que le terme t1 est plus général que le terme t2 ou que le terme t2 est plus spécifique que le terme t1. Une hiérarchie informelle est définie par le couple (T,R).

Les hiérarchies informelles sont des hiérarchies explicites organisant des catégories à partir de la notion générale de généralisation / spécification. Elles ont fait leur apparition sur le Web comme par exemple la hiérarchie proposée par Yahoo. Cependant, ces hiérarchies ne sont pas formelles car la hiérarchisation des catégories ne respecte pas la stricte notion de subsomption. Tout d'abord les termes employés pour désigner les catégories ne permettent pas de définir clairement le sens de la catégorie. Prenons par exemple l'extrait de la hiérarchie Yahoo suivante :

Accueil > Mode & Accessoires > Pour la Femme > Tous les Accessoires Femme > Pierres / Perles > Perle.

Le terme Perle définissant la catégorie la plus profonde dans cette branche de la hiérarchie ne fait pas référence à la notion de perle dans son ensemble ; elle devrait être désignée par les termes « accessoires féminins contenant des perles ».

De plus, au sens strict de la subsomption, les individus de cette dernière catégorie devraient avoir les mêmes propriétés sémantiques que celles de la classe Pierres/Perles. Ce type de hiérarchies regroupant des catégories disjointes (accessoires en pierre ou en perles) rend problématique l'héritage des propriétés.

1.3.2.2.4 Thésaurus

Soit T un ensemble de termes et $\mathfrak{R}$ un ensemble de relations de T x T. Un thésaurus est défini par le couple (T, $\mathfrak{R}$).

Un thésaurus est un ensemble de termes organisés suivant un nombre restreint de relations [Foskett 1980]. Les relations entre termes les plus typiques sont présentées dans la figure 9 [Foskett 1980] [Miles 2005] [Soergel 2004]. Elles définissent des relations entre termes

synonymes (terme préféré, terme à utiliser à la place de), entre termes préférés (terme plus spécifique, terme plus générique, terme lié à). Afin d'uniformiser leur format de représentation, différentes normes spécifient les thésaurus monolingues (ISO 2788:1986, AFNOR NF Z47-100:1981 , ANSI Z39) et multilingues (AFNOR NF Z47-101 :1990, ISO 5964 :1985).

t_1 **Terme préféré dans** $t_2, t_3,\ldots,t_N$	t_1 est le terme préféré pour désigner l'ensemble des synonymes $t_2, t_3,\ldots,t_N$.
t_1 **Note** texte	Remarque sur le terme t_1 (usage exceptionnel, contexte d'utilisation)
t_1 **Utiliser plutôt** t_2	t_2 est utilisé pour désigner t_1
t_1 **Utilisé pour** t_2	t_1 est utilisé pour désigner t_2
t_1 **Plus spécifique que** t_2	Le terme désigné par t_1 est plus spécifique que le terme désigné par t_2
t_1 **plus générique que** t_2	Le terme désigné par t_1 est plus générique que le terme désigné par t_2
t_1 **est lié à** t_2	t_1 est un terme lié ou associé à t_2

Figure 9: les relations entre termes les plus typiques dans un thésaurus

Les thésaurus sont principalement utilisés pour assister les documentalistes dans la tâche d'indexation manuelle de documents. Ils sont reconnus pour présenter différents avantages dans ce contexte [Foskett 1977]. Ils offrent tout d'abord une vue générale sur les termes et relations d'un domaine. Ils définissent ensuite un vocabulaire standardisé pour l'indexation. Ils permettent d'assurer qu'un seul terme d'un ensemble de synonymes soit choisi pour l'indexation (terme dit « à utiliser »). Ils sont également utilisés lors de la spécification d'une requête pour spécifier ou généraliser une recherche documentaire à partir des termes dits plus spécifiques ou plus génériques.

De nombreux auteurs [Tudhope 2001] [Soergel 1974] [Fischer 1998] considèrent que les relations de généricité /spécificité définissant la hiérarchie des thésaurus ne suivent pas l'engagement sémantique impliqué par la relation de subsomption. Ils considèrent que les relations *termes spécifiques, termes plus génériques* regroupent différentes relations sémantiques telles que les relations de généricité mais aussi « partie de » et « instance de ». Fischer explique cette ambiguïté par le fait que la définition de ces relations « terme plus spécifique », « terme plus générique » est orientée par l'utilisation faite des thésaurus, c'est-à-dire l'aide au travail du documentaliste (indexation, recherche), et non par la formalisation de la connaissance du domaine [Fischer 1998]. Il prend comme référence la définition donnée chez Soergel [Soergel 1974] : *« Le terme A est considéré comme étant plus générique que le terme B si pour toute recherche inclusive sur le terme A tous les éléments traitant de B doivent être retrouvés. Inversement B est plus spécifique* ».

La définition introduit donc de la subjectivité et implique un jugement de l'expert sur le résultat d'une recherche. Les thésaurus sont depuis de nombreuses années utilisés en RI [Baeza-Yates 1999]. Cependant, par le manque de formalisation et l'objectif de leur conception (l'aide aux documentalistes), ils présentent un degré d'ambiguïté et les termes qui les composent doivent être interprétés par une personne pour pouvoir capturer la sémantique implicite qu'ils sous-entendent.

Les thésaurus à facettes sont une catégorie de thésaurus. Les termes sont alors organisés suivant plusieurs hiérarchies mutuellement exclusives représentant chacune une facette du domaine représenté. Les termes appartiennent donc à une seule facette mais des relations non hiérarchiques peuvent exister entre les différents termes des différentes facettes [Spiteri 1999]. Le thésaurus AAT [getty] est un exemple de ce type de thésaurus. Il représente le domaine de l'architecture périodes, les agents, les activités, les matériaux et les objets.

1.3.2.3. Ressources conceptuelles :

Les ressources conceptuelles [Hernandez 2005] témoignent d'un engagement sémantique qui repose sur la notion de concepts. Différents niveaux sémiotiques sont à prendre en considération dans une ressource conceptuelle [Maedche 2002]. Le niveau lexical couvre tous les termes ou labels définis pour désigner les concepts. Le niveau conceptuel représente les concepts et la sémantique qui leur est associée à partir des relations conceptuelles entre eux. Une ressource conceptuelle est définie à partir d'une structure qui décrit son niveau conceptuel et d'un lexique correspondant au niveau lexical.

1.3.2.3.1. Hiérarchie de concepts :

Une structure d'une hiérarchie de concepts est le couple $S_H\{C, \leq^C\}$

Où : C'est un ensemble de concepts,

$\leq^C$: C x C est un ordre partiel sur C, il définit la hiérarchie de concepts
$\leq^C (c_1, c_2)$ signifie que c_1 subsume c_2 (relation orientée)

Le lexique d'une hiérarchie de concepts est le couple $L_H : \{L^C, F\}$

F est une fonction appelée référence tel que $F \rightarrow L^c$ pour les concepts,

- Pour $l \in L^C$, $F(l)=\{c / c \in C\}$
- Pour $c \in C$, $F^{-1}(c)=\{l / l \in L^C\}$

Une hiérarchie de concepts est le couple (S_H, L_H).

1.3.2.3.2. Ontologie dites « légères » :

La structure et le lexique définissant une ontologie légère sont les suivants :

La structure est un tuple $S_O := \{C, R, A, T, \leq^C, \sigma_R, \sigma_A,\}$ où :

- C, R, A, T sont des ensembles disjoints contenant les concepts, les relations associatives, les relations d'attribut, les types de données
- $\leq^C$: C x C est un ordre partiel sur C, il définit la hiérarchie de concepts
  $\leq^C (c_1, c_2)$ signifie que c_1 subsume c_2 (relation orientée)
- $\sigma_R : R \rightarrow C \times C$ est la signature d'une relation associative
- $\sigma_A : A \rightarrow C \times T$ est la signature d'une relation d'attribut

Le lexique est un tuple $L_O : \{L^C, L^R, F, G\}$

- L^c et L^R sont des ensembles disjoints des labels (termes) des concepts, des instances et des relations

- F et G sont deux relations appelées référence,
  $F \rightarrow L^c$ pour les concepts, $G \rightarrow L^r$ pour les relations :
  - Pour $l \in L^c$, $F(l)=\{c / c \in C\}$
  - Pour $c \in C$, $F^{-1}(c)=\{l / l \in L^C\}$
  - Pour $l \in L^R$, $G(l)=\{r/r \in R\}$
  - Pour $r \in R$, $G^{-1}(r)=\{l / l \in L^R\}$

Ces relations permettent d'accéder aux concepts, relations et instances désignés par un terme et réciproquement.

Une ontologie légère est le couple $O=(S_O,L_O)$. Elle est dite « légère » car l'engagement sémantique qu'elle suit n'est pas totalement formel dans la mesure où aucun axiome n'est spécifié. Cependant de telles ontologies présentent différents avantages. Elles sont facilement interprétables par l'homme. Leur construction, leur vérification et leur mise à jour demandent moins d'effort. Enfin, il est plus facile de trouver un consensus lors de leur spécification [Kiryakov 2004]. Contrairement aux ontologies lourdes, les ontologies légères ne possèdent pas d'axiomes.

1.3.2.3.3. Ontologies lourdes :

La structure d'une ontologie lourde est un tuple $S_O := \{C, R, A, T, A_x, P_{log}, \leq^C, \sigma_R, \sigma_A\}$.

Ce tuple est défini à partir de la structure d'une ontologie légère à laquelle est ajouté un ensemble d'axiomes A_x. Les axiomes sont décrits à partir des primitives logiques P_{log} définies par le langage logique considéré.

Une ontologie lourde est définie par le couple (S_O,L_O)

TOVE [Gruninger 1995b] et PIF [Lee 1995] sont des exemples d'ontologies rigoureusement formelles. Leur avantage repose sur la réduction considérable des interprétations possibles des concepts et donc la minimalisation des ambiguïtés. Cependant, elles demandent de lourds efforts de conception [Uschold 2003] et ne peuvent couvrir que des domaines précisément définis.

1.3.2.3.4. Modèle d'un domaine :

Afin de compléter la connaissance liée à un domaine, le modèle d'un domaine est représenté à partir d'une ontologie légère ou lourde O.

Il se formalise par le tuple suivant $(O, I, V, f_C, f_T, f_R, f_A)$ avec :

- O l'ontologie de domaine considéré ;
- I l'ensemble des instances ;
- V l'ensemble des valeurs des types de données ;
- $f_C : I \rightarrow C$ est la fonction d'instanciation d'un concept ;
- $f_T : V \rightarrow T$ est la fonction d'instanciation d'un type de donnée ;
- $f_R : I \times I \rightarrow R$ est la fonction d'instanciation d'une relation associative ;
- $f_A : I \times V \rightarrow A$ est la fonction d'instanciation d'une relation d'attribut.

1.3.3. Langages de représentation des ontologies conceptuelles :

L'objectif étant aussi la conception et la prise en compte par le SRI de ressources conceptuelles, nous avons analysé les différents langages de représentation des ontologies conceptuelles.

Les langages dédiés aux ontologies sont principalement issus des formalismes liés aux réseaux sémantiques. Nous les décrivons dans la section suivante. Nous nous concentrons ensuite sur les langages de représentation qui en sont issus.

1.3.3.1. Réseaux sémantiques et langages associés :

1.3.3.1.1. Réseau sémantique :

Un réseau sémantique Est une représentation graphique d'une conceptualisation d'une (ou plusieurs) connaissance humaine [Quillian 1968]. Il est représenté sous la forme d'un graphe étiqueté et orienté. Un arc lie un nœud de départ à un nœud d'arrivée. Chaque nœud peut être relié par un ou plusieurs arcs. Les inférences possibles dépendent de la nature des liens. Cependant, ce type de définition ne concerne que la structure du graphe et ne permet pas d'ajouter de l'information sémantique. De nombreuses études [Woods 1975] [Brachman 1977], ont montré que ce type de graphe manque de précision sémantique et mène à des confusions entre les relations et aussi entre les classes et individus. Elles ont mené à la définition de nouveaux formalismes tels que les frames, les logiques de description et les graphes conceptuels.

1.3.3.1.2. Les frames :

Minsky [Minsky 1975] a présenté comme étant une structure de données capable de représenter des objets structurés. Un frame représente donc une classe ou un objet. Les frames sont organisés dans une hiérarchie suivant un lien de spécification. Les composants du frame sont appelés «slots», ils sont considérés comme des attributs de la structure. Ils peuvent être de plusieurs natures : valeur de l'attribut (qui peut être vide), ensemble de valeurs, restriction de valeurs, valeur par défaut, une propriété avec un autre frame, une combinaison des différents cas. L'intérêt des frames est qu'ils permettent de représenter la façon de penser d'experts en fournissant une représentation structurée et concise des relations utiles [Fikes 1985]. L'information peut être partagée entre plusieurs frames grâce à l'héritage.

Chaise	
Est-un-sous-ens de:	Meuble
Est-un-super-ens de:	Tabouret
Nombre de pied:	4
Fonction:	Permet à quelqu'un de s'assoir

Figure 10: Les Frames [Minsky 1975]

1.3.3.1.3. Les logiques de description :

Issues des frames reposent sur trois notions de base : les concepts représentant des classes (ensemble d'objets), les rôles (relations liant deux objets) et les individus (objets représentant les classes qu'ils instancient). Pour décrire ces éléments, deux structures sont utilisées : la *T-BOX* et la *A-BOX*. La *T-BOX* (boîte terminologique) comprend la description des concepts et des rôles. Cette description est structurée à l'aide du lien hiérarchique *sorteDe*. Deux concepts particuliers figurent au minimum dans la *T-BOX* : le concept le plus générique (anything) et le concept le plus spécifique (nothing). La *A-BOX* (boîte assertionnelle) est constituée des individus, de leur description et des règles qui leur sont attachés. Les inférences reposent sur la reconnaissance d'instances de concepts à partir de leur définition, la détection des concepts plus généraux ou plus spécifiques, et la classification ordonnant les concepts dans la hiérarchie. Les logiques de description sont plus flexibles que les frames et reposent sur une sémantique et une

syntaxe rigoureuses [Baader 1991]. Elle est utilisable en RI car elle permet de traiter des données erronées ou incomplètes tout en offrant la possibilité d'ordonner hiérarchiquement les données. Cependant, elles nécessitent l'élaboration manuelle de ressources de connaissances formalisées à partir de cette logique.

1.3.3.1.4. Les graphes conceptuels :

Ont été présentés par Sowa en 1984 [Sowa 1984] et utilisent une notation à base de graphes. Ils ont été définis comme un langage pivot entre le langage naturel et la logique du premier ordre. Ils visent à formaliser les relations entre prédicats et arguments dans une phrase. Ils sont composés de deux types de nœuds étiquetés : les nœuds concepts et les nœuds relations. Les nœuds concepts et les nœuds relations sont respectivement typés par des types de nœuds et des types de relations, organisés suivant un ordre partiel. Les graphes conceptuels peuvent être vus comme des schémas permettant de représenter graphiquement des formules logiques, ou bien des schémas sans contraintes, servant juste d'interface « graphique » à la représentation de formules ou bien comme des graphes munis d'opérations de graphes permettant le raisonnement et leur manipulation en s'appuyant sur la théorie des graphes. Les graphes conceptuels ont été utilisés dans les systèmes d'information pour la représentation de requêtes et de documents chez Guarino [Guarino 1999]. Ils sont élaborés manuellement et une ressource lexicale (Wordnet) est utilisée pour les mettre en correspondance avec les requêtes de l'utilisateur.

1.3.3.2. Langages de représentation d'ontologie :

Différents langages de spécification d'ontologies issus des formalismes précédemment présentés sont apparus à partir des années 1990, tels que CycL et KIF [Genesereth 1994], LOOM [MacGregor 1991], F-Logic [Kifer 1995] et OCML.

1.3.3.2.1. XML [Bradley 2001] :

Est un langage permettant de générer des balises pour la structuration de données et de documents. Il permet la représentation et l'échange de documents semi-structurés.

XML-schéma [Fallside 2001] **:**
Permet de définir la structure, les contraintes, et la sémantique de documents XML. Ce langage n'est pas vu comme un langage d'ontologies car il a été créé pour vérifier la structure de documents XML. Les primitives qu'il met en place sont plutôt orientées application que concept. En effet, la sémantique définie dans le document est interprétable dans le contexte de l'opération faite sur le document mais ne permet pas d'établir des inférences en dehors de ce contexte. XML et XML-schéma sont considérés comme des langages définissant le format de « message » alors qu'un langage d'ontologies a pour but de « représenter » la connaissance.

1.3.3.2.2. RDF [Lassila 1999] :

Permet d'encoder, d'échanger et de réutiliser des méta-données structurées. Il a été créé pour gérer les méta-données de documents XML mais peut également être utilisé pour des ontologies. Il permet de définir des ressources avec des propriétés et des états.

RDF-Schéma : définit les relations entre ces ressources. Le pouvoir sémantique de ces deux langages est limité car les axiomes ne peuvent pas être directement décrits. Le type des relations (symétrique, transitive, …) ne peut être spécifié.

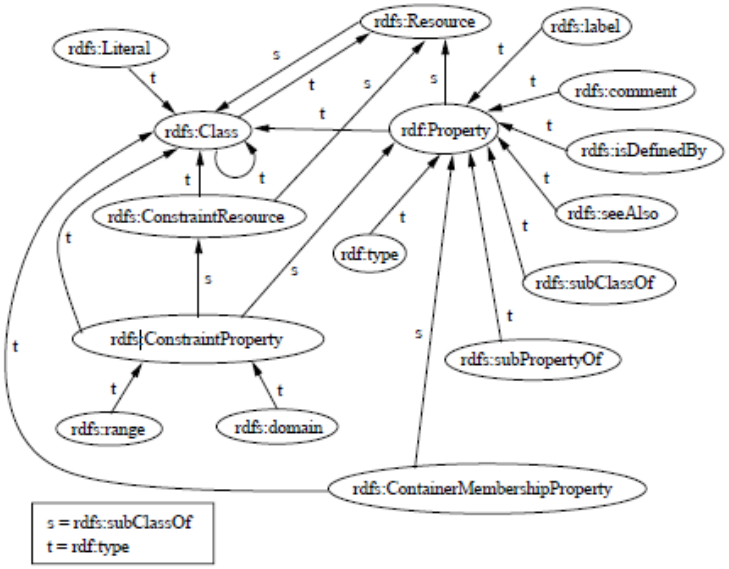


Figure 11: Le Schéma RDF

1.3.3.2.3. OIL (Ontology Inference Layer) :

Est à la fois un langage de représentation et d'échange pour les ontologies. Il combine les primitives des langages reposant sur les frames avec une sémantique formelle et des possibilités de raisonnement issues de la logique de description.
Pour être utilisé sur le Web, il repose sur les standards RDF(S) et XML. La description de l'ontologie est divisée en trois couches : la couche objet (instances concrètes), la couche de premier méta-niveau (définition de l'ontologie) et la couche de second meta-niveau (définition des caractéristiques de l'ontologie). OIL permet de définir des classes et des relations et un nombre limité d'axiomes. Les relations sont considérées comme des classes et peuvent être organisées hiérarchiquement.

1.3.3.2.4. XOL (XML based Ontology Exchange Language) [Karp 1999]:

A été créé pour échanger des ontologies se rapportant à la biologie moléculaire mais est applicable à d'autres domaines. Cependant, les relations entre concepts ne peuvent pas être spécifiées correctement.

1.3.3.2.5. SHOE (Simple HTML Ontology Extensions) [Luke 2000]:

Est une extension de HTML qui permet de rajouter de la sémantique dans ce type de documents. Il permet de définir des primitives pour spécifier et étendre les ontologies et annoter les documents Web. Chaque page déclare quelle ontologie elle utilise. L'inconvénient de ce langage est que les annotations des documents sont stockées à leur niveau et ne peuvent être centralisées.

1.3.3.2.6. DAML+OIL [Horrocks 2001] :

A été proposé par le W3C pour représenter des méta-données et des ontologies. DAML a été transformé en DAML+OIL en intégrant certaines propriétés de OIL. Il repose sur RDF et RDF schéma et fournit en plus des primitives plus riches issues de la logique de description. Les frames définis dans OIL ont été pour la plupart supprimées et remplacées par les assertions faites à l'aide d'un ensemble limité d'axiomes. Le résultat est que le langage est mieux adapté que RDF à l'utilisation et la maintenance d'ontologies mais présente des limites quant à la construction d'ontologie.

1.3.3.2.7. TOPIC MAPS :

Ont été créés par la Convention for Application of HyTime (CapH) [topicmaps] dont le but était de développer une application automatique d'indexation de livres. Les TM ont été acceptés par le groupe SGML d'ISO en 1996 et standardisés en janvier 2000. Topic Maps est un standard permettant de formaliser la sémantique sous la forme de méta-données. Il est défini à partir de thématiques (topics), d'occurrences de ces thématiques et d'associations non directionnelles entre les thématiques. Le rôle de chacun des membres de l'association a donc besoin d'être spécifié. Un mécanisme propre aux Topics Maps permet de préciser le contexte dans lequel l'association est valable ou intéressante. RDF et Topics Maps ont été créés dans le même but : décrire et organiser des méta-données. Ils sont compatibles et « traduisibles » de l'une à l'autre forme. Cependant, les TM présentent de nombreux intérêts : le mécanisme de spécification du contexte n'existe pas dans RDF, les TM permettent de connaître la relation et le rôle de l'objet dans la relation alors qu'avec RDF il est difficile de savoir si la source est un concept en relation avec l'objet ou contenant de l'information sur l'objet.

1.3.3.2.8. OWL Ontologie Web Language [McGuinness 2004] :

Est le standard actuellement proposé par le W3C pour représenter les ontologies. Il a été créé pour être utilisé par les applications cherchant à traiter le contenu de l'information et non plus uniquement à présenter l'information. OWL se veut plus représentatif du contenu du Web que XML, RDF et RDF-Schéma en apportant un nouveau vocabulaire avec une sémantique formelle.

OWL est une révision de DAML+OIL définie d'après l'expérience acquise lors de la création et l'utilisation de ce langage. OWL ajoute du vocabulaire pour décrire les propriétés et classes, comme par exemple la disjonction de classe, la cardinalité (exactement un), l'égalité, les types de propriétés plus riches, les caractéristiques de sous langages d'expressivité croissante : OWL lite, OWL DL, OWL Full. OWL Lite est fait pour des besoins préliminaires permettant de définir une hiérarchie et des contraintes simples. Il permet de définir facilement des thésaurus ou taxonomies. OWL DL et Full reposent sur OWL Lite auquel sont ajoutés des constructeurs supplémentaires. OWL DL supporte des besoins d'expressivité maximaux tout en garantissant une complétude de calculs et de décidabilité nécessaires aux systèmes de raisonnement. Il repose sur les éléments OWL auxquels il associe un grand nombre de restrictions.

OWL DL est conçu pour pouvoir supporter la logique de description. Cette logique appartient à un domaine de recherche qui a pour but d'aider au raisonnement sur une base de connaissances. OWL Full permet un maximum d'expressivité avec la liberté de syntaxe d'RDF. Il n'impose pas de séparation entre classe, propriété, individu et valeur des données. Il permet donc d'augmenter le sens du vocabulaire pré-défini (en OWL ou RDF). Il lève les contraintes imposées par OWL DL pour rendre certaines valeurs disponibles et utilisables dans des bases de données ou de connaissances, mais il ne supporte pas les raisonnements liés à la logique de description. L'utilisation du langage de représentation OWL dans le cadre d'un processus de RI permet d'une part de faire reposer le SRI sur un standard mais surtout d'utiliser un langage incrémental.

2. Conception et construction d'ontologies à partir de textes :

L'utilisation d'ontologies en informatique vise à intégrer une couche de connaissances aux systèmes afin de permettre des traitements élaborés de l'information qu'ils manipulent.

La conception d'ontologies est une tâche difficile qui nécessite la mise en place de procédés élaborés afin d'extraire la connaissance d'un domaine, manipulable par les systèmes informatiques et interprétable par les êtres humains. Deux types de conception existent : la

conception entièrement manuelle et la conception reposant sur des apprentissages. Plusieurs principes et méthodologies ont été définis pour faciliter la génération manuelle. Cependant, ce procédé de génération est très coûteux en temps et pose surtout des problèmes de maintenance et de mise à jour [Ding 2002]. La conception automatique d'ontologies commence à émerger comme un sous-domaine de l'ingénierie des connaissances. Face à la masse croissante de documents présents sur le Web et aux avancées technologiques dans le domaine de la recherche d'information, de l'apprentissage automatique et du traitement automatique des langues, de nouveaux travaux portent sur la recherche d'un procédé plus automatique de génération d'ontologies. Ce mécanisme mène généralement à la conception d'ontologies dites légères. Dans Maedche [Maedche 2001], différents types d'approches sont distingués en fonction du support sur lequel elles se basent : à partir de textes, de dictionnaires, de bases de connaissance, ….

2.1. Méthodologies de conception d'ontologies :

2.1.1. Conception manuelle d'ontologies :

Plusieurs principes ont été définis pour la construction d'ontologies [Gruber 1993a] [Uschold 1996]. Ces principes insistent sur la nécessaire clarté de la définition des éléments que l'ontologie doit contenir (rôle et portée de l'ontologie, définition des concepts, limitation des ambiguïtés) ainsi que sur la séparation des phases de conception et d'implantation dans un langage formel de l'ontologie. L'ensemble de ces principes reste cependant abstrait. Des méthodologies ont également été définies pour cadrer le développement d'ontologies de domaine.

Dans le projet TOVE [Gruninger 1995a], l'ontologie de domaine est construite à partir des scénarios d'entreprises pour lesquels elle sera utilisée. Cette méthodologie reste sommaire et aucune étape n'est décrite par rapport aux techniques qui peuvent y être employées. De plus, elle est spécialisée sur la spécification d'ontologies pour les entreprises. En revanche, les méthodologies METHONTOLOGY [Fernandez 1997] et KACTUS (modelling Knowledge About Complex Technical systems for multiple USe) sont conçues pour être appliquées dans des cadres plus généraux. Dans KACTUS, la méthodologie vise à réutiliser des ontologies existantes et propose des mécanismes permettant cette réutilisation. Ce principe est intéressant dans la mesure où il évite de construire une ontologie à partir de rien. Cette problématique existe d'ailleurs dans divers domaines comme dans la conception des systèmes d'information avec la définition de patrons [Rieu 1999].

2.1.1.1. Méthodologies:

METHONTOLOGY s'applique à clarifier les différentes étapes de la construction en respectant des activités de gestion de projets (planification, assurance qualité), de développement (spécification, conceptualisation, formalisation, implémentation, maintenance) et des activités de support (intégration, évaluation, documentation). Les différentes étapes proposées sont les suivantes :

- La première étape, étape de spécification, permet de produire un document de spécification de la future ontologie. Ce document décrit, entre autres, l'objet de l'ontologie, ses utilisateurs, ses utilisations, le degré de formalisation à employer ;
- La deuxième étape d'acquisition de connaissances mène à l'identification des termes de l'ontologie et leur définition. Des techniques d'acquisition de connaissances, comme les réunions de brainstorming, les interviews d'experts, les analyses de textes, sont listées. Toute technique d'acquisition est donc a priori utilisable ;

- L'étape suivante de conceptualisation vise à structurer la connaissance du domaine en un modèle conceptuel. La représentation est à ce stade informelle ;
- L'étape d'intégration qui suit, permet d'envisager quelles sont les ontologies existantes qui pourraient être intégrées dans la construction de l'ontologie. Des ontologies génériques ou de haut niveau peuvent être utilisées comme structuration des concepts de base. D'autres ontologies peuvent également être utilisées pour la définition de termes communs ;
- La phase suivante est celle de l'implantation. Elle consiste à représenter formellement l'ontologie à partir de langage comme LOOM, Ontolingua mais aussi Prolog ou C++ ;
- La phase d'évaluation intervient alors pour vérifier et valider l'ontologie en question ainsi que son environnement logiciel et sa documentation. Les problèmes de cohérence, d'incomplétude et de répétition sont alors vérifiés ;
- La dernière étape repose sur la documentation. Les auteurs insistent sur ce point en précisant que, généralement, les documentations sont incomplètes pour la compréhension globale de l'ontologie et sa réutilisation. Ils proposent alors de pallier ce manque en imposant la rédaction de documentations à la fin de chacune des phases de la construction de l'ontologie.

La particularité de la méthodologie METHONTOLOGY est de s'attacher fortement à la maintenance de l'ontologie de domaine et à son évaluation. De plus, les auteurs insistent sur le fait que les concepteurs doivent s'efforcer autant que possible d'utiliser des ontologies existantes.

Un autre type de méthodologie vise à aider à la construction d'ontologie formelle en proposant des règles pour évaluer la cohérence logique des liens de subsomption modélisés [Guarino 2002] [Kassel 2002]. La méthodologie **OntoClean** [Guarino 2002] repose sur la définition de caractéristiques des concepts pour structurer des ontologies en imposant certaines contraintes sur l'utilisation des liens de subsomption [Guarino 2000]. Ces caractéristiques aussi appelées méta-propriétés sont :

- L'*identité* : un concept porte une propriété d'identité si cette propriété permet de conclure quant à l'identité de deux instances de ce concept. Cette propriété peut porter sur des attributs du concept ou sur d'autres concepts. Par exemple, le concept « étudiant » porte une propriété d'identité liée au « numéro » de l'étudiant, deux étudiants étant identiques s'ils ont le même numéro ;
- La *rigidité* : une propriété d'un concept est rigide si elle est essentielle pour chacune des instances du concept, c'est-à-dire que chacune des instances détient cette propriété pour exister. Par exemple, la propriété « être un humain » est rigide, mais « être un étudiant » est non rigide ;
- L'*unité*: un concept composé de plusieurs concepts est un concept unité si chacune de ses instances forme «un tout». Par exemple, le concept «eau» n'est pas unité car une de ses instances est une quantité d'eau qui ne peut pas être reconnue en tant qu'entité isolée. Le concept «océan» est unité car ses instances «océan atlantique» sont des entités à part entière;
- La *dépendance* : un concept C1 est dépendant d'un concept C2 si, pour toute instance de C1, il existe une instance de C2 qui ne soit ni partie ni constituant de l'instance de C1. Par exemple, « parent » est un concept dépendant de « enfant » (et inversement), car l'existence d'un parent suppose celle d'un enfant. Mais « couteau » et « manche » ne sont pas dépendants, car le manche fait partie du couteau.

Ces quatre méta-propriétés font peser des contraintes sur les liens de subsomption entre concepts. Par exemple, un concept portant une propriété d'identité ne peut subsumer un concept qui n'en porte pas. La méthodologie OntoClean propose donc de typer les concepts d'une

ontologie à l'aide de ces caractéristiques, puis de tester la cohérence de la hiérarchie des concepts en vérifiant que les contraintes induites par ces caractéristiques ne sont pas violées.

La méthodologie **OntoSpec** [Kassel 2002] reprend ces méta-propriétés et incite le concepteur à utiliser certaines propriétés en particulier lors de l'élaboration de l'ontologie. Elle prend en amont un ensemble d'entités conceptuelles exprimées par des termes ainsi qu'un ensemble de définitions en langue naturelle. Par un processus de transformation reposant sur la définition des différentes méta-propriétés, OntoSpec permet d'élaborer des ontologies semi-informelles pouvant ensuite être représentées dans le langage formel choisi par le concepteur.

2.1.1.2. Différents outils de conception manuelle :

Différents outils ont été proposés pour aider à la conception manuelle d'ontologies. Ces outils permettent d'éditer une ontologie, d'ajouter des concepts et des relations, etc. Ils intègrent différents langages de formalisation (RDF, OWL). Certains doivent être installés en local alors que d'autres sont distribués sur le Web. Ces outils sont décrits plus spécifiquement :

2.1.1.2.1. OntoEdit (Ontology Editor) [Sure 2002] :

Est un environnement de construction d'ontologies. Il a été développé par la compagnie Ontoprise. Il permet l'édition des hiérarchies de concepts et de relations dans le cadre de la logique des frames, ainsi que l'expression d'axiomes algébriques. Des outils graphiques dédiés à la visualisation d'ontologies sont inclus dans l'environnement. OntoEdit intègre, dans sa version commerciale, un serveur destiné à l'édition d'une ontologie par plusieurs utilisateurs ainsi qu'un plug-in permettant le test de la cohérence d'une ontologie. OntoEdit gère de nombreux formats de représentation de connaissances dont OWL, RDFS et FLogic.

2.1.1.2.2. Protégé :

Est une interface modulaire, développée au Stanford Medical Informatics de l'Université de Stanford, permettant l'édition, la visualisation, le contrôle (vérification des contraintes) d'ontologies [Noy 2000]. Le modèle de connaissances de Protégé est issu du modèle des frames et contient des classes (concepts), des slots (propriétés) et des facets (valeurs des propriétés et contraintes), ainsi que des instances des classes et des propriétés. Protégé autorise la définition de métaclasses, dont les instances sont des classes, ce qui permet de créer son propre modèle de connaissances avant de bâtir une ontologie. De nombreux plug-in sont disponibles ou peuvent être créés par l'utilisateur. Parmi ceux-ci, citons le plug-in permettant d'utiliser le langage OWL et les plug-ins de visualisation.

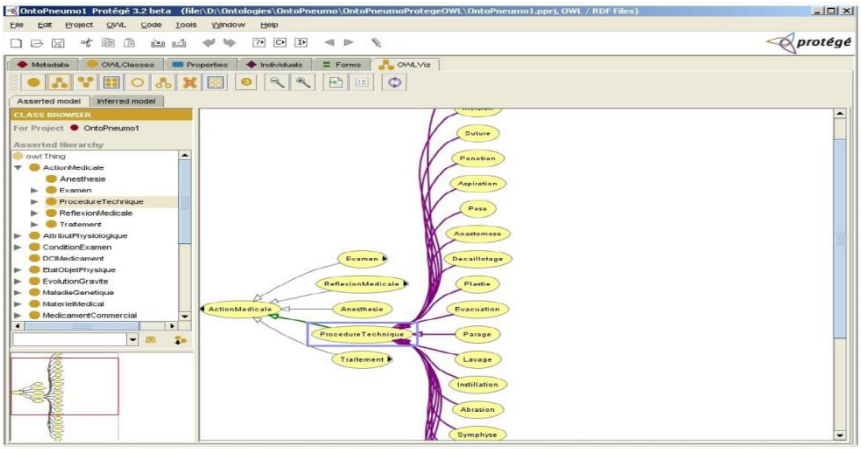


Figure 12: Vue extraite de protégé

2.1.1.2.3. L'ODE (Ontology Design Environment) :

Développé à l'Université de Polytechnique de Madrid permet de mettre en place la méthodologie METHONTOLOGY. Son successeur pour le Web WebODE a pour ambition de couvrir l'ingénierie ontologique à travers les différentes activités liées au cycle de vie d'une ontologie : acquisition de connaissances à partir du Web, édition d'ontologies, test de la consistance d'une ontologie, alignement et fusion d'ontologies, import et export dans des formats variés. Le modèle de représentation de connaissances utilisé associe un modèle de type frame (concepts et attributs) avec des relations entre concepts.

Des propriétés conceptuelles (en particulier algébriques) peuvent être associées aux relations. Les axiomes d'une ontologie sont des tautologies du domaine, mais on peut aussi inclure dans l'ontologie des règles susceptibles d'être utilisées pour raisonner dans un moteur d'inférence de type Prolog.

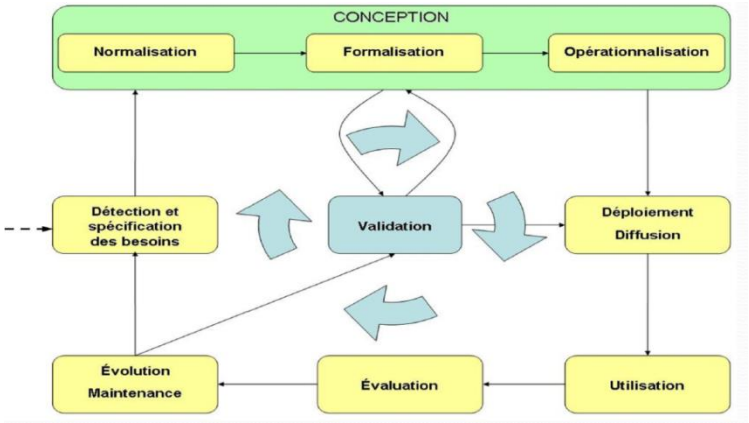


Figure 13: Cycle de vie d'une ontologie [Banyex 2007]

2.1.1.2.4. WebOnto :

Développé au Knowledge Media Institute de l'Open University, offre une interface graphique d'édition collaborative, de tests et de parcours d'ontologies sur le Web [Domingue 1998]. Le modèle de connaissance utilisé est celui du langage OCML, un langage à base de Frames.

2.1.1.2.5. OilEd (Oil Editor) :

Est un éditeur d'ontologies utilisant le formalisme DAML+OIL et les Logiques de Description [Bechhofer 2001]. Il est essentiellement dédié à la construction d'ontologies dont on peut ensuite tester la cohérence à l'aide de FACT, un moteur d'inférences bâti sur OIL. Il permet l'export d'ontologies sous les formats RDF, DAML+OIL, OWL et d'autres langages moins consensuels comme SHIQ.

2.1.1.2.6. ONTOLINGUA :

Développé au Knowledge Systems Laboratory de l'Université de Stanford, est un serveur d'édition d'ontologies permettant la construction collaborative d'ontologies [Farquhar 1997]. Une ontologie y est directement exprimée dans un formalisme également nommé ONTOLINGUA.

2.1.2. Conception d'ontologies en utilisant TERMINAE :

La méthodologie TERMINAE [Aussenac-Gilles 2000a] [Aussenac-Gilles 2000b] est une méthode qui repose sur l'analyse de corpus linguistique. Elle tente de répondre aux manques des autres méthodologies en proposant une approche pour sélectionner les concepts, leurs propriétés, les relations et leur regroupement. Elle repose pour cela sur l'utilisation d'outils de traitement automatique des langues analysant les termes de textes et les relations lexicales. Les

termes sont regroupés suivant leur contexte et facilitent la création de concepts et de relations sémantiques. Les concepts et relations sont ensuite formalisés dans un modèle.

Cette méthodologie est composée de plusieurs étapes :

- La première consiste en la description des besoins (utilisation de l'ontologie, connaissance à représenter…) ;

- L'étape suivante conduit à construire un corpus sur lequel les outils de traitement automatique de langues seront réalisés. Cette étape est fondamentale car de la qualité du corpus dépendra la qualité des traitements. Le corpus doit couvrir entièrement le domaine traité par l'application. Cette phase nécessite l'intervention d'un expert pour récolter les différents types de documents significatifs ;

- La troisième étape correspond à l'étude linguistique. Des outils sont utilisés sur le corpus afin d'extraire les termes et leurs relations lexicales et syntaxiques. Le choix des outils est laissé à l'utilisateur. Une application de la méthode est proposée par les auteurs à partir des outils LEXTER [Bourigault 1996] et Caméléon [Séguéla 1999]. Le premier extrait les termes candidats à partir de leurs dépendances syntaxiques. Le second extrait des relations entre termes à partir de patrons linguistiques. Les outils utilisés nécessitent l'intervention d'experts du domaine afin de sélectionner et de valider les candidats. A la fin de cette étape, un ensemble de termes, de relations lexicales entre ces termes et de regroupements est obtenu ;

- La phase suivante, appelée phase de normalisation, vise à conceptualiser les résultats de l'étape précédente. Les termes à conserver sont sélectionnés en fonction de leur contexte et définis à partir d'une définition en langage naturel. Les concepts sont ensuite identifiés ainsi que les relations sémantiques entre eux. Ils sont représentés sous forme d'un réseau sémantique ;

- La dernière étape est celle de la formalisation. Le réseau sémantique précédemment obtenu est traduit et enrichi dans un langage formel. Des méthodes de formalisation telle que celle définie dans Kassel [Kassel 2002] peuvent être utilisées dans cette étape. Cette méthodologie a l'avantage de répondre à certaines questions et d'axer le choix des concepts et des relations de l'ontologie sur l'extraction de termes d'un corpus de référence. Cependant, elle ne spécifie pas comment les concepts doivent être sélectionnés ni quelles sont les propriétés adéquates. Cette méthodologie est générale et demande l'intervention d'experts.

Un outil est associé à cette méthodologie. Il a été développé au LIPN et permet, à travers les outils LEXTER [Bourigault 1996], Syntex [Bourigault 2000] et Caméléon [Séguéla 1999], d'extraire d'un corpus textuel les candidats termes d'un domaine. Il offre un support méthodologique qui permet de faire évoluer progressivement une ontologie en conservant des liens entre les textes et les niveaux linguistiques et conceptuels. Le modèle de représentation de TERMINAE est celui des Logiques de Description mais une traduction des ontologies dans le langage OWL est possible.

2.2. Méthodes de construction d'ontologies de domaine à partir de textes :

Une méthodologie définit le cadre général de la conception d'ontologies mais nécessite l'ajout de méthodes pour la mettre en œuvre concrètement.

La construction d'ontologies à partir de textes vise à cette mise en œuvre à partir d'éléments qui peuvent être extraits de ces textes. Elle aboutit généralement à la conception d'ontologies légères de domaine.

La première phase décrit la constitution d'un corpus de référence. La deuxième phase concerne l'extraction des termes et des concepts du domaine. Les deux phases suivantes extraient les relations taxonomiques et associatives de la structure de l'ontologie. Enfin, des techniques permettent de mettre à jour une ontologie.

2.3. Constitution du corpus :

Afin de mettre en place la construction d'ontologies à partir de textes, il est tout d'abord nécessaire de constituer l'ensemble des documents sur lequel reposera cette élaboration [Condamines 2005]. Plusieurs cas de figures peuvent amener à élaborer ce corpus [Lame 2002]. S'il existe des documents dans lesquels la connaissance peut être capturée, les documents préexistants sont rassemblés.

L'enjeu est alors de collecter des documents existants afin de couvrir le domaine d'intérêt. Une solution est d'interroger le Web à partir de requêtes décrivant le domaine qui devra être traité dans l'ontologie. Chez Agirre [Agirre 2000] par exemple, l'objectif est de mettre à jour WordNet. Pour cela, un ensemble de documents de référence est extrait du Web pour chacun des concepts à mettre à jour à partir de requêtes formées des termes qui décrivent le concept et ses hyperonymes dans WordNet. Une alternative est de choisir un corpus existant et de le valider pour servir de corpus de référence. Dans le cas des travaux portant sur la génération d'ontologies pour la RI, le corpus est généralement composé de l'ensemble des documents à indexer comme par exemple chez Ok Koo [Koo 2003].

Si un tel ensemble de documents n'existe pas, des documents doivent être créés spécialement à cet effet. Ce cas de figure se présente quand l'ontologie doit capturer de la connaissance tacite sur un domaine comme, par exemple, lorsque l'ontologie traite de la mémoire d'une entreprise. Le savoir-faire des experts du domaine n'est pas explicitement présenté dans des documents. La connaissance des experts est alors capturée à partir de documents textuels relatant des interviews. La construction de ce type de corpus revient à faire passer les connaissances du tacite à l'explicite.

2.4 Extraction de termes :

Les termes candidats pour représenter les concepts d'une ontologie peuvent être extraits selon deux approches : syntaxique ou statistique. L'approche syntaxique analyse le rôle grammatical des mots dans ces textes, alors que l'approche statistique repose sur la fréquence d'apparition des mots dans les textes.

2.4.1. Techniques syntaxiques d'extraction de termes :

Les techniques syntaxiques extraient des termes à partir des relations grammaticales entre les mots dans les phrases des documents. Ces termes peuvent être composés d'un seul mot ou d'une suite de mots. Les expressions extraites syntaxiquement, aussi appelées syntagmes, exploitent le rôle des mots dans les documents dont elles sont issues. Elles déterminent des composants de la phrase très précis qui doivent être détectés en fonction de la grammaire de la langue utilisée. Elles contiennent un ou plus d'un mot et sont plus petites qu'une phrase [Caropreso 2000]. On peut extraire des expressions nominales telles que «هيئة عمومية», des expressions verbales : « تتولى وصاية », « يسير من طرف », des expressions adjectivales : «المدى الطويل ». Les expressions extraites syntaxiquement prennent en compte des relations linguistiques entre les mots, elles sont donc plus significatives sémantiquement que la juxtaposition des mots formant les expressions statistiques.

Différents analyseurs syntaxiques existent. Riloff [Riloff 1996] s'appuie sur l'utilisation de patrons syntaxiques définis manuellement. Termino [David 1990], quant à lui, repose sur le

découpage des textes en unités lexicales pour l'identification de syntagmes nominaux. Syntex [Bourigault 2000] s'appuie sur un apprentissage des relations de dépendance entre mots pour extraire les syntagmes de différents types (nominaux, verbaux,…) et les organiser suivant un réseau de dépendance. La particularité de Syntex est de s'appuyer sur un apprentissage endogène du corpus qui lui permet d'être plus performant qu'un analyseur reposant uniquement sur des règles définies manuellement. Il peut s'adapter à des collections spécialisées dans différents domaines tels que le droit et la chirurgie [Bourigault 2002a] [Le Moigno 2002].

2.4.2. Techniques statistiques d'extraction de termes :

2.4.2.1. Extraction des termes :

L'extraction de termes se base sur l'utilisation d'un anti-dictionnaire pour supprimer les mots vides puis sur la radicalisation des termes restants pour supprimer les variantes lexicales.

L'utilisation d'un anti-dictionnaire vise à éliminer les mots ayant un contenu informationnel vide. Ces mots apparaissent dans la plupart des documents et ne sont pas discriminants. Ces mots qui peuvent être des articles, prépositions, conjonctions voire même des verbes sont appelés mots vides et sont regroupés dans un anti-dictionnaire. L'utilisation d'un anti-dictionnaire permet de réduire considérablement le nombre de termes extraits. Par exemple, si les termes de l'anti-dictionnaire apparaissent dans plus de 80 % des documents, le nombre de termes diminue de 40% [Baeza-Yates 1999]. Afin de supprimer les différentes variantes lexicales d'un terme et de ne considérer qu'une forme unique, la racine du terme est extraite. Ce procédé est appelé radicalisation [Frakes 1992]. Il existe différentes méthodes de radicalisation selon la langue et le domaine:

L'utilisation de tables de correspondance entre le terme et le radical et le schème en langue arabe, Par exemple, le terme « مرسوم » son radical « رسم » et son schème c'est « مفعول ».

Ainsi, l'extraction de termes individuels (composés d'un seul mot) est la plus utilisée.

Cependant, l'extraction d'expressions permet d'obtenir des termes ayant une meilleure sémantique. Les expressions extraites statistiquement représentent une séquence de mots juxtaposés. Seules les expressions présentes fréquemment dans la collection sont extraites [Mitra 1997]. Une fois sélectionnées, elles peuvent être également radicalisées. Chacun des termes de l'expression est mis sous sa forme radicalisée et les termes de l'expression sont, par exemple, ordonnés alphabétiquement [Mitra 1997]. Ces techniques, bien que définies dans le cadre de l'indexation de documents pour la RI, peuvent être appliquées dans le cadre de la construction d'ontologies.

2.4.2.2 Sélection des termes :

Les termes sont ensuite sélectionnés à partir de leurs occurrences dans les documents. Des mesures issues de la RI peuvent être utilisées. Dans ce cas, la fréquence du terme dans les détails permet de prendre en compte les termes qui apparaissent souvent dans le corpus mais principalement dans quelques documents. L'entropie, quant à elle, analyse la répartition des termes dans les documents, ce qui permet, en fonction de la formule, de sélectionner les termes soit rares soit redondants [Brini 2005].

Plusieurs conclusions contradictoires ont été tirées sur l'efficacité comparative de ces différentes mesures.

2.5. Extraction de liens de subsomption :

Les liens de subsomptions dans une ontologie permettent d'organiser les concepts hiérarchiquement.

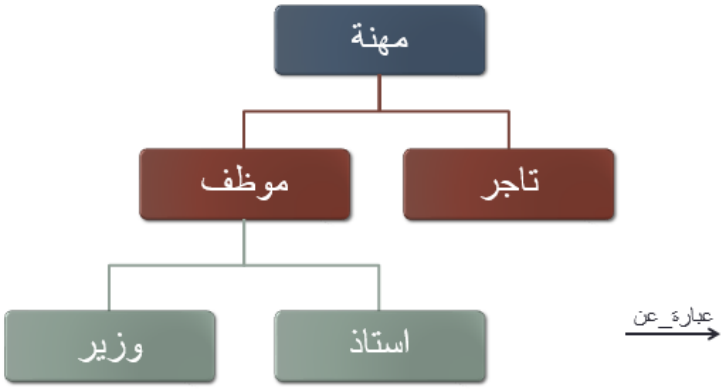


Figure 14: Les Relation de Subsomption

Pour extraire ce type de liens, différentes méthodes issues de la RI et de l'IC existent. En réalité, certaines d'entre elles permettent d'extraire des liens non pas entre concepts mais entre termes et les liens obtenus ne respectent pas la définition stricte de la subsomption. Elles s'appuient sur des approches statistiques ou linguistiques. Les approches statistiques regroupent et structurent les termes par rapport à leurs occurrences dans les différents documents. Les approches linguistiques reposent sur une analyse syntaxique du contenu des documents et regroupent les termes ou des concepts par rapport à leur contexte d'apparition.

2.5.1. Approches statistiques :

Nous décrivons plusieurs méthodes statistiques permettant d'extraire des relations taxonomiques entre termes. Ces méthodes se basent sur l'analyse des cooccurrences entre termes dans les documents. La cooccurrence correspond à l'apparition simultanée de deux termes dans un texte (document ou fenêtre de N mots). Les méthodes présentées dans cette section représentent l'ensemble des cooccurrences dans une matrice. Cette matrice est ensuite utilisée pour regrouper hiérarchiquement les termes, soit par application des méthodes de classification automatique, soit en s'appuyant sur des mesures de probabilité.

2.5.1.1. Méthodes de regroupement hiérarchique de termes :

Manning [Manning 1999] indique que le regroupement non supervisé en classes permet la détection de relations de généralisation. Ainsi, Maedche [Maedche 2000] propose d'appliquer les méthodes hiérarchiques ascendantes ou descendantes à la matrice de cooccurrence des termes extraits des documents. Dans le cas d'une classification hiérarchique ascendante, au départ, chaque classe est composée d'un terme. Les regroupements se font en associant deux classes qui ont les profils les plus proches. Le profil d'une classe correspond à ses occurrences avec l'ensemble des termes. Les regroupements successifs aboutissent à la formation d'une seule classe. La classification hiérarchique descendante procède, au contraire, par divisions successives. Elle part d'une classe formée de tous les termes et la divise en deux classes. Elle détermine la classe la moins cohérente, au sens de la mesure de similarité entre classes et la divise elle aussi en deux. Ceci est répété jusqu'à l'obtention des partitions d'un seul élément. Dans les deux cas, de nombreuses mesures de similarités ont été définies ou utilisées dans la littérature [Jain 1999] [Murtagh 1998].

2.5.1.2. Méthode reposant sur la probabilité de cooccurrence :

Pour Sanderson [Sanderson 1999], l'association des termes repose sur une relation parent-enfant où le terme parent est plus général que le terme enfant. Cette relation entre termes est

extraite d'après la cooccurrence asymétrique de termes. La relation est caractérisée par les deux règles suivantes :

$$P(x/y) \geq 0.8 \text{ (seuil empirique)}$$

$$Et \qquad P(y/x) < P(x/y)$$

où p(x/y) est la probabilité d'obtenir le terme x dans un document sachant que le document contient le terme y, inversement pour p(y/x).

La première règle assure que les deux termes apparaissent suffisamment dans les mêmes documents (en l'occurrence dans 80% des cas). D'après la deuxième règle, x subsume y si les documents dans lesquels il apparaît sont un sous-ensemble des documents où apparaît x. Le terme apparaissant le plus souvent est choisi comme parent. Les relations extraites à partir des deux règles citées sont ensuite nettoyées en supprimant les termes qui co-occurrent dans moins de deux documents et en supprimant les relations redondantes par rapport à la propriété transitive de la relation. Si les relations a subsume b, a subsume c et b subsume c sont extraites, la relation a subsume c peut être supprimée parce qu'elle est déductible des deux autres.

2.5.2. Approches linguistiques :

Les approches linguistiques s'appuient sur une analyse syntaxique des documents pour extraire des relations taxonomiques entre termes issus des documents. L'analyse syntaxique permet soit de définir des patrons d'extraction, soit de procéder au regroupement conceptuel.

2.5.2.1. Approches reposant sur la définition de patrons d'extraction :

L'idée d'utiliser des patrons lexico-syntaxiques sous la forme d'expressions régulières afin d'extraire des relations sémantiques a été introduite par Hearst [Hearst 1992]. Dans ces travaux, les patrons syntaxiques définissant la relation taxonomique sont construits manuellement. Les relations sont ensuite extraites automatiquement du corpus. Un exemple de patron est le suivant «*SN {,SN}*{,}ou autres SN*» où SN dénote la présence d'un syntagme nominal. En appliquant ce patron sur l'extrait suivant « الولايات ، البلديات وغيرها من المباني العمومية» «préfectures, mairies ou autres bâtiments publics», les relations taxonomiques sont déduites entre les couples المباني البلدية/المباني العمومية , الولاية/العمومية .

Le Système Prométhée développé par Morin propose d'étendre ces travaux en se basant sur un apprentissage permettant d'extraire automatiquement les patrons lexico-syntaxiques correspondant à une relation sémantique donnée [Morin 1999]. L'apprentissage consiste tout d'abord à donner au système un ensemble de couples de termes vérifiant cette relation. Un corpus est ensuite analysé et l'ensemble des patrons que suivent ces couples est traité par le système. Les patrons sont triés et sélectionnés à partir d'une mesure de similarité permettant de choisir les patrons les plus représentatifs de l'ensemble. Les patrons retenus sont ensuite validés par un expert. Le système permet également d'étendre les relations déduites entre termes simples (un seul mot) à de nouvelles relations entre termes composés de plusieurs mots dont deux des mots sont liés. Si, par exemple, les termes *مادة* et *قانون* sont liés sémantiquement, le même lien sémantique pourra être déduit entre les termes composés «*أحكام المادة*» et «أحكام القانون». Cette étape, appelée normalisation sémantique, repose sur l'analyse syntaxique des syntagmes et restreint les cas où la formation des termes composés est possible. Une relation sémantique entre les mots t_1 et t_2 n'implique pas toujours une relation entre les syntagmes $t_1t_{1'}$ et $t_2t_{2'}$ où les t_i sont des mots. Une relation sémantique est déduite entre les syntagmes $t_1t_{1'}$ et $t_2t_{2'}$ si les trois règles suivantes sont respectées :

1. Une relation sémantique lie (t_1 et t_2) ou ($t_{1,}$ et $t_{2,}$) et les mots non sémantiquement liés sont identiques ou sont morphologiquement liés ;

2. t_1 et $t_{1,}$ sont deux mots têtes et t_2 et $t_{2,}$ sont deux arguments avec des rôles thématiques semblables ;

3. t_1 $t_{1,}$ et $t_2t_{2,}$ partagent la même relation sémantique entre leurs composants.

Les auteurs se basent sur une hypothèse qui dit que si 1 et 2 sont vrais alors 3 l'est aussi. Afin de mettre en place les règles 1 et 2, une analyse morphologique des termes est réalisée à partir de l'outil d'acquisition FASTER [Jacquemin 1999]. Les rôles thématiques sont extraits à partir de la nature des termes et des prépositions les encadrant. Un ensemble de patrons, réalisés manuellement à partir des résultats du logiciel d'acquisition ACABIT, définit les rôles thématiques semblables. Voici un exemple d'application. Une relation peut être déduite entre *الادارة العامة* et *الهيئة العمومية* car tout d'abord les termes *الهيئة* et *الادارة* sont donnés comme étant liés et les termes *العامة* et *العمومية* sont morphologiquement similaires. De plus, un patron syntaxique permet de déterminer que les deux syntagmes sont des arguments du même rôle thématique. Le système a été testé pour la détection de relations taxonomiques à partir d'un ensemble de couples d'apprentissage extraits de documents ou bien d'un thésaurus. Les résultats par les deux apprentissages sont similaires et montrent qu'environ 60% des relations taxonomiques détectées à partir des patrons appris sont correctes et qu'à peu près 80% des relations étendues par la normalisation sont justes. Le système pourrait permettre d'extraire d'autres types de relations.

Les patrons ont aussi été utilisés par Maedche [Maedche 2000] pour créer une ontologie à partir de dictionnaires dans le domaine de l'assurance et des télécommunications. Les auteurs présentent leur méthode comme étant plus originale que la précédente car les patrons sont générés au niveau des concepts et non pas au niveau des termes. Pour cela, ils considèrent les concepts comme étant le premier mot de la définition des entrées d'un dictionnaire. Les patrons sont ensuite définis à partir de ce terme.

2.5.2.2. Regroupements conceptuels :

Faure et Nedellec [Faure 1998] ont présenté le système ASIUM. Dans ce système, les relations taxonomiques sont acquises par un traitement syntaxique des documents. Des classes sont formées à partir des termes qui apparaissent après le même verbe et la même préposition en appliquant un algorithme de regroupement conceptuel. Les classes sont successivement agrégées pour former de nouveaux concepts et une hiérarchie constituant l'ontologie. Les classes formées doivent être labellisées par un expert pour identifier le concept qu'elles représentent. Les classes sont composées de groupements de mots suivant le patron: <verbe> <rôle syntaxique |préposition : nom*>*, comme par exemple « <ينشر > <sujet : المرسوم > <في : الجريدة >. Les couples <rôle syntaxique : nom> ou <préposition : nom> sont appelés « mots têtes ». La mesure de similarité qui permet d'évaluer la distance entre les classes, et donc de les regrouper, dépend de la proportion de mots têtes communs dans les différentes classes en prenant en compte leur fréquence d'apparition dans les documents.

Dans OntoLearn [Velardi 2002], des sous-graphes conceptuels sont créés pour faciliter la génération manuelle de taxonomies. Les sous-graphes sont formés en regroupant les syntagmes ayant la même tête et en développant la hiérarchie au fur et à mesure que des termes sont ajoutés à la queue des syntagmes.

2.6. Détection de relations non taxonomiques :

Une autre phase dans l'élaboration d'ontologies consiste à extraire des relations non taxonomiques entre concepts. La difficulté est qu'elle doit non seulement extraire des relations entre concepts mais également permettre de labelliser les relations afin de désigner la relation sémantique. Les approches présentées ici sont des contributions qui rentrent dans ce cadre. Il s'agit d'approches statistiques reposant sur la cooccurrence et d'approches syntaxiques.

2.6.1. Cooccurrences des verbes :

S. Koo [Koo 2003] propose de construire une ontologie pour aider à la Recherche d'Information. Cette ontologie porte sur le domaine de l'économie ; elle est construite semi automatiquement à partir de l'analyse des documents du Wall Street Journal des collections TREC. Le principe consiste à extraire les noms et noms propres apparaissant fréquemment. Ces termes sont appelés les termes centraux. Afin de déceler des relations entre termes, les termes apparaissant dans une fenêtre de quatre mots autour des termes centraux sont extraits. Les termes co-occurrents fréquemment sont proposés pour être reliés dans l'ontologie, les verbes apparaissant dans le contexte sont proposés pour être labels de la relation.

2.6.2. Analyse syntaxique :

Plusieurs approches reposant sur l'analyse syntaxique des documents permettent d'extraire des relations non taxonomiques.

L'approche présentée dans Velardi [Velardi 2002] consiste à analyser syntaxiquement les documents et à produire des triplets (Acteur Procédé Objet) à partir du patron syntaxique (Sujet Verbe Objet). Les couples vérifiant ce patron sont sélectionnés si au moins un élément du triplet appartient déjà à l'ontologie. Les couples ayant une faible plausibilité sont supprimés. La plausibilité prend en considération le pouvoir consensuel des termes. Aucun résultat n'est donné sur les performances de la méthode.

Bourigault [Bourigault 2002b] propose également des mécanismes pour extraire des relations (non typées). Le principe utilisé s'appuie sur la proximité entre les différents syntagmes extraits par Syntex et sur l'analyse distributionnelle [Harris 1968]. Cette analyse consiste à regrouper des syntagmes en fonction du contexte (mots par lesquels ils sont régis et mots qu'ils régissent) qu'ils partagent. Les syntagmes déduits de l'analyse syntaxique sont rapprochés s'ils sont formés autour de la même relation et de la même tête. La proximité est calculée ensuite grâce à différentes formules qui prennent en compte la "productivité" d'un contexte (c'est-à-dire le nombre de termes qui apparaissent dans ce contexte), et la "productivité" d'un terme (nombre de contextes différents dans lesquels apparait le terme). L'analyse distributionnelle permet donc de rapprocher des termes deux à deux en fonction des contextes qu'ils partagent mais aussi de rapprocher les contextes en fonction des termes qu'ils ont en commun. Le module UPERY reposant sur cette analyse a été intégré à Syntex. Ce module permet donc de rapprocher des termes par rapport à leur contexte et de détecter une relation sémantique entre eux. Cependant, la nature de cette relation n'est pas spécifiée.

2.6.3. Approche reposant sur les règles d'association :

Maedche [Maedche 2000] et Sugiura [Sugiura 2004] proposent d'extraire des relations non taxonomiques entre concepts à partir de règles d'association. L'algorithme d'exploration de données présenté dans Srikant [Srikant 1995] permettant de trouver des associations généralisées entre éléments tels que الأسواق et شراء أشياء est utilisé. Le principe de l'algorithme consiste à déterminer un ensemble de transactions $T := \{t_i / i=1..n\}$ où chaque transaction ti est

constituée d'un ensemble d'éléments $X_i := \{a_{ij}/j=1..m, a_{ij} \in C\}$ et chaque element a_{ij} référence un concept appartenant à l'ensemble C.

Dans le cas de la conception d'ontologies, les éléments sont les termes. Dans Maedche [Maedche 2000], les transactions retenues sont l'apparition de groupes de noms ou d'entités nommées dans un même syntagme (contenant une préposition ou bien apposant les mots). L'approche proposée a été évaluée sur un corpus des télécommunications et permet d'extraire des relations intéressantes. En revanche, dans, les transactions sont le fait de retrouver deux mêmes termes dans une phrase. Testée sur un corpus du droit, cette méthode permet d'obtenir des relations suivant des taux de rappel et de précision relativement bas (respectivement 27% et 24%). L'algorithme [Srikant 1995] génère des règles d'association $X_k \Rightarrow Y_k$ telles que le support et la confiance de ces règles excèdent un seuil fixé. Le support de l'association $X_k \Rightarrow Y_k$ est défini par le pourcentage de transactions qui contient $X_k \cup Y_k$. La confiance correspond au pourcentage de transactions où Y_k apparaît dans la transaction si on a X_k. Cet algorithme est initialement étendu pour prendre en compte les ancêtres dans une taxonomie des éléments a_{ij}. Le support et la confiance sont alors calculés pour les transactions où Y_k ne contient aucun ancêtre de X_k. Puis les règles d'association obtenues sont filtrées pour ne contenir aucune règle d'association faisant intervenir $X_k \Rightarrow Y_k$ $\underline{X_k} \Rightarrow \underline{Y_k}$ où $\underline{X_k}$ est un ancêtre de X_k et $\underline{Y_k}$ est un ancêtre de Y_k. Cependant cette méthode ne permet pas de fournir des labels aux relations extraites.

Les travaux présentés dans Maedche [Maedche 2000] été étendus dans Kavalec [Kavalec 2004] afin d'extraire les labels. Les labels sont extraits parmi les verbes co-occurrents fréquemment autour des deux concepts c_1 et c_2. Des expérimentations ont été menées sur le corpus ***SemCor***. Ce corpus présente l'avantage que chacun des termes est désambiguïsé à partir d'un sens précis qui peut être retrouvé dans WordNet. L'utilisation de ce corpus a permis de regrouper les termes co-occurrents à partir des concepts les généralisant dans WordNet. Les résultats obtenus montrent que la moitié des labels proposés permettent de désigner correctement les relations sémantiques entre concepts.

3. Techniques de mise à jour d'ontologies :

Différentes techniques de mise à jour de l'ontologie ont été proposées. Ces techniques visent à extraire de nouveaux termes et à les intégrer dans l'ontologie. Elles se basent sur la détection d'indices lexicaux et statistiques liant les nouveaux termes détectés dans le corpus et les labels de concepts présents dans le corpus. Plusieurs approches ont été définies pour extraire ces indices.

La méthode proposée dans Faatz [Faatz 2002] consiste à définir une mesure calculant la distance dans les textes entre les labels de concepts de l'ontologie liés par une propriété et d'utiliser cette mesure pour détecter de nouveaux labels. Les labels extraits des documents sont alors supposés être liés par la propriété aux concepts existants. La mesure proposée est inspirée de la divergence de Kullback qui calcule la dissimilarité entre deux mots. Elle est étendue pour prendre en compte des indices dans les documents témoins de la relation sémantique considérée. La mesure est optimisée à partir des labels des concepts liés par cette relation dans l'ontologie. Deux corpus de référence sont utilisés pour extraire les nouveaux termes. Un premier corpus est constitué sur le Web par la spécification de requêtes à partir des labels des concepts dans l'ontologie. Le deuxième corpus est composé de documents du domaine considéré. Le cas d'application présenté dans l'article consiste à définir une mesure permettant d'extraire la relation sémantique «est un». Cette propriété est prise en compte dans la mesure par la

fréquence de cooccurrence des labels de deux concepts dans une fenêtre de cinq mots autour des termes (seuls les termes co-occurrent au moins deux fois sont considérés). La méthode est testée dans le domaine de la médecine à partir d'une ontologie spécialisée. Sept concepts de l'ontologie sont considérés pour optimiser la mesure. Les résultats obtenus sur le corpus constitué de documents issus du Web permettent d'obtenir un nombre plus restreint de nouveaux labels que ceux provenant du corpus spécialisé. Cependant, l'analyse de ces différents labels montre que ceux proposés à partir du premier corpus sont plus pertinents. Bien qu'elle n'ait été testée que pour la relation «est un», l'avantage de cette approche est de permettre l'apprentissage de la distance représentée par n'importe quelle relation de l'ontologie. Cependant, sa mise en pratique pour tous types de relations est difficile car elle implique de maîtriser les indices du corpus qui permettent de déceler la relation.

Dans Maedche [Maedche 2002], une autre méthode est définie. Cette méthode vise à détecter de nouveaux termes mais propose uniquement de les intégrer à l'ontologie à partir de relations taxonomiques. Une matrice de cooccurrence est constituée à partir de termes désignant les concepts de l'ontologie et les termes co-occurrents autour d'eux dans une fenêtre de trois mots dans une même phrase. Ces termes sont ensuite hiérarchisés à partir d'une méthode hiérarchique ascendante et d'une méthode hiérarchique descendante.

La mesure utilisée pour effectuer les rapprochements entre termes est définie pour prendre en compte l'organisation des termes déjà labels de concepts dans l'ontologie. Les nouveaux termes rapprochés par les méthodes hiérarchiques sont proposés pour être ajoutés à l'ontologie comme sous-concepts des concepts à partir des labels avec lesquels ils sont liés. L'évaluation a consisté à prendre une ontologie existante portant sur le domaine du tourisme et à supprimer des concepts puis d'essayer de les retrouver et de les replacer correctement. Le nombre moyen de réponses correctes ainsi que le degré d'erreur de placements sont comparés aux résultats obtenus par l'algorithme *des k plus proches voisins*[1] sans prendre en compte l'ontologie existante. L'algorithme de classification des *k plus proches voisins* a l'avantage de ne mettre en place aucun apprentissage et considère qu'un nouvel objet doit être ajouté à la classe existante dont il est le plus similaire parmi les *k plus proches*. Les résultats ne permettent pas d'établir d'amélioration des performances en prenant en compte l'ontologie existante. Les concepts pères sont seulement mieux retrouvés à partir de l'algorithme ascendant. Les résultats obtenus par cette méthode ne permettent pas de mettre en valeur son intérêt.

Un autre type d'approche vise à mettre à jour les ontologies à partir de la méthode des signatures de thématique. Habituellement utilisée pour la génération de résumés, cette approche consiste à trouver un ensemble de termes relatifs à une thématique et à pondérer le lien entre chaque terme et la thématique. Appliquée à la construction d'ontologies, cette méthode permet de réaliser la signature de concepts. La méthode proposée dans Agirre [Agirre 2000] vise à mettre à jour WordNet par rapport à un corpus donné, en supprimant les sens inutiles des concepts et en en proposant de nouveaux, extraits de documents du Web. La première étape consiste à rechercher des documents du Web relatifs aux concepts de WordNet, les requêtes sont formulées à partir des termes contenus dans les synsets ainsi que les hyperonymes trouvés dans WordNet. Une collection est ainsi créée pour chaque concept.

La deuxième étape consiste à calculer pour chaque terme des documents sa fréquence d'apparition dans les différentes collections. Les termes qui ont une statistique différente dans une collection sont retenus pour constituer la signature du concept. Les termes de la signature sont retenus comme labels du concept. Les termes qui sont partagés par différentes signatures permettent de trouver des liens entre les concepts. Dans Agirre [Agirre 2000], l'ontologie ainsi

[1] http://www.lri.fr/~aze/enseignements/bibs/2008-2009/M2/docs/cours/kppv.pdf

mise à jour est testée sur le corpus **Semcor**[2] pour une tâche de désambiguïsation. Les résultats qu'elle permet d'obtenir sont plus précis que ceux obtenus par l'utilisation de WordNet non modifié. La limite de cette approche est qu'elle détecte de nouveaux liens entre concepts à partir de leur signature mais ne permet pas de déterminer comment ces liens doivent être interprétés et où les concepts doivent être ajoutés dans l'ontologie.

Dans Alfonseca [Alfonseca 2002], la méthode de la signature des thématiques est également utilisée pour mettre à jour WordNet. La même démarche d'interrogation du Web est réalisée pour extraire la collection d'apprentissage. La différence est que cette approche vise à proposer de nouveaux concepts et leur placement dans l'ontologie à partir des termes co-occurrent autour des termes définissant les concepts dans WordNet. La méthode permet donc d'affiner l'ontologie en déterminant où il faut ajouter les nouveaux concepts. Le principe de la méthode repose sur l'hypothèse de la distribution sémantique «le sens d'un mot est corrélé au contexte dans lequel il apparaît». Les termes apparaissant fréquemment dans les documents autour des termes issus de WordNet sont retenus pour être de nouveaux concepts. Une signature de thématique est réalisée pour chacun des concepts (ceux de WordNet ainsi que les nouveaux concepts proposés). Un algorithme parcourt ensuite l'ontologie de sa racine vers ses fils en calculant, au niveau de chaque concept, la similarité entre la signature de ce concept et celle du concept à ajouter. Au niveau de chaque nœud, le fils choisi est celui qui a la similarité la plus forte. L'algorithme s'arrête lorsque le score d'un concept est supérieur à celui de ses fils. Le procédé est entièrement automatique. Il a l'avantage d'extraire de nouveaux termes et de les intégrer directement dans l'ontologie par la création de nouveaux concepts définis à partir de plusieurs labels (les termes de la signature des nouveaux concepts). Bien qu'ils soient obtenus automatiquement, les résultats doivent cependant être validés par un expert.

Les méthodes de mise à jour d'une ontologie se sont essentiellement concentrées sur la détection de nouveaux termes à ajouter à l'ontologie et sur l'intégration de ces termes par la création de concepts rattachés à l'ontologie par la relation «est un». La détection de relations associatives est un élément important dans la génération d'ontologies car elles permettent de spécifier le sens des concepts.

4. Utilisation des ontologies en RI :

Un des enjeux actuels de la RI est de développer des systèmes capables d'intégrer plus de sémantique dans leurs traitements. L'objectif est double: «comprendre» les contenus des documents et «comprendre» le besoin de l'utilisateur pour pouvoir les mettre en relation.

Les ontologies sont utilisées pour représenter des descriptions partagées et plus ou moins formelles de domaine et ainsi ajouter une couche sémantique aux systèmes informatiques. C'est donc naturellement que des travaux sur l'intégration des ontologies dans les SRI se développent. Une première solution vise à construire une ontologie à partir du ou des corpus sur lesquels les tâches de RI vont être réalisées [Saias 2003] [Koo 2003] Cette solution assure a priori l'adéquation entre l'ontologie construite, le corpus et la tâche à réaliser. Cette solution reste cependant coûteuse et ne prend pas en compte l'existence de ressources qui pourraient être réutilisées. De plus, grâce à l'intérêt croissant pour les ontologies dans le domaine des systèmes d'information, de plus en plus d'ontologies sont maintenant accessibles. Une seconde solution est alors la réutilisation de ressources existantes. Dans ce cas-là, les ontologies sont généralement choisies uniquement à partir du domaine de connaissance qu'elles abordent [Vallet 2005] [Baziz 2005] [Hearst 1997].

[2] http://www.gabormelli.com/RKB/SemCor_Corpus

4.1. Similarités entre concepts dans une ontologie :

L'évaluation du lien sémantique entre deux concepts dans une ontologie est un problème de longue date dans le domaine de l'intelligence artificielle et de la psychologie. La similarité sémantique est une évaluation du lien sémantique entre deux concepts dont le but est d'estimer le degré par lequel les concepts sont proches dans leur sens [Resnik 1999].

La définition donnée par Lin de la similarité sémantique repose sur trois suppositions [Lin 1998]. La similarité entre deux concepts est liée aux caractéristiques qu'ils ont en commun (plus ils ont de caractéristiques communes, plus les concepts sont similaires) et à leurs différences (plus deux concepts sont différents, moins ils sont similaires). La similarité maximale est obtenue lorsque deux concepts sont identiques.

La majorité des travaux portant sur le calcul de similarité dans une ontologie considèrent que la similarité peut être évaluée uniquement à partir des liens taxonomiques (ou lien «est un») [Rada 1989] [Resnik 1995] [Wu 1994] [Jiang 1997]. D'autres, au contraire, estiment que ce calcul doit intégrer les autres types de liens [Lord 2003] [Thieu 2004].

4.1.1. Similarité dans une taxonomie :

Différentes mesures ont été définies pour permettre le calcul de la similarité entre concepts. Ces mesures sont classées par rapport aux caractéristiques des concepts permettant d'évaluer la similarité. Ces caractéristiques reposent soit sur la distance entre les concepts à travers leurs liens dans l'ontologie, soit sur l'information contenue par les concepts, soit sur les deux.

L'exemple de hiérarchie de concepts présenté dans la figure est utilisé pour illustrer les différentes mesures. Les concepts sont représentés par des rectangles et les flèches symbolisent la relation « est un ».

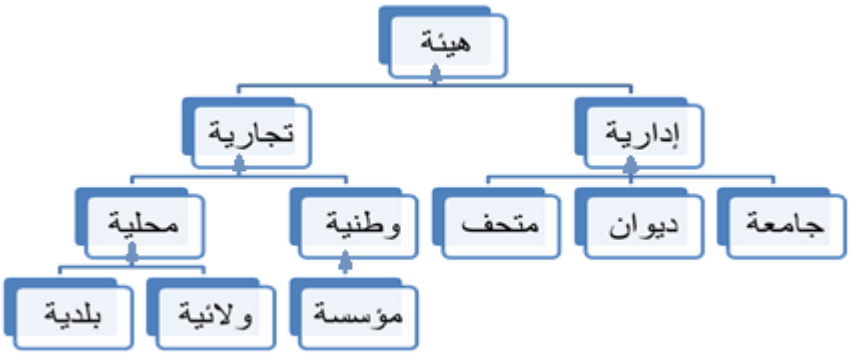


Figure 15 : Exemple de taxonomie

4.1.1.1. Mesures reposant sur la distance :

Les mesures reposant sur la distance considèrent que la similarité entre deux concepts peut être calculée à partir du nombre de liens qui séparent les deux concepts. Plusieurs variantes existent en fonction du chemin pris en compte pour calculer la distance entre les concepts. La mesure du **edge counting** [Rada 1989] évalue la distance sémantique à partir du nombre de branches séparant les concepts par le plus court chemin dans la hiérarchie.

A partir de l'exemple de la figure 15, la distance se calcule de la façon suivante :

- $Dist_{edge}$(ولائية , بلدية)=2
- $Dist_{edge}$(إدارية, ديوان)=1 $Dist_{edge}$(محلية , تجارية)=1

- $Dist_{edge}$(بلدية, إدارية)=4 $Dist_{edge}$(بلدية, مؤسسة)=4

La similarité sémantique entre deux concepts correspond à l'inverse de la distance entre deux concepts. Plus deux concepts sont distants, moins ils sont similaires. A partir de la mesure de distance précédente, Leacock [Leacock 1998] a proposé la formule suivante pour calculer la similarité.

Elle est issue de la proposition et assure la normalisation de cette dernière.

$Sim_{edge}(c1,c2) = -\log\frac{Distedge(c1,c2)}{2*\text{Max}}$

Où max étant la profondeur maximale de la taxonomie

L'utilisation de la fonction –log permet de normaliser la similarité entre [0,1] (1 signifiant que les concepts sont totalement similaires).

- sim_{edge}(بلدية , ولائية)= $-log\frac{2}{2*4}$ =0,6
- sim_{edge}(إدارية, ديوان)= 0,9 sim_{edge}(تجارية , محلية)=0,9
- sim_{edge}(بلدية, إدارية)=0,3 sim_{edge}(بلدية, مؤسسة)=0,3

Wu et Palmer [Wu 1994] ont proposé une autre mesure de similarité prenant en compte à la fois la profondeur des concepts dans la hiérarchie de concepts et la structure de la hiérarchie de concepts.
Pour calculer la similarité entre deux concepts c1 et c2, la formule suivante est utilisée :

$Sim_{Wu}(c_1,c_2) = \frac{2*depth\,(c)}{depth\,(c1)+depth\,(c2)}$

où depth(c_i) correspond au niveau de profondeur du concept c_i et c représente le concept le plus spécifique qui généralise c_1 et c_2.

La valeur de la similarité est comprise entre 0 et 1 (1 signifiant que les concepts sont totalement similaires).

- sim_{Wu}(بلدية , ولائية)= 2*3/(4+4)=3/4 (0.75) (محلية *étant le plus spécifique subsumeur)*
- sim_{Wu}(إدارية, ديوان)= 2*2/(2+3)=4/5 (0.8) sim_{Wu}(تجارية , محلية)=4/5(0.8)
- sim_{Wu}(بلدية, إدارية)=2*1/(2+4)=1/3(0.33) sim_{Wu}(بلدية, مؤسسة)=2*2/(4+4)=1/2(0.5)

Cette mesure est plus pertinente que les mesures précédentes reposant uniquement sur le chemin le plus court entre les deux concepts, car elle prend en compte l'organisation hiérarchique des concepts, c'est-à-dire le concept généralisant les deux concepts considérés. La similarité entre *بلدية* et *إدارية* devient plus basse par sim_{Wu} que celle entre *بلدية* et *مؤسسة* , alors qu'elles étaient identiques par les mesures reposant sur le «edge counting». Cette différence est pertinente dans la mesure où *مؤسسة* et *بلدية* ont un concept commun les spécifiant *(تجارية)* plus proche ces deux concepts dans la hiérarchie que *بلدية* et *إدارية*. Cependant, cette mesure admet que la distance sémantique entre deux concepts reliés par la relation «est un» est égale, alors que ce n'est pas forcément le cas. La distance sémantique portée par le lien «est un» entre *إدارية* et *ديوان* et celle portée par ce même type de lien entre *تجارية* et *محلية* est considérée de la même façon alors que ces deux relations ne témoignent pas du même niveau de spécificité (on passe directement de la catégorie *إدارية* à *ديوان* sans spécifier les catégories intermédiaires comme cela était le cas pour la catégorie *تجارية*). Les choix arbitraires pris lors de la construction de la hiérarchie de concepts influencent donc la valeur de la similarité. Afin de prendre en compte le fait que les liens dans

une ontologie ne représentent pas la même distance sémantique, plusieurs solutions ont été proposées.

L'une d'elles consiste à prendre en compte la densité locale de l'ontologie au niveau des concepts considérés pour contrecarrer la subjectivité dans le choix des relations. L'utilisation de la densité repose sur l'observation suivant laquelle les concepts appartenant à une partie dense en concepts de l'ontologie sont sémantiquement plus proches que ceux appartenant à une partie éparse de l'ontologie [Mc Hale 1998]. Dans la densité est représentée par le nombre de liens fils du concept. Plus un concept a de fils, plus la similarité entre le concept et ses fils est élevée. Le point faible de ce critère est que la distribution des concepts dans l'ontologie doit refléter la distribution des concepts dans le domaine. Cependant, comme nous l'avons vu précédemment, les choix faits lors de l'élaboration de la hiérarchie de concepts sont généralement liés à la connaissance utile pour la tâche pour laquelle elle est construite et ne reflète pas forcément le domaine dans son ensemble.

Une autre solution consiste à prendre en compte le contenu en information des concepts.

4.1.1.2. Mesures reposant sur le contenu en information des concepts :

Les approches reposant sur le contenu en information supposent que l'information détenue par les concepts puisse être quantifiée. La similarité est alors calculée à partir de l'information partagée par les concepts. Les différentes mesures proposées pour évaluer la similarité entre concepts reposent sur ce calcul.

4.1.1.2.1. Calcul du contenu en information d'un concept :

Afin de calculer l'information contenue par les concepts, un corpus de référence est choisi. Les concepts sont alors pondérés par une fonction correspondant à l'information portée par un concept dans le corpus. Le contenu en information du concept est défini par ses occurrences dans le corpus ainsi que celles des concepts qu'il subsume. Il a pour objectif d'utiliser la probabilité d'obtenir un concept dans un corpus documentaire afin de contrecarrer la subjectivité dans le choix des relations « est un » de l'ontologie.

Le contenu en information d'un concept c se calcule de la façon suivante [Resnik 1995] :

$$CI(c) = -\log(p(c))$$

$$Freq(c) = \sum_{n \in word(c)} count(n) \qquad p(c) = \frac{freq(c)}{N}$$

Où *word(c)* est l'ensemble des termes ou labels représentant le concept c et les concepts subsumés par c, count(n) le nombre d'occurrences du terme n dans le corpus et N le nombre total d'occurrences des labels de concepts retrouvés dans le corpus. L'utilisation de la fonction -log permet de réduire le contenu en information d'un concept ayant une forte probabilité d'apparition dans le corpus.

Un inconvénient du calcul du contenu en information d'un concept proposé par Resnik est que la probabilité d'obtenir un concept est calculée à partir des labels retrouvés dans le corpus sans vérifier que chacune des occurrences se rapporte effectivement au concept. Ceci peut poser problème dans le cas où le label est ambigu et désigne différents concepts. Le contenu en information d'un concept peut ainsi être amplifié par la forte occurrence d'un label désignant un autre concept dans le corpus.

Une solution plus adaptée serait de désambiguïser le label dans les documents et de comptabiliser uniquement les occurrences qui se rapportent au concept en question. Cette solution ne peut pas être mise en place dans la mesure où Resnik [Resnik 1999] utilise ces

mesures justement pour désambiguïser les termes. Afin de prendre en compte les labels ambigus, une autre formule a été proposée dans [Sanderson 1995]. Dans cette formule (freq2), la fréquence d'apparition d'un label est normalisée par rapport au nombre de concepts auxquels il se rapporte.

$$\text{Freq2(c)} = \sum_{n \in word(c)} \frac{count(n)}{nbclasse(n)}$$

Où nbclasse(n) correspond au nombre de concepts dont le terme n est label

4.1.1.2.2. Interprétation du contenu en information :

Plusieurs mesures de similarité ont été définies à partir du contenu en information des concepts [Resnik 1995] [Jiang 1997]. Elles reposent sur le principe que l'information commune aux deux concepts est capturée dans le contenu en information du concept le plus spécifique qui subsume les deux concepts.

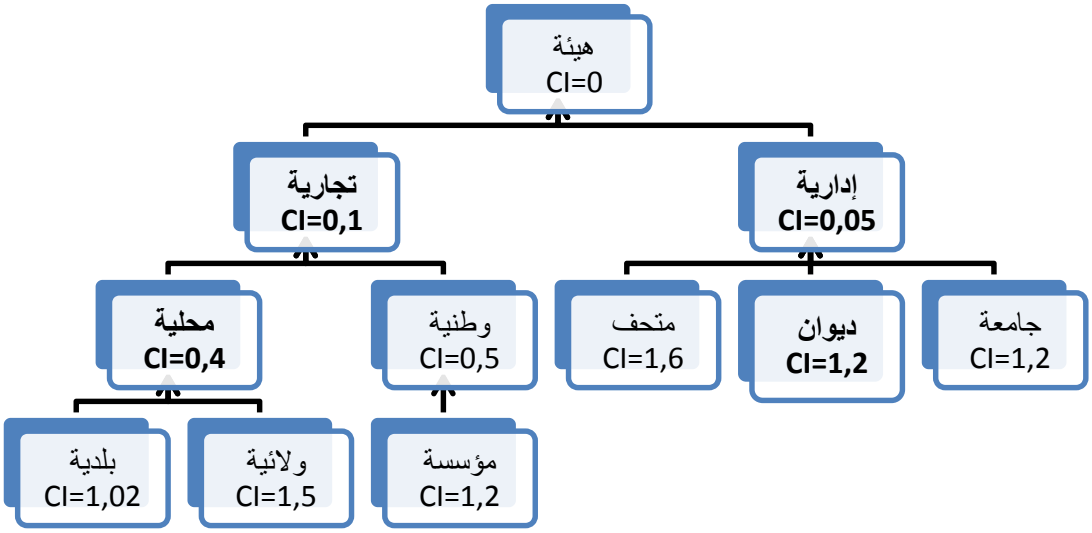


Figure 16: Hiérarchie de concepts augmentée par le contenu en information des concepts

En prenant par exemple les concepts ***محلية*** et ***تجارية***, l'information partagée par ces deux concepts est le contenu en information du concept ***تجارية***, c'est-à-dire 0,1.
L'information commune aux concepts ***إدارية*** et ***ديوان*** est 0,05. Ainsi l'information partagée par deux concepts permet de capturer le poids sémantique du lien « est un ».
Afin de pouvoir calculer la similarité entre deux concepts, la probabilité d'obtenir dans le corpus le concept le plus spécifique subsumant c1 et c2 est définie de la façon suivante :

$$Pms(c1, c2) = \min_{c \in S(c1,c2)} \{P(c)\}$$

Où S(c1,c2) est l'ensemble de concepts qui subsument à la fois c1 et c2

La probabilité d'obtenir un concept prend en considération à la fois la probabilité d'obtenir un concept ainsi que tous les concepts qu'il subsume. Pms (ou min{p(c)}) revient donc à choisir le concept subsumant c1 et c2 ayant la plus faible probabilité, c'est-à-dire le concept le plus spécifique de c1 et c2 dans la hiérarchie au sens où son contenu en information est le plus faible. Quand l'ontologie permet l'héritage multiple et que plusieurs concepts subsument les deux concepts considérés, cette formule permet de choisir dans l'ensemble des concepts candidats le concept le plus spécifique au sens où c'est celui qui a la probabilité la plus faible.

4.1.1.2.3. Mesures :

La première mesure de similarité calculée à partir du contenu en information des concepts a été proposée dans Resnik [Resnik 1995]. Elle prend des valeurs entre [0,1].

- $Sim_{Resnik}(c1,c2)=[-\log(Pms(c1,c2)]$
- Sim_{Resnik}(بلدية, ولائية)= CI(محلية)=0,4 (محلية *étant le plus spécifique subsumeur)*
- Sim_{Resnik}(إدارية, ديوان)= 0,05 sim_{Wu}(محلية, تجارية)=0,1
- Sim_{Resnik}(بلدية, إدارية)=0 sim_{Wu}(بلدية, مؤسسة)=0,1

L'inconvénient de cette mesure est que deux couples de concepts qui ont le même subsumeur le plus spécifique ont la même similarité. Ceci est par exemple le cas entre (بلدية et مؤسسة) et (محلية et تجارية).

La mesure proposée par Lin vise à contrecarrer cet inconvénient [Lin 1998]. Dans le cas de l'évaluation de la similarité entre deux concepts dans une taxonomie, Lin étend ces mesures en considérant que la description de chacun de ces concepts est le contenu en information des concepts et que les caractéristiques communes aux deux concepts sont quantifiées par le contenu en information du concept le plus spécifique généralisant les deux concepts.

$$\underset{Lin}{\mathrm{Sim}}(c1,c2)=\frac{2*\log(pms(c1,c2))}{\log(p(c1))+\log(p(c2))}$$

Cette mesure a l'avantage de prendre en compte le concept le plus spécifique subsumant c1 et c2 ainsi que le contenu en information des concepts comparés. Contrairement à la mesure précédente proposée par Resnik, elle permet de différencier la similarité entre plusieurs couples de concepts ayant le même subsumeur le plus spécifique. Elle prend des valeurs entre [0,1].

$$\underset{Lin}{\mathrm{Sim}}(\text{بلدية},\text{ولائية})=\frac{2*CI(\text{محلية})}{\mathrm{CI}(\text{ولائية})+\mathrm{CI}(\text{بلدية})}=0{,}32$$

- Sim_{Lin}(إدارية, ديوان)= 0,08 sim_{Lin}(محلية, تجارية)=0,4
- Sim_{Lin}(بلدية, إدارية)=0 sim_{Lin}(بلدية, مؤسسة)=0,09

4.1.1.3. Mesures Mixtes :

Le principe des mesures mixtes est de considérer le plus court chemin reliant deux concepts dans l'ontologie et de pondérer ces liens à partir de leur poids sémantique. Le poids sémantique des liens prend notamment en compte le contenu en information des concepts [Richardson 1995] [Jiang 1997]. Dans Jiang [Jiang 1997], les liens du plus court chemin reliant les deux concepts sont pondérés à partir de quatre facteurs.

Le contenu en information du lien calculé à partir de la différence du contenu en information du concept fils c_{fils} et celui du père $c_{père}$

$$\mathrm{CI}(c_{fils}, c_{père})=\mathrm{CI}(c_{fils}) - \mathrm{CI}(c_{père})$$

La profondeur du concept père $c_{père}$ évaluée de la façon suivante :

$$\mathrm{Prof}(c_{père})= profondeur(c_{père}) +1 \ / \ profondeur\_maxi$$

La densité locale du concept père calculée à partir du nombre de nœuds fils du concept $E(c_{père})$ et le nombre moyen $\bar{E}$ de nœud fils pour un concept dans le réseau.

$$\mathrm{Densité}(c)= \bar{E} / E(c)$$

Le poids accordé au type de lien considéré $t(c_{fils}, \bar{E})$.

Le poids d'un lien est ensuite calculé de la façon suivante où α et β permettent de pondérer les différents facteurs avec $\alpha \geq 0$ et $0 \leq \beta \leq 1$:

$$\text{Poids}(\text{cfils}, \text{cpère}) = (ß + (1 - ß)\frac{\bar{E}}{E(\text{cpère})})(\frac{d(\text{cpère}) + 1}{d(\text{cpère})})^{\alpha}[CI(\text{cfils}) - CI(\text{cpère})]T(\text{cfils}, \text{cpère})$$

La distance entre deux concepts c1 et c2 est alors calculée par :

$$Dist_{Jiang}(c1,c2) = \Sigma_{c \in pcc(c1,c2)} Poids(c, père(c))$$

Où pcc(c1,c2) représente l'ensemble des concepts appartenant au plus court chemin entre c1 et c2.

Dans le cas où seul le contenu en information des liens est considéré (α et β sont nuls et le poids du lien est 1), la distance entre deux concepts peut être simplifiée par la formule suivante :

$$Dist_{Jiang}(c1,c2) = CI(c1) + CI(c2) - 2*CI(SPS(c1,c2))$$

Où SPS(c1,c2) représente le concept le plus spécifique subsumant c1 et c2

Cette dernière formule est équivalente à celle proposée par Lin. La seule différence est qu'elle permet de calculer la distance sémantique et non pas la similarité. Jiang propose de convertir la distance proposée en similarité en adaptant la formule de similarité proposée par Resnik.

$$Sim_{Jiang}(c1, c2) = (2*\max) - Dist_{Jiang}(c1,c2)$$

Où max représente la distance maximale obtenu par la formule $Dist_{Jiang}$

4.1.2. Similarité dans une ontologie faisant intervenir des liens associatifs :

Plusieurs travaux visent à permettre le calcul de la similarité entre concepts en ne limitant pas les liens considérés aux liens taxonomiques.

Dans Lord [Lord 2003], les mesures de Resnik, Lin et Jiang sont utilisées pour calculer la similarité entre concepts à partir des liens «est un» et «partie de» de l'ontologie des Gènes. Les mesures sont utilisées sans modification, en considérant les deux relations comme étant la relation père-fils permettant d'organiser hiérarchiquement les concepts. Le contenu en information d'un concept est calculé à partir de la probabilité d'obtenir les labels du concept dans le corpus ainsi que chacun de ses fils auxquels il est lié aussi bien par la relation «est un» que par la relation «partie de».

Cette considération a des limites [Hernandez 2005]. Dans ces mesures, le contenu en information du concept le plus spécifique lié aux deux concepts est considéré pour évaluer l'information commune à deux concepts. Cependant, le subsumeur de deux concepts ne peut être calculé qu'à partir du moment où une relation sémantique peut être établie entre celui-ci et les deux concepts considérés, ceci impliquant que le concept subsumeur contient effectivement des informations sur les deux concepts. Prenons l'exemple présenté dans la figure 18, il paraît intuitif que l'information commune à مادة et سلطة n'est pas détenue par تنظيم qui généralise effectivement سلطة mais pas مادة.

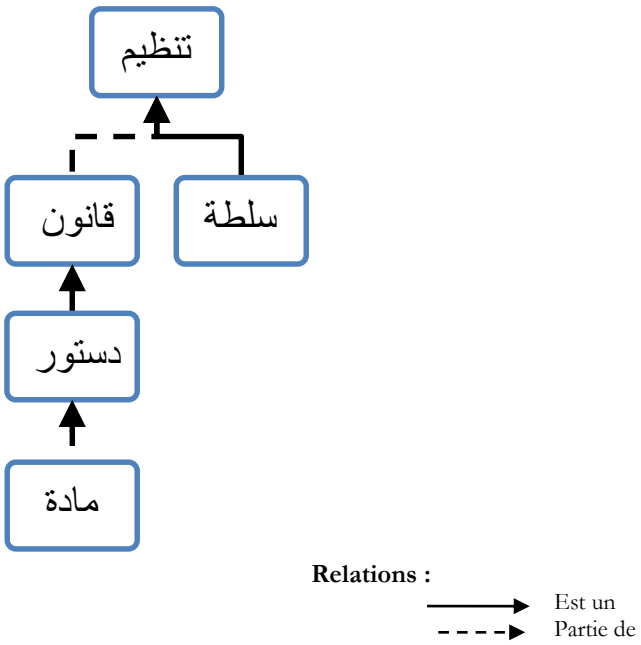


Figure 17: Exemple d'ontologie faisant intervenir deux relations « partie de » de familles différentes

L'adaptation des mesures reposant sur le contenu en information des concepts doit considérer le fait qu'une relation sémantique puisse être trouvée entre les deux concepts dont la similarité est calculée.

Dans Thieu [Thieu 2004], une adaptation de la mesure de Jiang est proposée pour prendre en compte les relations non taxonomiques. La distance sémantique impliquée par les relations taxonomiques est considérée de la même façon que dans la méthode initiale. L'adaptation porte uniquement sur le poids sémantique affecté au lien non taxonomique. Jiang propose de représenter le poids à partir de la différence du contenu en information du fils et du père. Or, dans le cas des relations non taxonomiques, le contenu en information n'est pas décroissant du père au fils car le père ne généralise pas le concept fils. De plus, le contenu en information des deux concepts calculé par la probabilité d'obtenir les labels des concepts et ceux de leur subsumeur ne contient aucun élément commun. Les auteurs proposent alors de pondérer le lien au moyen des différences de contenu en information du fils avec l'ensemble de ses pères taxonomiques. Bien que cette adaptation ait l'avantage de s'adapter à n'importe quel type de lien, elle ne prend pas en compte le contenu en information du nœud père et ceci implique une perte d'information. Le poids de tous les liens non-taxonomiques partant d'un même nœud est en effet égal.

D'autres mesures ne prenant pas en compte le contenu en information des concepts ont été proposées.

Dans Ehrig [Ehrig 2005], la mesure de similarité reposant sur le nombre d'arcs taxonomiques du plus court chemin entre les concepts, proposée par Rada, est utilisée pour calculer la similarité sémantique à partir de n'importe quel type de relation composant le plus court chemin. L'inconvénient de cette approche est d'admettre que les relations sémantiques représentent la même distance sémantique entre les concepts. Ceci pose problème parce que chaque type de relations implique sa propre sémantique. Aussi, comme nous l'avons vu précédemment, une même relation, telle que la relation «est un», peut impliquer un degré d'engagement différent.

4.2. Quelle ontologie choisir ?

La majorité des approches de RI visant à intégrer une ontologie dans leur procédé reposent sur des ontologies existantes [Hearst 1997] [Vallet 2005] [Baziz 2005]. Généralement, l'unique caractéristique prise en compte dans le choix de l'ontologie est le domaine de connaissance représentée dans l'ontologie qui doit couvrir le domaine traité dans le corpus ainsi d'analyser dans quelle mesure les ontologies sont réutilisables.

4.2.1. Réutilisabilité des ontologies :

La construction d'ontologies réutilisables est le but affiché d'un certain nombre de travaux [Gómez-Pérez 1996] [Fernandez 1997] [Uschold 1996].

Cependant, de nombreux auteurs considèrent que les ontologies sont non réutilisables. Bachimont affirme en effet que par leur méthode de construction et les travaux épistémologiques qui les supportent, leur réutilisation est impossible [Bachimont 1996]. De la même façon, Charlet considère que «*les ontologies sont des artefacts construits en fonction d'une tâche précise et ne peuvent être réutilisées, en tant qu'objet formel, pour une autre tâche.*». Il affirme également que les travaux sur la génération d'ontologies à partir de corpus montrent une dépendance forte entre la construction de corpus et la construction de la future ontologie. «*Le corpus est porteur via les expressions linguistiques qui en sont extraites, des futurs concepts de l'ontologie*» [Charlet 2002].

Selon Furst [Furst 2004], les ontologies sont destinées à être réutilisées. La sémantique qu'elles représentent est liée au cadre applicatif à partir duquel le sens des termes et concepts est défini. Cependant, la représentation ne dépend pas de l'opération faite avec l'ontologie. La sémantique de l'ontologie est liée au contexte mais la représentation n'implique pas que l'ontologie soit utilisée uniquement dans le contexte de sa création.

4.2.2. Evaluer la réutilisation d'une ontologie :

Afin d'évaluer la réutilisabilité des ontologies plusieurs démarches sont suivies :

La première consiste à considérer une ontologie existante et à décrire les étapes et le coût impliqués par le processus de réutilisation dans une application donnée. Cette démarche est suivie notamment par Uschold qui recommande la création d'ontologies à partir de la réutilisation d'ontologies existantes plutôt qu'en partant de rien [Uschold 1995]. Les travaux présentés dans Uschold [Uschold 1998] et Pinto [Pinto 2001] proposent une analyse du procédé impliqué par la réutilisation d'une ontologie formelle pour la construction d'une nouvelle et les étapes nécessaires à l'application de cette ontologie dans un nouveau système. Les conclusions de ces deux analyses montrent que l'automatisation de ce procédé est loin d'être envisageable dans la mesure où il nécessite des connaissances extérieures liées à l'ontologie et à la tâche à réaliser.

Une autre approche consiste à évaluer la réutilisabilité de l'ontologie par rapport à certains critères voire certaines mesures. Une première tentative consiste à reprendre les mesures d'évaluation des SRI basées sur les notions de précision (proportion de documents correctement retournés par rapport à l'ensemble des documents retournés par le système) et de rappel (proportion de documents correctement retournés par rapport à l'ensemble des documents pertinents dans la collection) [Salton 1971]. Cependant, l'évaluation d'une ontologie n'est pas aussi triviale. Les notions de rappel et de précision devraient être comprises de la façon suivante. La précision correspondrait à évaluer la quantité de connaissance correctement identifiée dans l'ontologie par rapport à toute la connaissance contenue dans l'ontologie en fonction de la tâche à réaliser. Le rappel serait la quantité de connaissance correctement définie dans l'ontologie par

rapport à la connaissance qui devrait être identifiée. Le problème est qu'il est impossible de déterminer ces ensembles de connaissances ; ils dépendent en effet des différentes interprétations et des différents types de connaissance que l'on souhaite représenter.

D'autres solutions ont donc été proposées pour permettre l'évaluation d'une ontologie. Elles peuvent être regroupées en deux types d'analyse: l'analyse qualitative ou l'analyse quantitative. Ces analyses peuvent être appliquées pour évaluer l'adéquation entre une hiérarchie de concepts et un corpus.

4.2.2.1. Analyse qualitative et analyse quantitative :

4.2.2.1.1. Une analyse qualitative :

Une analyse qualitative consiste à évaluer une ontologie ou ses parties et à mesurer son taux de pertinence. Cependant, se posent deux questions : qui parmi l'utilisateur de l'ontologie, un ou plusieurs experts du domaine et le concepteur de l'ontologie est censé donner ce taux? Quels sont les critères qui devront être pris en compte dans l'évaluation?

Guarino [Guarino 1998] et Gomez Perez [Gómez-Pérez 1999] répondent à cette dernière question en proposant des critères fondés sur les principes utilisés lors de la construction de l'ontologie. Les critères proposés par Gomez [Gómez-Pérez 1999] sont les suivants :

- La consistance de l'ontologie : la possibilité d'obtenir des conclusions contradictoires à partir des inférences possibles sur l'ontologie est ici évaluée,
- La complétude de l'ontologie : l'ontologie recouvre toute la connaissance qu'elle est censée représenter et chacune de ses définitions contient bien tous les éléments nécessaires,
- La concision de l'ontologie : l'ontologie ne contient pas de connaissance inutile ou redondante,
- L'expansibilité de l'ontologie : l'ajout de connaissance dans l'ontologie est possible,
- La sensibilité de l'ontologie : le changement d'une définition n'altère pas toutes les autres définitions.

Ces critères restent cependant très théoriques et nécessitent leur évaluation par les concepteurs de l'ontologie, ceux-ci étant capables de les évaluer à partir de la sémantique qu'ils associent à ces critères. Ces critères sont de plus indépendants de la tâche pour laquelle l'ontologie est réutilisée. Ce type de critères peut également amener à la construction d'ontologies non opérationnelles par l'absence de la prise en compte de la tâche [Milks 2002].

Lozano-Tello [Lozano-Tello 2004] propose de nouveaux critères en créant un cadre d'évaluation multidimensionnelle aidant au choix d'une ontologie pour une tâche donnée. Les critères pris en compte sont regroupés sous forme hiérarchique à partir de cinq dimensions : le contenu de l'ontologie, le langage de représentation, la méthodologie suivie lors de la construction, les outils pouvant utiliser l'ontologie, et le coût de la réutilisation. Chaque dimension est ensuite déployée par rapport à des caractéristiques, pouvant elles-mêmes être détaillées par des sous caractéristiques plus spécifiques. L'ensemble de ces critères est pondéré par un expert en développement d'ontologie en fonction du cadre de réutilisation de l'ontologie. Une analyse multidimensionnelle des valeurs affectées aux différents critères permet ensuite à l'expert de choisir quelle ontologie utiliser. L'inconvénient majeur de cette approche est qu'elle nécessite un investissement important de la part de l'expert dans l'évaluation des critères.

4.2.2.1.2. Analyse quantitative :

Une analyse quantitative consiste quant à elle à évaluer la réutilisabilité d'une ontologie par rapport à son efficacité dans la réalisation d'une tâche donnée. Une évaluation de ce type consisterait par exemple à prendre plusieurs ontologies différentes et à exécuter une même tâche

avec chacune d'entre elles, puis à comparer les résultats obtenus. Dans la mesure où une ontologie doit permettre l'interopérabilité entre applications, il est possible de concevoir son évaluation par rapport aux résultats qu'elle permet et non pas par rapport à l'interprétation qu'en fait un être humain [Brewster 2004].
Un exemple d'analyse quantitative est donné dans Porzel [Porzel 2004]. L'évaluation repose sur la définition d'une tâche, la considération d'une ou plusieurs ontologies dites légères, la définition d'une application réalisant la tâche par l'intermédiaire d'un algorithme prenant en compte l'ontologie, et la définition de références donnant les réponses correctes qui doivent être retournées par l'application. L'efficacité de l'ontologie est examinée par rapport aux différents niveaux de l'ontologie (vocabulaire, taxonomie, relation non taxonomique) suivant trois critères :

- Les erreurs d'insertion (concepts, relations taxonomiques, relations non taxonomiques non appropriés),
- Les erreurs de délétion (concepts, relations taxonomiques, relations non taxonomiques manquants)
- Les erreurs de substitution (concepts, relations taxonomiques, relations non taxonomiques ambigus).

La tâche choisie est l'étiquetage par des relations ontologiques des relations entre entités de l'ontologie trouvées dans des textes. Cette tâche peut être apparentée à un système de désambiguïsation dont le but est d'identifier les concepts abordés dans le texte et d'extraire les relations sémantiques liant ces concepts. Le point critique de cette approche est la réalisation de l'ensemble de références. Elle demande l'intervention d'experts et pénalise l'évaluation car elle implique qu'une personne réalise la tâche.

Dans Maedche [Maedche 2002], des mesures sont définies pour comparer deux ontologies entre elles et plus particulièrement un ensemble d'ontologies par rapport à une ontologie de référence. Les ontologies considérées dans ces travaux sont des ontologies dites légères. La comparaison des deux ontologies repose sur une analyse lexicale (comparaison des termes des ontologies) et une analyse conceptuelle (comparaison de l'organisation des concepts dans l'ontologie). Ces mesures donnent des résultats concluants dans la comparaison d'une ontologie de référence d'un domaine réalisée par des experts et des ontologies du même domaine réalisées par des étudiants.

Un autre type d'analyse quantitative consiste justement à comparer l'adéquation entre une ou plusieurs ontologies par rapport à un corpus. Ce type d'analyse permet d'évaluer si la connaissance contenue dans l'ontologie se rapporte à l'information contenue dans un corpus.

5. Conclusion

La capture de la connaissance nécessaire pour l'élaboration d'une ontologie peut être réalisée à partir de plusieurs principes et méthodologies. L'élaboration d'ontologies à partir de textes permet de faciliter la conception d'ontologies. Elle peut reposer soit sur une analyse statistique des termes apparaissant dans les documents, soit sur une analyse syntaxique qui consiste à analyser le rôle grammatical des mots qui les composent. Ces deux approches permettent tout d'abord d'aider à extraire les termes qui définiront le lexique de l'ontologie et qui seront les labels des concepts et des relations. Elles permettent également d'aider à définir la structure de l'ontologie à partir de l'extraction de relations taxonomiques et de relations non taxonomiques. Ces méthodes nous permettront de construire notre ontologie dédiée au Journal Officiel. .

Chapitre 3 :
Construction d'une ontologie pour le Journal Officiel

Chapitre 3 : Construction d'une ontologie pour le Journal Officiel :

1. Le Journal Officiel de la République algérienne démocratique et populaire « الجريدة الرسمية للجمهورية الجزائرية الديمقراطية الشعبية »:

Pour pouvoir modéliser les textes du Journal Officiel, nous allons présenter dans ce chapitre sa définition, son historique, ses caractéristiques, et nous terminerons par l'étude de ses approches informatiques afin de modéliser son contenu.

1.1 Définition :

Le Journal Officiel de la République algérienne démocratique et populaire (JORA) fait suite au Journal Officiel de l'État Algérien (JOEA). Il publie tous les textes juridiques algériens (lois et décrets, ordonnances, arrêtés ...) et d'autres informations officielles.

1.2 Historique :

De juin 1958 à la veille de l'indépendance de l'Algérie, la publication des textes juridiques (lois et décrets, ordonnances, arrêtés ...) la concernant, était assurée par le Recueil des actes administratifs de la Délégation Général du Gouvernement [JORADP]. Ce dernier fait suite au Journal Officiel de l'Algérie fondé en janvier 1927.

Au lendemain de l'indépendance, proclamée le 5 juillet 1962, parait pour la première fois le Journal Officiel de l'État algérien[3] (JOEA). Son premier numéro, de douze pages, et daté du 6 juillet 1962, est principalement consacré à l'indépendance du pays. Il publie les résultats du référendum d'autodétermination du 1er juillet 1962 avec une erreur typographique : il est mentionné 15 juillet 1962 (au lieu du 5 juillet 1962) ; la lettre du général de Gaulle, président de la République française, qui reconnait l'indépendance de l'Algérie, à Abderrahmane Farés, président de l'Exécutif Provisoire de l'État Algérien et la lettre du président de l'Exécutif Provisoire de l'État algérien qui prend acte.

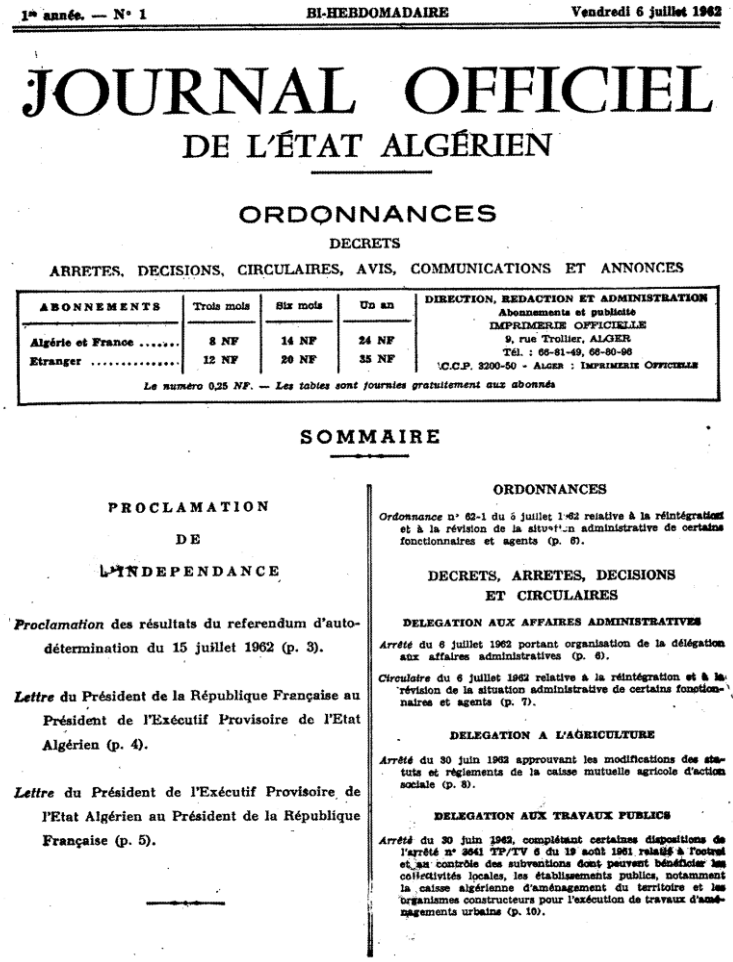
Le sommaire du premier numéro du Journal Officiel du 6 juillet 1962[4] est :

- Proclamation de l'Indépendance
- Proclamation des résultats du référendum d'autodétermination du 15 juillet 1962
- Lettre du Président de la République Française au Président de l'Exécutif Provisoire de l'Etat Algérien
- Lettre du Président de l'Exécutif Provisoire de l'Etat Algérien au Président de la République Française
- Ordonnances
- Ordonnances du 6 juillet relative à la réintégration et à la révision de la situation administrative de certains fonctionnaires et agents
- Décrets, arrêtés, décisions et circulaires
- Délégation aux affaires administratives

[3] : Journal Officiel de l'État Algérien du vendredi 6 juillet 1962

[4] : Journal Officiel de l'État Algérien du vendredi 6 juillet 1962

- Délégation à l'agriculture
- Délégation aux travaux publics
- Avis et communications
- Avis portant modification du régime commercial du point-d'arrêt de Tlétat-des-Douaïrs (ligne de Blida à Djelfa)
- Avis d'appel d'offre ouverts - Union des S.A.P. de Bougie
- Avis d'appel d'offre avec concours - Union des S.A.P de Bougie
- Situation de la banque d'Algérie au 30 avril 1962



1re année. — N° 1 BI-HEBDOMADAIRE Vendredi 6 juillet 1962

JOURNAL OFFICIEL
DE L'ÉTAT ALGÉRIEN

ORDONNANCES
DECRETS
ARRETES, DECISIONS, CIRCULAIRES, AVIS, COMMUNICATIONS ET ANNONCES

ABONNEMENTS	Trois mois	Six mois	Un an
Algérie et France	8 NF	14 NF	24 NF
Etranger	12 NF	20 NF	35 NF

DIRECTION, REDACTION ET ADMINISTRATION
Abonnements et publicité
IMPRIMERIE OFFICIELLE
9, rue Trollier, ALGER
Tél. : 66-81-49, 66-80-96
C.C.P. 3200-50 - ALGER : IMPRIMERIE OFFICIELLE

Le numéro 0,25 NF. — Les tables sont fournies gratuitement aux abonnés

SOMMAIRE

PROCLAMATION
DE
L'INDEPENDANCE

Proclamation des résultats du referendum d'autodétermination du 15 juillet 1962 (p. 3).

Lettre du Président de la République Française au Président de l'Exécutif Provisoire de l'Etat Algérien (p. 4).

Lettre du Président de l'Exécutif Provisoire de l'Etat Algérien au Président de la République Française (p. 5).

ORDONNANCES

Ordonnance n° 62-1 du 5 juillet 1962 relative à la réintégration et à la révision de la situation administrative de certains fonctionnaires et agents (p. 6).

DECRETS, ARRETES, DECISIONS
ET CIRCULAIRES

DELEGATION AUX AFFAIRES ADMINISTRATIVES

Arrêté du 6 juillet 1962 portant organisation de la délégation aux affaires administratives (p. 6).

Circulaire du 6 juillet 1962 relative à la réintégration et à la révision de la situation administrative de certains fonctionnaires et agents (p. 7).

DELEGATION A L'AGRICULTURE

Arrêté du 30 juin 1962 approuvant les modifications des statuts et règlements de la caisse mutuelle agricole d'action sociale (p. 8).

DELEGATION AUX TRAVAUX PUBLICS

Arrêté du 30 juin 1962, complétant certaines dispositions de l'arrêté n° 2641 TP/TV 6 du 19 août 1961 relatif à l'octroi et au contrôle des subventions dont peuvent bénéficier les collectivités locales, les établissements publics, notamment la caisse algérienne d'aménagement du territoire et les organismes constructeurs pour l'exécution de travaux d'aménagements urbains (p. 10).

Figure 18: Journal Officiel de l'État Algérien (JOEA) du vendredi 6 juillet 1962[joradp]

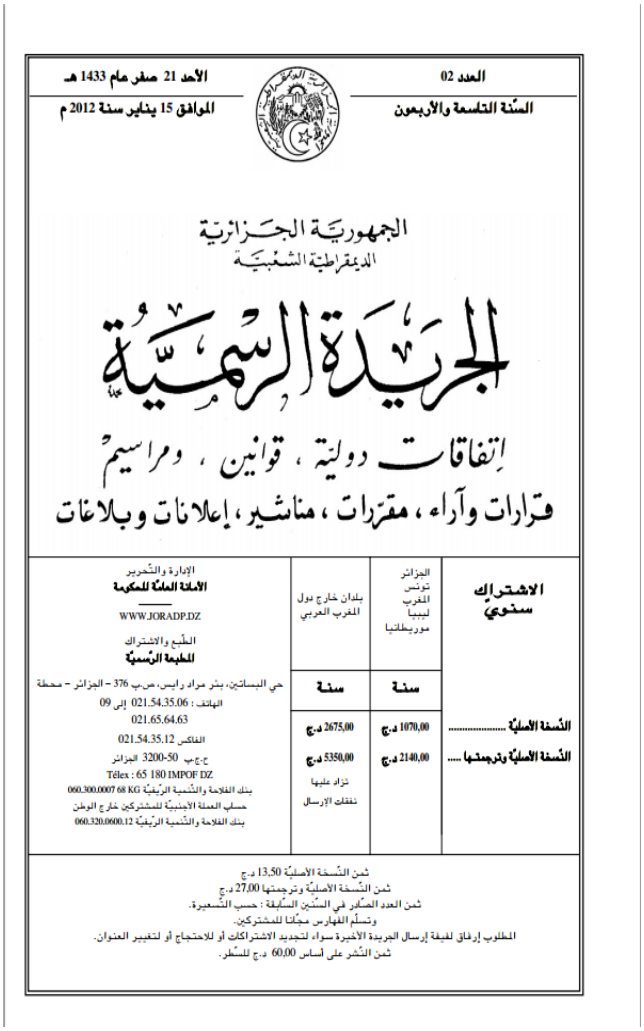
Ces textes sont les premiers textes officiels de la République algérienne.

Le deuxième numéro du Journal Officiel, daté du 17 juillet 1962[5], publie l'ordonnance n° 62-2 qui ordonne l'amnistie des faits commis avant le 20 mars 1962.

À compter du 26 octobre 1962[6], le JOEA devient le Journal Officiel de la République Algérienne Démocratique et Populaire (JORA).

[5] : Journal Officiel de l'État Algérien du mardi 17 juillet 1962.

[6] : Journal Officiel de la République Algérienne Démoctaique et Populaire du vendredi 26 octobre 1962.



الأحد 21 صفر عام 1433 هـ
الموافق 15 يناير سنة 2012 م

العدد 02
السّنة التاسعة والأربعون

الجمهورية الجزائرية
الديمقراطية الشعبية

الجريدة الرسمية

اتفاقات دولية ، قوانين ، ومراسيم
قرارات وآراء ، مقرّرات ، مناشير ، إعلانات وبلاغات

الاشتراك سنوي	الجزائر تونس المغرب ليبيا موريطانيا	بلدان خارج دول المغرب العربي
	سنة	سنة
النسخة الأصليّة	1070,00 د.ج	2675,00 د.ج
النسخة الأصليّة وترجمتها	2140,00 د.ج	5350,00 د.ج
		تزاد عليها نفقات الإرسال

الإدارة والتّحرير
الأمانة العامّة للحكومة
WWW.JORADP.DZ
الطّبع والاشتراك
المطبعة الرّسميّة
حي البساتين، بئر مراد رايس، ص.ب 376 – الجزائر – محطة
الهاتف : 021.54.35.06 إلى 09
021.65.64.63
الفاكس 021.54.35.12
ج.ج.ب 50-3200 الجزائر
Télex : 65 180 IMPOF DZ
بنك الفلاحة والتّنمية الرّيفيّة 060.300.0007 68 KG
حساب العملة الأجنبيّة للمشتركين خارج الوطن
بنك الفلاحة والتّنمية الرّيفيّة 060.320.0600.12

ثمن النّسخة الأصليّة 13,50 د.ج
ثمن النّسخة الأصليّة وترجمتها 27,00 د.ج
ثمن العدد الصّادر في السّنين السّابقة : حسب التّسعيرة.
وتسلّم الفهارس مجّانا للمشتركين.
المطلوب إرفاق لفيفة إرسال الجريدة الأخيرة سواء لتجديد الاشتراكات أو للاحتجاج أو لتغيير العنوان.
ثمن النّشر على أساس 60,00 د.ج للسّطر.

Figure 19: Journal Officiel de la République Algérienne Démocratique et Populaire[joradp]

1.3. Hiérarchie des textes juridiques publiés au Journal Officiel :

Tout le système juridique repose sur le principe de la hiérarchie des normes. Pour mieux comprendre ce principe, il faut s'imaginer le système comme une pyramide à plusieurs niveaux, dont chaque niveau est une source du droit s'imposant à tous [Mahiou 1984].

1.3.1. La constitution :

Au sommet de la pyramide, on trouve la constitution. Elle représente la loi fondamentale de l'Etat qui définit les droits et les libertés des citoyens ainsi que l'organisation et les séparations du pouvoir politique (législatif, exécutif, judiciaire). Elle précise l'articulation et le fonctionnement des différentes institutions qui composent l'Etat (Conseil constitutionnel, Parlement, gouvernement, administration...).

1.3.2. Les conventions et traités internationaux :

Les traités internationaux sont des règles de droit négociées par plusieurs États dans le but de s'engager mutuellement, les uns envers les autres, dans les domaines qu'ils définissent (défense, commerce, justice...).

1.3.3. Les textes législatifs :

Les **lois** votées par l'Assemblée Populaire Nationale et le Sénat ainsi que les **ordonnances** du Président de la République entre les périodes législatives.

1.3.4. Les textes exécutifs :

Sont des textes issus du pouvoir exécutif présenté par le président de la république (**décret présidentiel**) et le premier ministre, ex chef de gouvernement (**décret exécutif**) ; ainsi que les **arrêtés** ministériels concernant un secteur signé par le ministre et les **arrêtés** interministériels commun à plusieurs secteurs signé par les ministres concernés [Mahiou 1984].

1.3.5. Les textes des autres autorités :

Les textes du Conseil Constitutionnel, de la Cours du Compte, du Haut Conseil d'Etat, des Wali,… en forme de **déclaration, résolution, proclamation, rapport, arrêté de wali**,…

1.4. Les caractéristiques du Journal Officiel :

Le Journal Officiel est une publication gouvernementale quotidienne destinée à assurer la publication des lois, des ordonnances, des décrets et des comptes rendus des débats de l'assemblée nationale et du sénat.

Il est simultanément édité sur papier et sous forme électronique sous format PDF en mode texte pour les années récentes et en mode image pour les années antérieures à 2001.

Le Secrétariat Général du Gouvernement (SGG) algérien publie en ligne l'intégralité du Journal Officiel de la République Algérienne Démocratique et Populaire (الجريدة الرسمية للجمهورية الجزائرية الديمقراطية الشعبية) depuis sa proclamation en 1962 ainsi que le texte de la constitution algérienne dans leurs versions françaises et arabes.

Dans notre modeste travail nous nous intéressons à la version écrite en langue arabe.

1.4.1. Structure du Journal Officiel :

Le Journal Officiel est un recueil de textes législatifs, exécutifs et réglementaires, classés selon des règles particulières, afin d'autoriser un accès logique.

Le Journal Officiel est composé d'un ensemble d'éléments, repartis sur plusieurs pages numérotées. Il s'agit de la première page (page de titre), d'un sommaire et du corps du journal qui contient les textes juridiques publiés dans ce journal.

1.4.1.1. La page de titre :

La page de titre vise à identifier le journal. Elle comporte les éléments suivants :

- Le numéro du journal ;
- Date de publication du journal, en grégorien et hégirien;
- Nombres d'années de publication ;
- Titre, sigle et d'autres informations ;
- La page de titre non numérotée, mais comptée dans la pagination de l'ouvrage.

1.4.1.2. Le sommaire :

Le sommaire a pour objet de donner une vue d'ensemble de la structure du journal, il est placé après la page de titre. Il comporte tous les titres des textes publiés dans ce journal. Chaque texte est représenté par sa nature, son code pour la plus part, sa date en grégorien et hégirien, l'objet (à l'exception des décisions du conseil constitutionnel), ainsi que son numéro de page.

Le sommaire est divisé en parties, selon la nature des textes et il est compté dans la pagination.

21 صفر عام 1433 هـ
15 يناير سنة 2012 م

الجريدة الرّسميّة للجمهوريّة الجزائريّة / العدد 02

3

فهرس (تابع)

مراسيم فردية

Figure 20: Sommaire du Journal Officiel [joradp]

1.4.1.3. Le corps du journal :

Le corps du journal suit le sommaire. Il est rédigé sur plusieurs pages qui contiennent un entête.

On trouve au niveau de l'entête de la page :

- Le numéro de la page,
- La date de publication en deux formats,
- Le numéro du journal.

Les textes sont classés par nature, c'est-à-dire arrêté, publication ou autres.

Chaque texte, selon sa classe, contient :

1.4.1.3.1. Bloc de titre :

Le texte est identifié par son titre, qui comprend les éléments suivants :

- Nature du texte : Par exemple loi, ordonnance, décret, arrêté,…
- Numéro : Le numéro est indispensable dans certains textes publiés au Journal Officiel ;
- Date hégirienne et grégorienne : c'est la date de signature de ce texte ;
- Objet : c'est le sujet du texte. Généralement il commence par les termes يتضمن, يتعلق, يعدل, .. . Certains textes ne contiennent pas l'objet comme les décisions du conseil constitutionnel,….

1.4.1.3.2. Bloc des visas :

Contient les textes publiés précédemment et qui sont en liaison avec le sujet du texte en objet. Les visas cités sont classés par ordre hiérarchique.

1.4.1.3.3. Bloc corps du texte :

C'est le noyau du texte constitué de paragraphes et/ou articles numérotés.

1.4.1.3.4. Bloc Signature :

On y trouve l'autorité et le nom de la personne signataire.

Certains textes du Journal Officiel contiennent des annexes, dans la plus part des temps de type technique, des listes, des tableaux, des résultats, ou même des schémas.

6 الجريدة الرّسميّة للجمهوريّة الجزائريّة / العدد 02 — 21 صفر عام 1433 هـ / 15 يناير سنة 2012 م

– واعتبارا أن المادة 120 (الفقرات الأولى و2 و3) تعتبر ركنا أساسيا في إجراءات إصدار أي قانون، وبالتالي فهي سند دستوري لهذا القانون العضوي، موضوع الإخطار،

– واعتبارا بالنتيجة أن عدم إدراجها ضمن التأشيرات يعدّ سهوا يتعيّن تداركه.

3 – فيما يخصّ عدم الاستناد إلى المادة 126 من الدستور ضمن تأشيرات القانون العضوي، موضوع الإخطار،

– اعتبارا أن المادة 126 من الدستور تنصّ على ما يأتي : "يصدر رئيس الجمهورية القانون في أجل ثلاثين (30) يوما ابتداء من تاريخ تسلمه إياه.

غير أنه إذا أخطرت سلطة من السلطات المنصوص عليها في المادة 166، المجلس الدستوري، قبل صدور القانون يوقف هذا الأجل حتى يفصل في ذلك المجلس الدستوري وفق الشروط التي تحددها المادة 167 من الدستور".

– واعتبارا أن المادة 126 تعتبر أساسية في إصدار أي قانون، وبالتالي فهي سند دستوري لهذا القانون العضوي موضوع الإخطار،

– واعتبارا بالنتيجة أن إغفال المشرّع للإشارة إلى المادة 126 من الدستور ضمن تأشيرات هذا القانون العضوي، يعدّ سهوا يتعين تداركه.

4 – فيما يخصّ عدم تحديد الفقرة 2 في المادة 165 من الدستور :

– اعتبارا أنه بموجب الفقرة 2 من المادة 165 من الدستور يبدي المجلس الدستوري بعد أن يخطره رئيس الجمهورية رأيه وجوبا في دستورية القوانين العضوية بعد أن يصادق عليها البرلمان،

– واعتبارا أن المشرّع أشار ضمن تأشيرات القانون العضوي إلى المادة 165 من الدستور، لكنه لم يحدد الفقرة 2 منها وهي الفقرة الخاصة بالقوانين العضوية،

– واعتبارا بالنتيجة أن عدم تحديد الفقرة 2 عند إدراج المادة 165 ضمن تأشيرات القانون العضوي، موضوع الإخطار، يعدّ سهوا يتعيّن تداركه.

5 – فيما يخص الاستناد إلى المادتين 179 و180 من الدستور ضمن تأشيرات القانون العضوي، موضوع الإخطار، مأخوذتين معا لاتحادهما في العلّة :

– اعتبارا أن المادة 179 تنص على استمرار الهيئة التشريعية القائمة آنذاك حتى انتهاء مهمتها، و رئيس الجمهورية بعد انتهاء هذه المهمة، التشريع بأوامر بما في ذلك في المسائل التي أصبحت وفق دستور 1996 تدخل ضمن مجال القوانين العضوية،

– واعتبارا أن المادة 180 تنصّ على أنه حتى تنصيب المؤسسات التي نصّ عليها دستور 1996 يستمر سريان مفعول القوانين المتعلقة بمجال القوانين العضوية إلى أن تُعدّل أو تُستبدل وفق الإجراءات التي نصّ عليها الدستور، واستمرار المجلس الدستوري في ممارسة صلاحياته بتمثيله الذي كان عليه حتى تنصيب المؤسسات الممثلة فيه، واستمرار المجلس الشعبي الوطني في ممارسة السلطة التشريعية كاملة حتى تنصيب مجلس الأمة،

– واعتبارا بالنتيجة أن المادتين تتضمنان أحكاما انتقالية حققت الأهداف التي وضعها المؤسّس الدستوري من أجلها، مما يجعل هاتين المادتين لا علاقة لهما بالقانون العضوي، موضوع الإخطار.

6 – فيما يخصّ عدم الاستناد إلى ميثاق السلم والمصالحة الوطنية ضمن تأشيرات القانون العضوي، موضوع الإخطار :

– اعتبارا أن ميثاق السلم والمصالحة الوطنية، حدّد المبادئ و التدابير التي قامت عليها المصالحة الوطنية، وفوّض رئيس الجمهورية باتخاذ جميع التدابير قصد تجسيد ما جاء في بنوده،

– واعتبارا أن المشرّع أدرج ضمن تأشيرات القانون العضوي، موضوع الإخطار، الأمر الذي يحدّد إجراءات تنفيذ ميثاق السلم والمصالحة الوطنية، دون الإشارة إلى الميثاق الذي يشكل الأساس القانوني لهذا الأمر،

– واعتبارا أن ميثاق السلم والمصالحة الوطنية تمت تزكيته في استفتاء شعبي، وبعدّ التعبير المباشر عن الإرادة السيّدة للشعب، ومن ثم فإنه يحتل في تدرّج القواعد القانونية مرتبة أسمى من القوانين العضوية منها أو العادية، بالنظر إلى اختلاف إجراءات الإعداد والمصادقة والرقابة الدستورية،

– واعتبارا بالنتيجة أن عدم إدراج ميثاق السلم والمصالحة الوطنية ضمن التأشيرات يعدّ سهوا يتعيّن تداركه، بترتيب هذا النص مباشرة بعد مواد الدستور.

Figure 21: Textes du Journal Officiel [joradp]

1.4.2. Vocabulaire du Journal Officiel :

Le Journal Officiel est écrit en langage juridique, qui est un langage professionnel dont la singularité a éveillé ces dernières années un grand intérêt dans des disciplines comme la linguistique. Deux motifs expliquent cet intérêt : le premier est l'importance du langage dans la plus grande partie des processus juridiques (interprétation, application, …etc.) ; le second est la formalité de son registre découlant des caractères morphologiques, syntaxiques, lexicaux et pragmatiques.

La langue juridique est l'une des langues de spécialité les plus complexes. Ses termes sont l'ensemble des mots qui contiennent au moins un sens lié à un concept appartenant au système juridique. On peut remarquer qu'il existe deux vocabulaires juridiques : les termes exclusivement juridiques et les termes à double appartenance.

1.4.2.1. Concepts exclusivement juridique :

Les termes exclusivement juridiques n'ont d'emploi que dans le langage de droit. C'est un ensemble défini de mots réservés aux juristes comme : دستور، مرسوم، اتفاقية،..... Ils désignent avec précision leurs référents du fait qu'ils sont monosémiques. La technicité de la forme linguistique et la juridicité du référent empêchent que ces termes engendrent une dérivation lexicologique.

1.4.2.2. Concepts à double appartenance :

Ces concepts appartiennent au langage courant mais souvent ils acquièrent un sens différent dans le langage du droit. Dans cet ensemble, on constate qu'il y a des termes qui ont un sens juridique principal et un sens dérivé dans le langage courant comme : قانون, عدالة, قضاء, ... D'autres termes ont leur sens principal dans le langage courant et un sens dérivé dans le droit commun : اتهام، فرضية، قرينة،.... Certains de ces termes ont acquis un sens plus spécifique dans le langage juridique.

1.4.2.3. Sémantique juridique :

L'étude de la sémantique juridique est pertinente car elle nous a montré deux sortes de polysémie : externe et interne. La polysémie externe concerne les mots à double appartenance. La polysémie interne désigne la multiplication du sens d'un mot suivant le sous-domaine juridique comme : سبب، فعل، حرية،.... L'étude morphologique du langage juridique est également importante afin de mettre en évidence les particularités de son fonctionnement interne.

1.5. Classification des textes du Journal Officiel :

Les textes du Journal Officiel sont classifiés selon plusieurs critères, on peut citer la classification par secteur, par nature du texte juridique, ou par ministère.

1.5.1. Classification par secteur :

Les secteurs du Journal Officiel sont cités dans le tableau qui suit :

Nom du Secteur en français	Nom du Secteur en arabe (القطاع)
Sureté nationale	الأمن الوطني
Mines	المناجم
Finances	المالية
Communication	الاتصال
Moudjahidine	المجاهدين
Réforme administrative	الإصلاح الإداري
Travaux publics	الأشغال العمومية
Transport	النقل

Economie	الاقتصاد
Information	الإعلام
Construction	البناء
Aménagement du territoire	التهيئة العمرانية
Environnement	البيئة
Equipement	التجهيز
Commerce	التجارة
Recherche scientifique	البحث العلمي
Parlement	البرلمان
Education et enseignement supérieur	التربية والتعليم العالي
Postes et télécommunications	البريد
Solidarité	التضامن
Formation professionnelle	التكوين المهني
Culture	الثقافة
Intérieur et collectivités locales	الداخلية والجماعات المحلية
Défense nationale	الدفاع الوطني
Hydraulique	الري
Affaires sociales	الشؤون الاجتماعية
Affaires étrangères	الشؤون الخارجية
Affaires religieuses	الشؤون الدينية
Industries	الصناعة
Energie	الطاقة
Jeunesse et sports	الشباب والرياضة
Tourisme	السياحة
Santé	الصحة
Pêche	الصيد
Habitat	السكن
Travail	العمل
Foret	الغابات
Justice	العدل
Agriculture	الفلاحة
Droit de l'homme	حقوق الإنسان
Présidence de la république	رئاسة الجمهورية
Chef du gouvernement	رئاسة الحكومة

Tableau 1: Table des secteurs du Journal Officiel

1.5.2. Classification par ministère :

La liste contient tous les ministères du gouvernement, ceux qui sont en cours et les anciens. Citons par exemple le ministère des finances, ministère de l'intérieur, ministère des affaires étrangères,…

1.5.3. Classification par nature juridique des textes :

Les textes du Journal Officiel sont catégorisés selon leur nature juridique, Le tableau qui suit contient ces différentes natures.

Nature en français	Nature en arabe (النوع)
Ordonnance	أمر
Circulaire	منشور
Circulaire interministérielle	منشور وزاري مشترك
Résolution	لائحة
Délibération	مداولة
Délibération du haut conseil d'Etat	مداولة مجلس أعلى للدولة
Décret	مرسوم
Décret exécutif	مرسوم تنفيذي
Décret législatif	مرسوم تشريعي
Décret présidentiel	مرسوم رئاسي
Décision	مقرر
Décision interministérielle	مقرر وزاري مشترك
Proclamation	إعلان
Règlement	نظام
Convention	اتفاقية
Déclaration	تصريح
Rapport	تقرير
Instruction	تعليمة
Instruction interministérielle	تعليمة وزارية مشتركة
Barème	جدول
Avis	رأي
Loi	قانون
Arrêté	قرار
Arrêté de wali	قرار ولائي
Arrêté interministérielle	قرار وزاري مشترك

Tableau 2: Table Rubrique nature des textes du Journal Officiel

1.6. Approches informatiques du Journal Officiel :

1.6.1. JORADP :

Le développement des systèmes d'exploitation supportant la langue arabe a contribué à l'expansion des logiciels en arabe pour répondre aux différents besoins des utilisateurs arabes.

Le Secrétariat Général du Gouvernement met à la disposition du peuple algérien un serveur WEB qui permet [JORADP] :

- De consulter la Constitution ;
- D'accéder directement aux numéros du Journal Officiel publiés ;
- De consulter les autres publications élaborées par les services du Secrétariat Général du Gouvernement.

Figure 222: Le site du Journal Officiel Algérien (joradp) sur la page d'accueil[joradp]

Le JORADP (Journal Officiel de la République Algérienne Démocratique et Populaire) est accessible au format PDF en mode texte pour les années récentes et en mode image pour les années antérieures à 2001. Vous pouvez naviguer à l'intérieur du JORADP par année que vous sélectionnez à l'aide d'un menu déroulant situé à droite de la page d'accueil du site. Un calendrier graphique vous permet alors de choisir les mois et jours de publication du JORADP que vous souhaitez consulter.

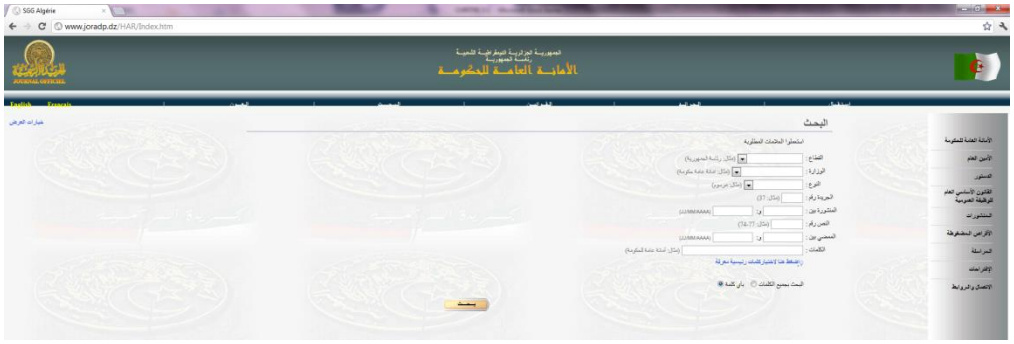


Figure 233 : Le site du Journal Officiel algérien (joradp) sur la page de Recherche des textes [joradp]

La rubrique « téléchargement » vous donne la possibilité de télécharger une version compressée au format .zip des fichiers PDF proposés à la lecture. Un moteur de recherche vous permet de faire des recherches dans le texte intégral de tout ou partie de la collection en ciblant un ministère, un secteur d'activités ou un numéro du journal. Le site du Secrétariat Général du gouvernement algérien vous donne également accès au texte intégral (format PDF) d'une série de codes juridiques [aldebaran].

Cette application est développée par le service informatique du secrétariat général du gouvernement. Ce service a pour principale mission l'informatisation des services du Secrétariat Général du Gouvernement et la mise en place d'une base de données des textes juridiques.

La base de données est utilisée aussi bien par les structures du Secrétariat Général du Gouvernement que par les administrations et institutions publiques et autres utilisateurs à travers la connexion sur le réseau Internet, sur le site WWW.JORADP.DZ

Ce service est également chargé d'éditer sur CD ROM la collection des Journaux Officiels en langues arabe et française.

Chaque CDROM contient une année de publication en langue arabe et sa traduction française sous format image PDF. Le CDROM inclut aussi un programme permettant le choix du journal et de la langue.

Figure 244: CD-ROM Journal Officiel[7]

1.6.1.1. Base de données référentielle :

SCALER est un outil simple d'utilisation, permettant une recherche multicritère bilingue (Arabe et Français) sur les textes publiés au Journal Officiel de la République Algérienne Démocratique et Populaire depuis 1962 à ce jour. Il permet l'accès par :

- Nature du texte (Loi, Ordonnance, Décret,...) ;
- Secteur ;
- Ministère initiateur ;
- Numéro du journal ;
- Date de publication ;
- Numéro du texte ;
- Date de signature du texte ;
- Et enfin combinaison de mots-clés prédéfinis qui permet d'extraire tous les textes contenant ces mots-clés.

Figure 255: L'application Scaler[8]

1.6.2. Besoin d'une ontologie de domaine pour le Journal Officiel :

Du fait que l'utilisateur qui cherche ou exploite des données ou des textes du Journal Officiel soit limité par un moteur de recherche qui compare des mots sans prendre en compte leur sémantique (sens) et exécute uniquement une recherche strictement syntaxique et donc sans réflexion car « دخل » et « أجر » représentent le même concept (la même chose), que nous appelons des classes en terminologie du Web sémantique. Plus précisément, on peut dire que

[7] : Image cd-rom scaler (Base de données).

[8] : Image cd-rom scaler (Application).

la relation de spécialisation sur les classes n'est pas gérée. Par exemple, « مؤسسة عمومية » est une spécialisation de la classe générale « هيئة عمومية ». Ainsi, pour raisonner, il ne faut plus se baser sur les mots mais sur les classes.

Ces limites incitent à développer un standard pour modéliser le contenu du Journal Officiel ; Sa manipulation du Journal Officiel souligne la nécessité d'établir des liens entre les textes constituants, la terminologie juridique et technique, Le droit qui lui-même se lie aux différentes sciences telles que l'économie, les finances, … et les sciences de la langue arabe. Ce qui ne peut être rendu possible que par l'élaboration d'un modèle standard des documents de chacune de ces sciences et de permettre la communication entre ces différents modèles.

Pour avoir une référence électronique qui facilite la réutilisation, la communication entre les différents concepts concernant le contenu du Journal Officiel, la recherche et l'indexation.

Pour cela, nous nous intéresserons à la modélisation sémantique des textes du Journal Officiel, et de ce fait nous avons choisi les ontologies et nous avons procédé à l'élaboration d'un modèle sémantique des documents du Journal Officiel avec cette technologie ; le modèle obtenu doit faciliter la recherche d'une manière sémantique dans le contenu du Journal Officiel.

1.6.3. Représentations ontologiques pour le Journal Officiel :

La sémantique peut avoir plusieurs dimensions, entre autre la dimension du sens des mots et les relations qui peuvent exister entre eux. Pour la modélisation de la sémantique, on a besoin d'informaticiens, de linguistes qui connaissent le mieux la langue du document, de cogniticiens qui sachent modéliser les connaissances et d'experts dans le domaine, en relation avec le contenu du document.

Modéliser la sémantique consiste à définir les concepts et les relations entre eux. Cela servira à faciliter la recherche non pas syntaxiquement (équivalence lettre par lettre) mais sémantiquement (équivalence du sens). Une telle approche servira aussi à faciliter la communication d'une part entre les savants des sciences juridiques, et d'autre part entre eux et les savants des autres sciences telles que les sciences de la langue arabe.

A l'heure actuelle, et à notre connaissance, il n'y a aucune modélisation sémantique des textes du Journal Officiel en Algérie.

2. Construction de l'ontologie pour le Journal Officiel :

Les besoins de réutilisation et de communication entre experts du domaine du Journal Officiel ainsi que la recherche par concept impliquent la construction d'une ontologie. Dans ce qui suit, nous exposerons les différentes possibilités d'ontologie, nous procéderons ensuite à sa conception et nous terminerons par l'évaluation et l'enrichissement de l'ontologie obtenue.

2.1. Modélisation sémantique du contenu du Journal Officiel :

Il existe des applications appropriées au Journal Officiel, mais pour concevoir une nouvelle application en relation avec le contenu du Journal Officiel, on sera contraint de repartir de zéro, car on ne pourra pas utiliser l'application déjà existante du fait qu'elle ne repose pas sur un modèle standard. Ce qui constitue une perte d'effort, de temps et de précision ; pour cela, il est nécessaire de développer un modèle standard qui favorisera la modularité, la réutilisation et le partage de données dans les systèmes informatiques et permettra de modéliser l'aspect sémantique des textes du Journal Officiel.

Il est intéressant d'élaborer un modèle à base d'ontologie qui permettra la réduction du coût et du temps nécessaires à la construction de nouvelles applications en réutilisant les connaissances déjà modélisées.

L'objectif de modéliser les textes du Journal Officiel en utilisant les ontologies est de permettre :

- Le développement d'un modèle standard complet qui favorise la modularité, la réutilisation et le partage de données dans les systèmes informatiques ;
- La communication entre les applications du Journal Officiel ;
- La recherche par concepts, ce qui facilitera grandement la tâche des utilisateurs ;
- Le développement d'applications utilisant les techniques sémantiques.

Plusieurs types d'ontologies sont possibles pour la modélisation du Journal Officiel (voir figure 27) :

Figure 266: Ontologies possibles pour le Journal Officiel

- Ontologie du domaine du Journal Officiel qui répond aux besoins du modèle standard sémantique, réutilisable et modulaire ; elle contiendra les concepts du domaine des textes et les contraintes sémantiques et servira de base à la construction d'autres ontologies en relation avec le contenu du Journal Officiel.

- Construction d'une ontologie concernant les connaissances contenues dans le Journal Officiel ; cette ontologie pourra servir à élaborer une ontologie de la langue juridique. Par exemple, nous suggérons la construction d'une ontologie des entreprises, administrations, autorités… citées dans le Journal Officiel.

- Dans le Web Sémantique, les ontologies y sont utilisées pour déterminer les index conceptuels décrivant les ressources en relation avec le contenu du Journal Officiel ou y faisant référence. Ces index à base d'ontologie seront utilisés pour la recherche par concept ce qui donne à l'utilisateur une grande liberté et facilité pour exprimer ses requêtes.

- L'interopérabilité entre applications distinctes sur le Journal Officiel sera possible grâce à une ontologie qui assure le rôle de médiateur entre ces différentes applications. Ainsi, elles peuvent s'échanger même si elles sont distantes et développées sur des bases différentes. Cela permettra aussi l'intégration de données provenant d'autres applications en relation avec le Journal Officiel.

Parmi ces possibilités d'ontologie, nous allons concevoir une ontologie du domaine pour le Journal Officiel en langue arabe.

2.2. Construction de l'ontologie pour le Journal Officiel :

Il n'existe aucune ontologie pour le Journal Officiel. Développer une ontologie à partir de zéro est coûteux et difficile ; il est avantageux d'utiliser toutes informations en relation avec le domaine à modéliser.

2.2.1. Constitution du Corpus l'ontologie :

La mise en place de la construction d'ontologies à partir de textes du Journal Officiel nécessite en premier lieu la constitution de l'ensemble des journaux officiels sur lesquels repose cette élaboration. L'ensemble des textes existent déjà sous forme papier et sous format PDF. La connaissance peut être alors capturée de ces documents préexistants qui sont déjà rassemblés.

La collection de ces textes couvre notre domaine d'intérêt.

2.2.2. Extraction de termes :

Les termes candidats pour représenter les concepts de notre ontologie ont été extraits selon deux approches qui ont été présenté dans le chapitre précédent. Syntaxique où nous avons procédé à l'extraction des termes à partir des relations grammaticales entre les mots dans les phrases des documents et qui se composent d'un ou de plusieurs mots tels que «هيئة عمومية», « المجلس الشعبي البلدي», ou selon l'approche statistique qui repose sur la fréquence d'apparition des mots dans les textes.

2.2.3. Extraction de liens de subsomption :

Pour extraire ce type de liens, nous avons suivi les deux approches expliquées précédemment, approches statistiques ou linguistiques. Nous avons regroupé et structuré les termes par rapport à leurs occurrences dans les différents textes du Journal Officiel ou par rapport à leur contexte d'apparition.

2.2.4. Détection de relations non taxonomiques :

Dans cette phase nous avons extrait les relations non taxonomiques entre les différents concepts du Journal Officiel. Où il fallait labelliser ces relations après leur extraction.

2.3. Choix du langage de description de l'ontologie :

Dans ce travail nous avons choisi OWL comme langage de description de l'ontologie du domaine des textes du Journal Officiel.
Ce choix se justifie par les points forts suivants d'OWL :

- L'utilisation d'un standard reconnu tel qu'OWL permettra d'optimiser l'interopérabilité avec d'autres systèmes. Des outils en code source libre (*Protégé*, SWOOP, etc.) peuvent être utilisés pour gérer le modèle, cela permettra de limier les efforts de développement nécessaires.

- L'utilisation d'OWL permet de partager facilement des ontologies sur le Web ; il est également possible de vérifier la cohérence des ontologies ainsi reliées, d'identifier et de résoudre des conflits éventuels et de déduire des informations nouvelles à partir des relations exprimées.

- L'utilisation d'un standard reconnu tel que le modèle OWL limitera les besoins en formation ; il suffit de se référer à la documentation OWL publique, plutôt que d'avoir à réaliser de multiples documentations nouvelles pour un système propriétaire.

- L'OWL n'a pas été conçue dans un esprit fermé permettant de borner les frontières d'une ontologie ; au contraire, la conception d'OWL a pris en compte la nature distribuée du web sémantique et de ce fait, a intégré la possibilité d'étendre des ontologies existantes, ou d'employer diverses ontologies existantes pour compléter la définition d'une nouvelle ontologie.

- L'OWL a désormais le statut de recommandation du W3C, ce qui signifie qu'il est devenu une spécification stable de grande qualité technique et qu'il est voué à un large déploiement au service de l'interopérabilité sur Internet.

2.4. Implémentation de l'ontologie obtenue :

Dans ce qui suit nous exposerons une implémentation Protégé d'une ontologie spécifiant une partie du domaine du Journal Officiel. Nous détaillons la structuration de l'ontologie en termes de classes, d'individus, et des relations faisant la liaison entre ces classes et ces individus.

En premier lieu, nous commençons par présenter la hiérarchie en termes de classes héritées de la classe racine *owl.thing* en mettant en évidence les classes primitives ou définies et les classes abstraites dont l'héritage des sous classes et possible. Ensuite nous présentons les relations entre les classes ainsi que leurs propriétés et leur dépendance. Enfin nous dénombrerons les instances des classes en termes d'individus et nous mettons en évidence l'application des relations définies entre les classes à ces instances. Parallèlement nous présentons les graphes générés pour mieux montrer les détails de l'ontologie.

2.4.1. Les Classes de l'ontologie :

Notre ontologie « ontoJO.owl » contient une sous classes principale qui hérite de la classe racine owl.thing 'الجريدة الرسمية'. Les classes الوزارة، النوع، القطاع، موضوع sont des sous classes de la classe 'الجريدة الرسمية'.

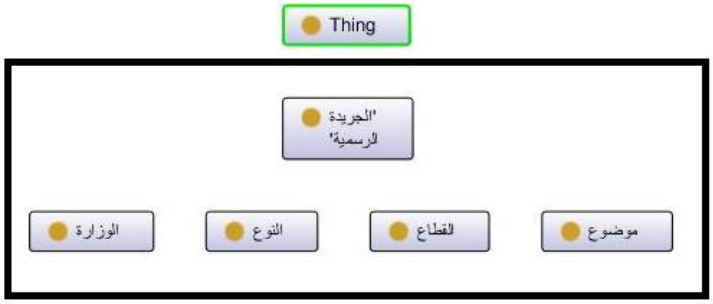


Figure 277: Structuration de base de l'ontologie ontoJO

Nous nous intéressons à la classe موضوع dans cette partie du fait que cette classe doit engendrer toutes les sous-classes qui représentent les concepts du Journal Officiel et permet d'élaborer les relations entre eux.

Dans la figure 30 nous avons illustré la hiérarchie de la classe موضوع. Nous rappelons que cette représentation n'est pas complète et que notre choix de ces sous classes est dû au fait que nous avons travaillé sur un échantillon de vocabulaire du Journal Officiel et pas sur la totalité.

Figure 288: La sous classe موضوع de l'ontologie ontoJO

Si on prend la classe هيئة par exemple, c'est une classe générique qui intègre tous les Etablissements au sens large du terme. صناعية تجارية intègre ceux qui sont de type industriel commercial, إدارية de type administratif et qui elle-même intègre deux types de هيئة إدارية, مركزية et لامركزية.

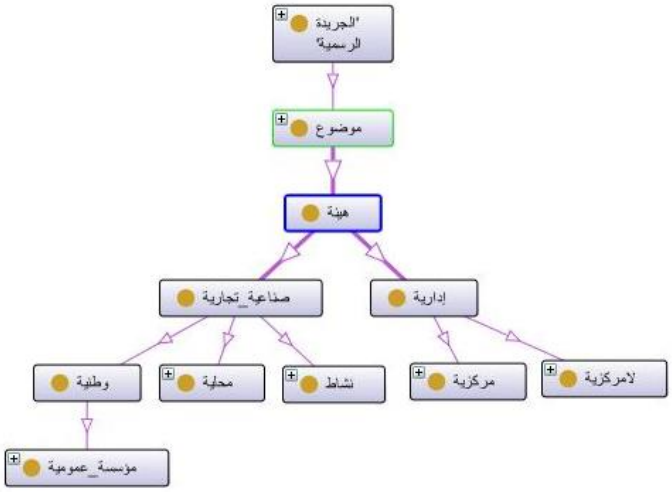


Figure 29: La sous classe هيئة de l'ontologie ontoJO

Nous avons choisi deux entreprises publics de type industriel commercial et sous la classe nationale « وطنية » pour illustrer notre exemple, Sonatrach et Entreprise Nationale de Transport Maritime de Voyageurs « المؤسسة الوطنية للنقل البحري للمسافرين ».

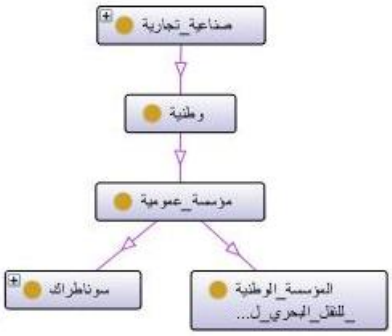


Figure 290: exemple de sous classes de l'ontologie ontoJO

Ainsi les classes obtenues sont :

N°	Nom	Type	Sous-classe de
01	Thing	class	
02	'الجريدة الرسمية'	class	Thing
03	'المجلس الشعبي البلدي'	class	لامركزية
04	'المجلس الشعبي الولائي'	class	لامركزية
05	'النقل بواسطة القنوات'	class	نشاط
06	النّقل البحري	class	نشاط
07	'طيران الطاسيلي'	class	سوناطراك
08	'وزارة الداخلية'	class	وزارة
09	'وزارة الطاقة والمناجم'	class	وزارة
10	وزارة النقل	class	وزارة
11	أحياء	class	موضوع

12	إدارية	class	هيئة
13	إصطناعي	class	صنف
14	التسويق	class	نشاط
15	القطاع	class	الجريدة الرسمية
16	المؤسسة_الوطنية_للنقل_البحري_للمسافرين	class	مؤسسة عمومية
17	المصب	class	نشاط
18	المنبع	class	نشاط
19	النوع	class	الجريدة الرسمية
20	الوزارة	class	الجريدة الرسمية
21	انسان	class	أحياء
22	بلدية	class	منطقة
23	تسمية	class	مادة
24	حالة	class	مادة
25	حيوان	class	أحياء
26	دولة	class	منطقة
27	دولي	class	مستوى
28	سائلة	class	حالة
29	سوناطراك	class	مؤسسة عمومية
30	صلبة	class	حالة
31	صناعية تجارية	class	هيئة
32	صنف	class	مادة
33	طبيعي	class	صنف
34	غازية	class	حالة
35	غالسي	class	سوناطراك
36	كوجيز	class	سوناطراك
37	لامركزية	class	إدارية
38	مؤسسة_عمومية	class	وطنية
39	ما_بين_البلديات	class	محلية
40	مادة	class	موضوع
41	محلي	class	مستوى
42	محلية	class	تجارية
43	مركزية	class	إدارية
44	مسافر	class	انسان
45	مستوى	class	مكان
46	مكان	class	موضوع
47	منطقة	class	مكان
48	موضوع	class	الجريدة الرسمية
49	ميدغاز	class	سوناطراك
50	نبات	class	أحياء
51	نشاط	class	تجارية
52	نفتيك	class	سوناطراك
53	نفطال	class	سوناطراك
54	هيئة	class	موضوع
55	وزارة	class	مركزية
56	وطني	class	مستوى
57	وطنية	class	تجارية
58	ولائية	class	محلية
59	ولاية	class	منطقة

Tableau 3 : Classes de l'ontologie de domaine du Journal Officiel

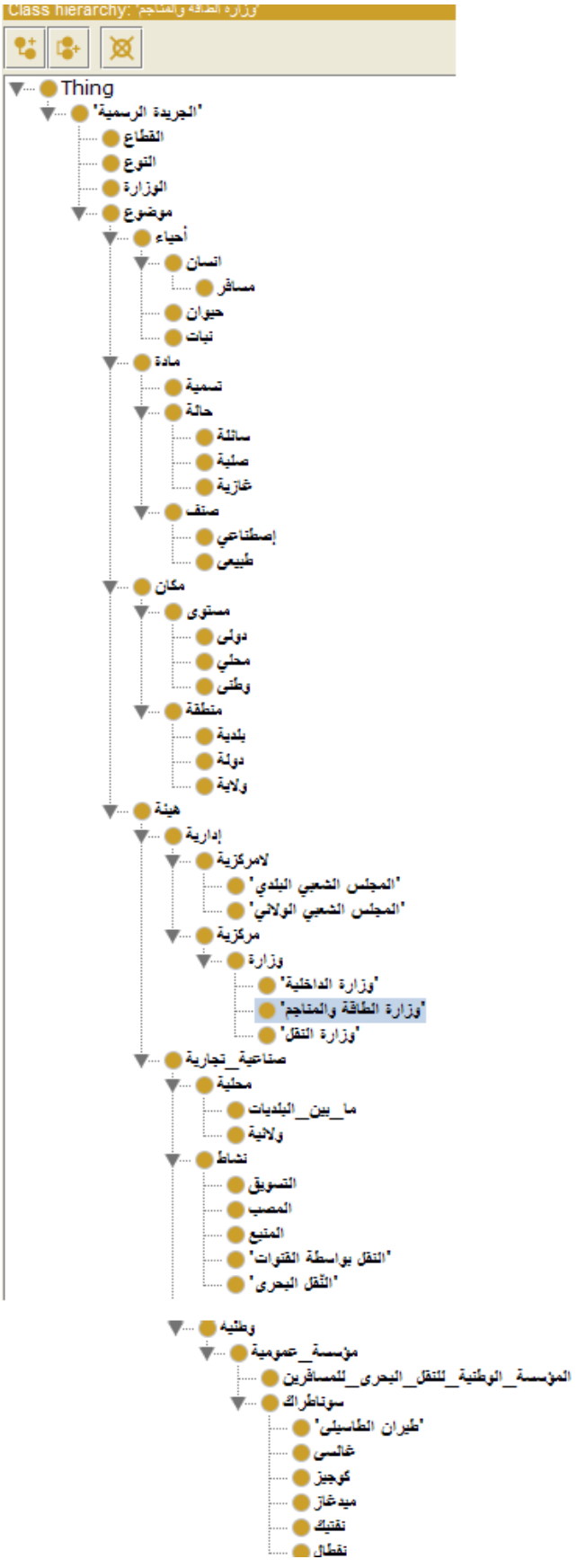


Figure 301 : La hiérarchie des classes sous Protégé de ontoJO

2.4.2. Relations de l'ontologie obtenue (Object Property):

Nous allons présenter les différents graphes relationnels entre les classes décrites dans la section précédente.

Nous commençons par la relation entre المؤسسة الوطنية للنقل البحري للمسافرين et وزارة النقل , l'Entreprise Nationale de Transport Maritime de Voyageurs est *sous tutelle* du ministère de transport ; Nous avons défini donc la propriété *تحت وصاية* ainsi que son inverse *وصية على* pour montrer que le ministère du transport et tutelle de l'Entreprise Nationale de Transport Maritime de Voyageurs.

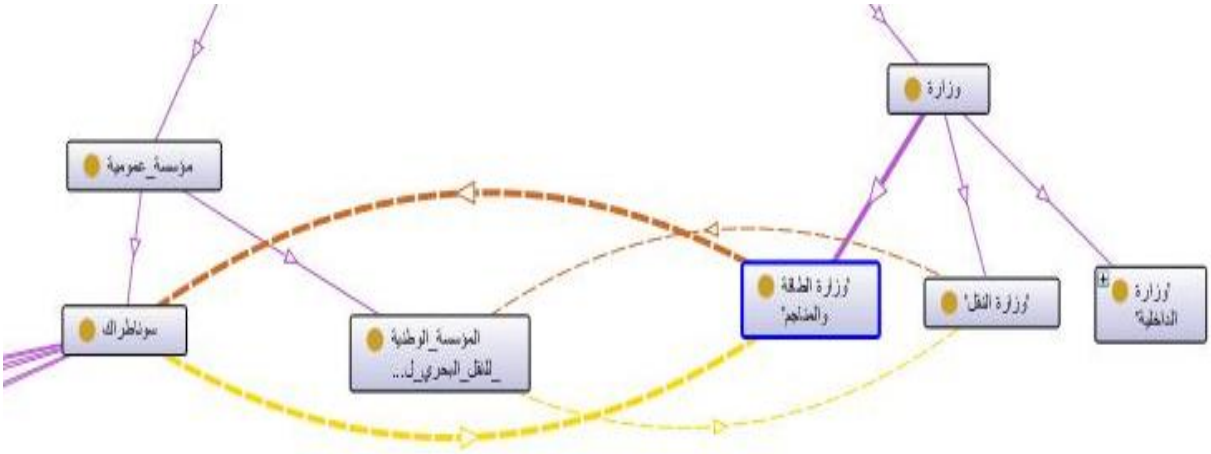


Figure 312: relations تحت وصاية et وصية على de l'ontologie ontoJO

Nous avons utilisé la même propriété pour montrer la relation qui existe entre sontrach et le ministère d'énergie et des mines, en employant la restriction **only** pour montrer que cette entreprise ne dépend que de ce ministère.

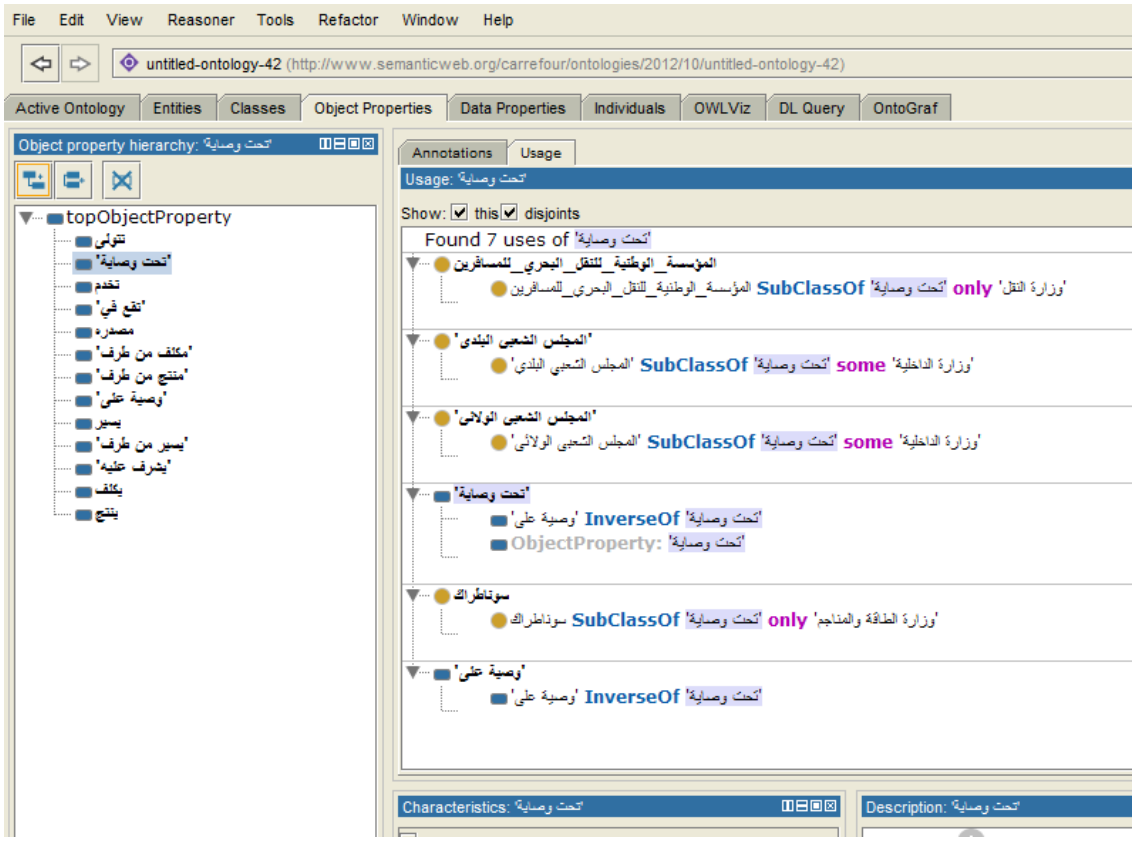


Figure 323: relations وصية على et تحت وصاية de l'ontologie ontoJO

Range (Co-domaine) : Valeurs autorisées pour la propriété, permettant de poser des contraintes sur les ressources (ou littéraux). Le Co-domaine d'une propriété définit le type des objets autorisés pour la propriété.

Domain (Domaine) : Type de ressources sur lesquels peut porter la propriété.

N°	Nom	Type	Domaine	Co-domaine
01	تحت وصاية	ObjectProperty	المؤسسة الوطنية للنقل البحري للمسافرين	وزارة النقل
			المجلس الشعبي البلدي	وزارة الداخلية
			المجلس الشعبي الولائي	وزارة الداخلية
			سوناطراك	وزارة الطاقة والمناجم
02	تقع في	ObjectProperty	المؤسسة الوطنية للنقل البحري للمسافرين	الدار البيضاء
			سوناطراك	حيدرة
03	مكلف من طرف	ObjectProperty	كوجيز	ميدغاز
04	منتج من طرف	ObjectProperty		
05	وصية على	ObjectProperty	وزارة النقل	المؤسسة الوطنية للنقل البحري للمسافرين
			وزارة الداخلية	المجلس الشعبي البلدي

				المجلس الشعبي الولائي
			وزارة الداخلية	
			وزارة الطاقة والمناجم	سوناطراك
06	يسير من طرف	ObjectProperty	بلدية	المجلس الشعبي البلدي
			ولاية	المجلس الشعبي الولائي
07	يشرف عليه	ObjectProperty		
08	تتولى	ObjectProperty	كوجيز	النّقل البحري
09	تخدم	ObjectProperty	المؤسسة الوطنية للنقل البحري للمسافرين	مسافر
10	يسير	ObjectProperty	المجلس الشعبي البلدي	بلدية
			المجلس الشعبي الولائي	ولاية
11	يكلف	ObjectProperty	ميدغاز	كوجيز
12	ينتج	ObjectProperty	ميدغاز	غاز
13	مصدره	ObjectProperty		

Tableau 4 : Relations de l'ontologie du domaine du Journal Officiel

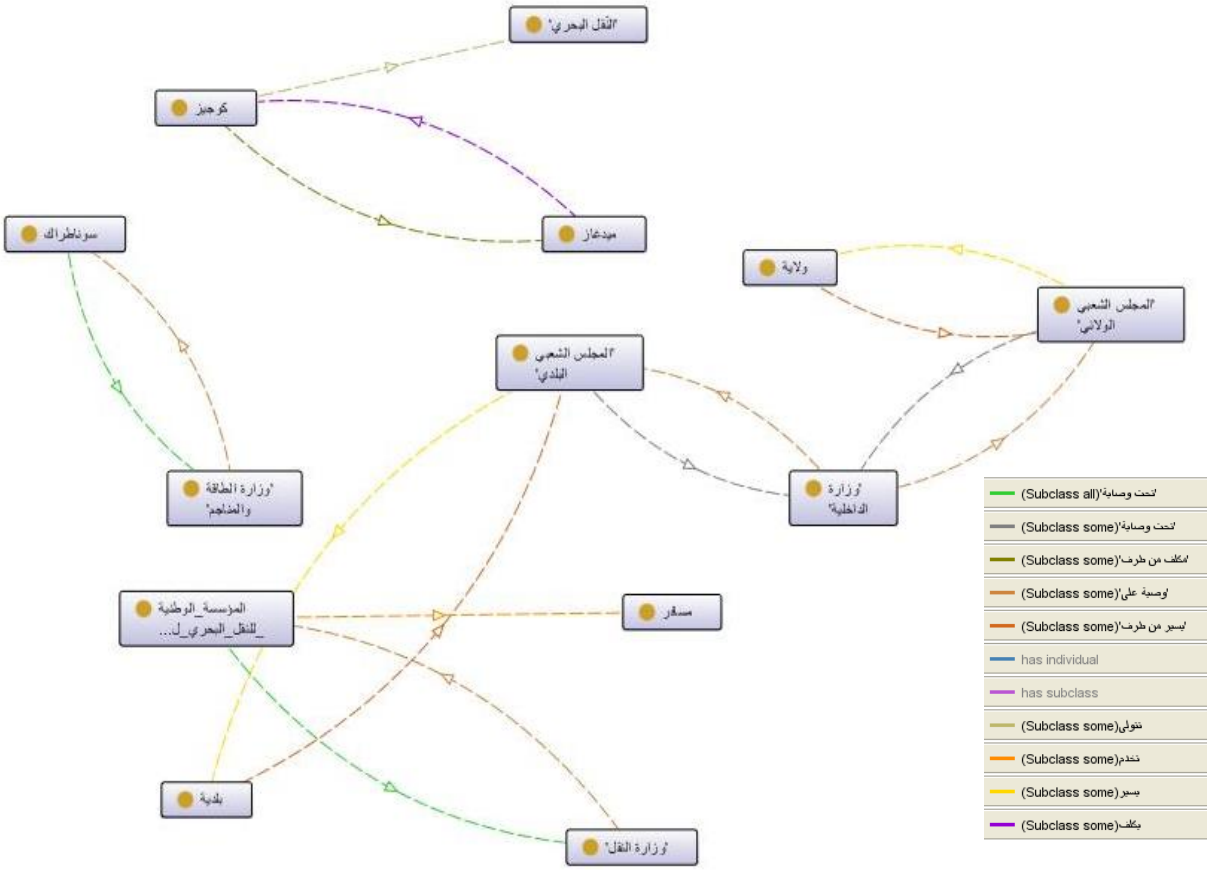


Figure 334: Graphe des relations de l'ontologie ontoJO

2.4.3. Les individus :

Nous avons introduit dans la base quelques noms de commune et de wilaya ainsi que quelques désignations de matières telles que غاز، بحر , pour illustrer la notion d'individu.

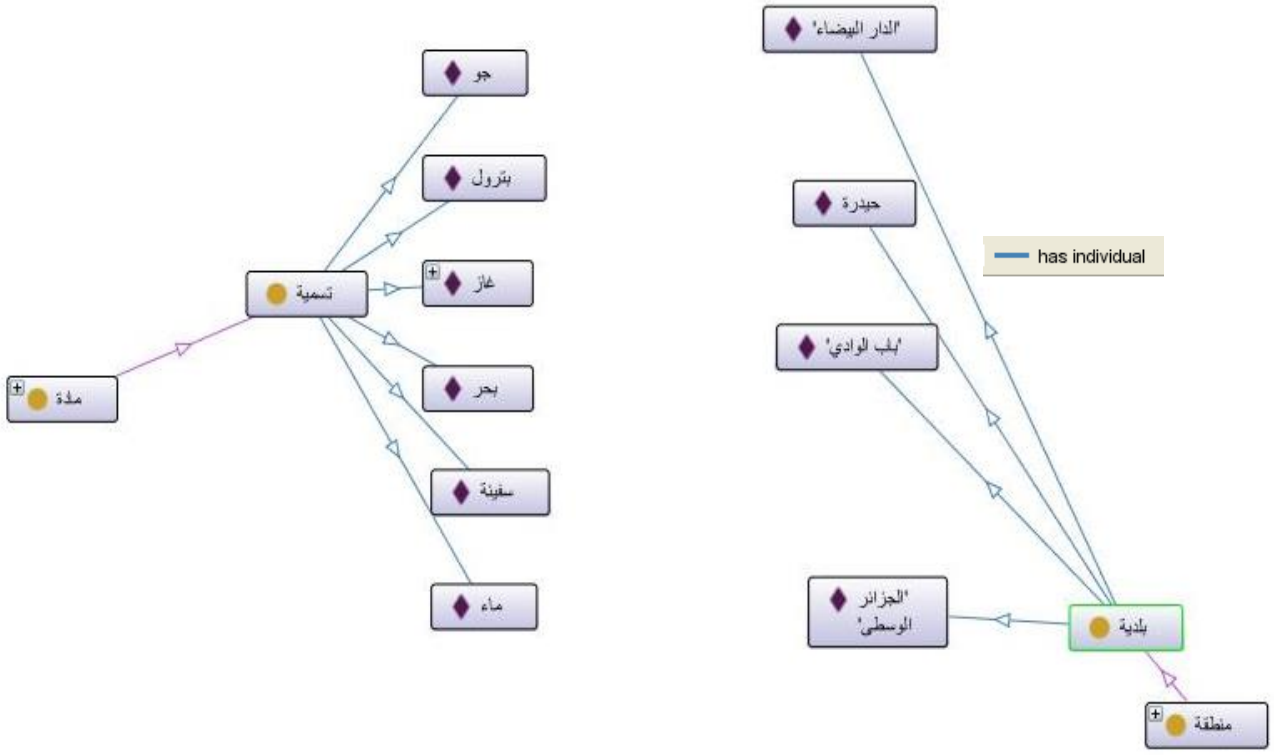


Figure 345: Les individus des classes تسمية et بلدية dans l'ontologie ontoJO

2.4.4. Schéma de l’ontologie obtenue :

L’ontologie de domaine du Journal Officiel peut être schématisée comme suit :

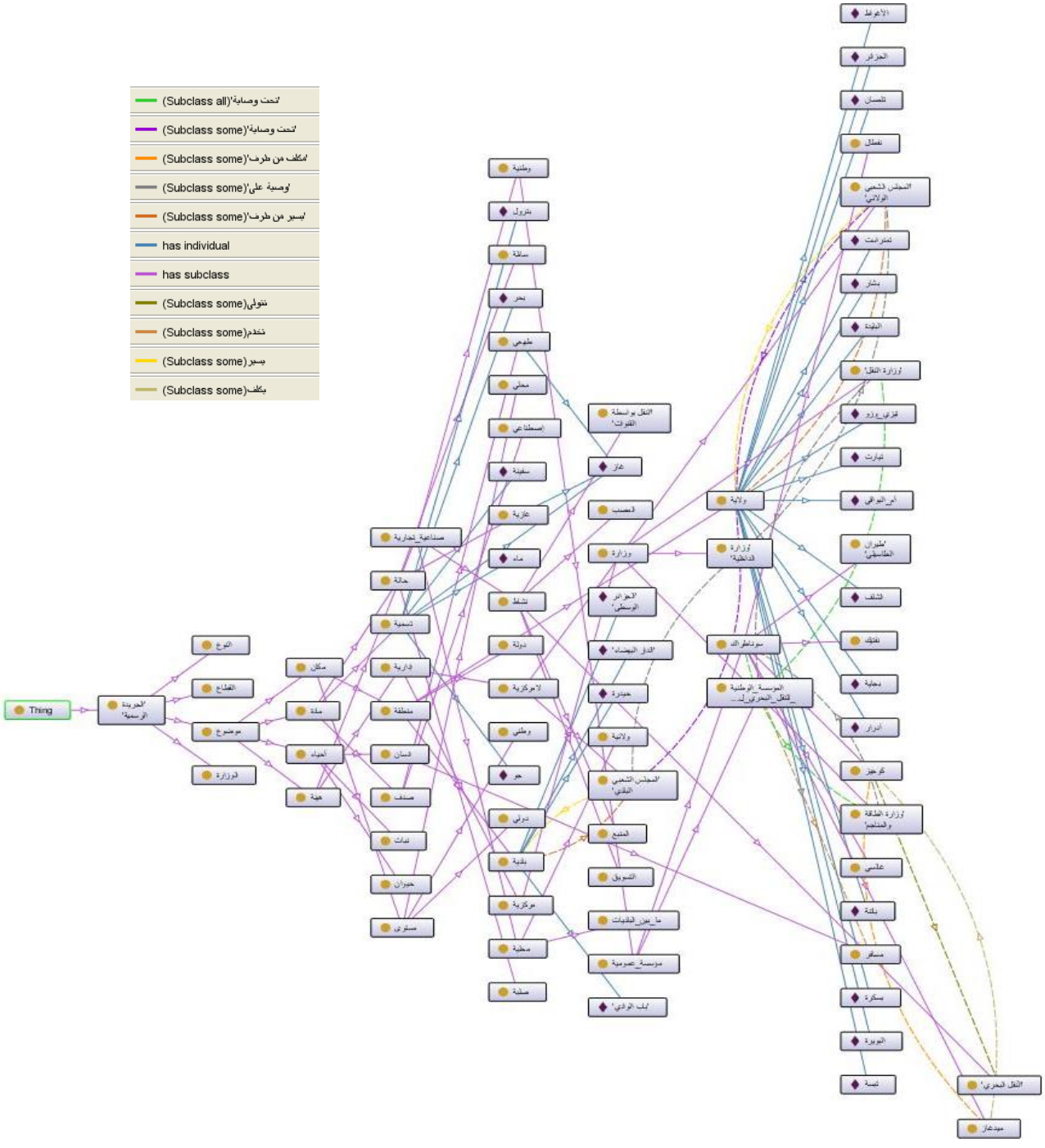


Figure 356: Schéma de l’ontologie du domaine du Journal Officiel « ontoJO »

Nous venons de présenter comment nous pouvons intégrer notre description du domaine du Journal Officiel et l'ontologie que nous avons construite dans le logiciel d'implémentation d'ontologies « Protégé ».

2.5. Evaluation et enrichissement de l'ontologie obtenue :

2.5.1. Evaluation de l'ontologie obtenue :

La question des critères pour l'évaluation en ingénierie ontologique n'est pas résolue dans le sens où elle ne fait pas encore l'objet d'un consensus au sein de la communauté.
La validation sert à s'assurer que l'ontologie modélise vraiment le domaine du Journal Officiel pour lequel le modèle a été créé.
Pour valider l'ontologie construite il faut s'assurer des principes identifiés ci-dessous :

- La cohérence ou consistance. Ce critère cherche à savoir s'il est possible d'obtenir des conclusions contradictoires à partir de définitions valides en entrée.
- La complétude. Il existe des informations qui ne sont pas prises en considération par le schéma actuel. D'où l'incomplétude de l'ontologie. L'ontologie doit être donc enrichie par la suite.

2.5.2. Enrichissement de l'ontologie obtenue :

Enrichir l'ontologie consiste à rajouter de nouvelles connaissances à celles qu'on a déjà, On peut modifier directement l'ontologie, pour préciser la définition de certains concepts, de certaines relations, ajouter/supprimer certains concepts ou certaines relations.

L'ontologie peut être enrichie ainsi :

- En utilisant la dimension multilingue, une application réalisée autour de l'ontologie du domaine du Journal Officiel en une seule langue, peut être obtenue en d'autres langues sans aucun coût supplémentaire en utilisant l'équivalent de chaque concept dans la langue voulue. Il faut prendre en compte les différentes formes linguistiques sous lesquelles peut apparaître un concept (les synonymes) en utilisant la relation « same-as » entre les concepts ;

- La construction d'une ontologie se fait en plusieurs allers-retours (cycle); nous devrons donc sonder les nouveaux besoins auxquels l'ontologie doit répondre et mettre à jour ainsi cette ontologie sera enrichie davantage.

2.6. Indexation du contenu du Journal Officiel :

Le texte du Journal Officiel est un texte spécial car la compréhension du sens des documents nécessite de fois la connaissance des sciences juridiques et de la langue arabe pour ceux qui consultent le Journal Officiel en langue arabe.

Pour indexer le contenu du Journal Officiel nous avons le choix entre l'indexation syntaxique et l'indexation par sujet.

2.6.1. Indexation syntaxique du Journal Officiel (textuelle) :

Ce type d'indexation se base sur la capture des mots et l'organisation de ces mots selon l'ordre alphabétique dans un index (la racine ou le mot en script standard peut être utilisé comme entrée à l'index); ce type d'indexation se base uniquement sur les mots sans tenir compte du sens des mots.

2.6.2. Indexation du Journal Officiel par sujet :

L'indexation par sujet appelée également indexation thématique est considérée comme une indexation sémantique du Journal Officiel, car elle se base sur la compréhension du texte à indexer et dont sont extraites les idées contenues dans chaque texte, chacune de ces idées étant par la suite exprimées en utilisant des expressions abrégées et précises composées de mots spécifiques correspondant aux sujets du contenu du Journal Officiel.

Ce type d'index nous permet de retrouver les textes en relation avec un sujet donné même si elles contiennent des mots différents du mot recherché, mais ayant le même sens que ce dernier.

Par exemple, si nous cherchions les textes en relation avec les hôpitaux (مستشفى أو مستشفيات) , cet index nous indiquera alors tous les textes concernant ce sujet y compris les textes ne contenant pas le mot hôpital (مستشفى) mais en relation avec le sujet recherché, telles que le texte suivant :

مرسوم تنفيذي رقم 163ـــ11 مؤرخ في 13 جمادى الأولى عام 1432 الموافق لـ 17 أبريل سنة2011 ، يعدل و يتمم المرسوم رقم 80ـــ59 المؤرخ في 21 ربيع الثاني عام 1400 الموافق لـ 8 مارس سنة 1980 والمتضمن احداث المراكز الطبية التربوية والمراكز المتخصصة في تعليم اِلاطفال المعوقين وتنظيمها وتسييرها

Et si on s'intéresse aux textes en relation avec le sujet sécurité nationale (الأمن الوطني), alors cet index nous indiquera les textes qui sont en relation avec ce sujet et qui contiennent les mots suivants : الشرطة , الجيش الوطني , الحماية المدنية ,…etc.

مرسوم تنفيذي رقم 11 – 163 مؤرخ في 13 جمادى الأولى عام 1432 الموافق 17 أبريل سنة 2011، يعدل ويتمم المرسوم رقم 80–59 المؤرخ في 21 ربيع الثاني عام 1400 الموافق 8 مارس سنة 1980 والمتضمن إحداث المراكز الطبية التربوية والمراكز المتخصصة في تعليم الأطفال المعوقين وتنظيمها وسيرها.

إنّ الوزير الأول،

– بناء على تقرير وزير التضامن الوطني والأسرة،

– وبناء على الدستور، لا سيّما المادتان 85–3 و125 (الفقرة 2) منه،

– وبمقتضى المرسوم رقم 80–59 المؤرخ في 21 ربيع الثاني عام 1400 الموافق 8 مارس سنة 1980 والمتضمن إحداث المراكز الطبية التربوية والمراكز المتخصصة في تعليم الأطفال المعوقين وتنظيمها وسيرها، المعدل والمتمم، لا سيما المادة 3 منه،

– وبمقتضى المرسوم الرئاسي رقم 10–149 المؤرخ في 14 جمادى الثانية عام 1431 الموافق 28 مايو سنة 2010 والمتضمن تعيين أعضاء الحكومة،

– وبعد موافقة رئيس الجمهورية،

"المادة 14 : يضم مجلس إدارة المراكز الطبية التربوية والمراكز المتخصصة في تعليم الأطفال المعوقين الذي يرأسه مدير النشاط الاجتماعي والتضامن للولاية أو ممثله :

– ممثلا عن مديرية التربية للولاية،

– ممثلا عن مديرية الصحة والسكان للولاية،

– ممثلا عن مديرية التكوين والتعليم المهنيين للولاية،

– ممثلا عن مديرية الشباب والرياضة للولاية،

– ممثلا عن مستخدمي التعليم ينتخبه نظراؤه،

– ممثلا عن مستخدمي التربية ينتخبه نظراؤه،

– ممثلا عن مستخدمي الإدارة ينتخبه نظراؤه،

– ممثلا عن جمعية أولياء التلاميذ الذين ينشطون في نفس ميدان نشاطات المؤسسة.

يمكن مجلس الإدارة أن يستعين بأي شخص من شأنه مساعدته في أشغاله.

يحضر مدير المؤسسة اجتماعات مجلس الإدارة بصوت استشاري ويتولى أمانته.

المادة 3 : تتمم قائمة المراكز المتخصصة في تعليم الأطفال المعوقين سمعيا بإحداث مدرسة (1) لصغار الصم يحدد مكان إنشائها ومقرها طبقا للجدول أدناه :

اسم المؤسسة	مكان إنشائها	
	البلدية	الولاية

Figure 37 : Exemple d'indexation par sujet (sémantique) [joradp]

Les granules sont indexés par des concepts qui reflètent leur sens plutôt que par des mots bien souvent ambigus. Il est nécessaire de retrouver dans l'ontologie les concepts présents dans la collection pour indexer les documents à partir de toutes les thématiques abordées.

Les ontologies de domaine peuvent par leur formalisation représenter des ressources impliquant un engagement sémantique plus fort. Nous entendons donc par indexation sémantique, l'indexation de granules documentaires. L'indexation sémantique se fait en deux étapes. La première étape consiste à identifier les concepts ou instances de l'ontologie dans les granules. La deuxième pondère les concepts pour chaque document en fonction de la structure conceptuelle dont ils sont issus.

L'indexation sémantique s'inscrit également dans la démarche orientée Web Sémantique. Les précurseurs de cette nouvelle version du Web considèrent que les ressources participant au Web Sémantique seront toutes reliées entre elles par des relations sémantiques. Plus précisément, les données présentes sur le Web Sémantique seront modélisées sous forme d'ontologies où chaque ressource apparaît comme un élément de ces ontologies au même titre que la connaissance qui les décrit. L'objectif est donc d'ajouter au contenu du Web une structure formelle et de la sémantique (à travers des métadonnées et de la connaissance) dans le but de permettre une meilleure gestion et un meilleur accès aux informations. L'ontologie peut être vue comme une représentation des métadonnées explicitement ou implicitement présentes dans les textes.

La phase d'indexation a pour but aussi de représenter les informations relatives à la date de création, de publication, numéro du journal, sa taille,…. Les métadonnées présentes dans les documents (auteurs, date de production), les index (les descripteurs du contenu du document), l'identifiant du document par le système (emplacement) et une vue sur le contenu (résumé ou extraits). Un enjeu actuel du Web Sémantique est de définir des techniques permettant de les extraire. La démarche orientée Web Sémantique a donc un double objectif : indexer le contenu des textes à partir des ressources permettant d'en extraire les concepts et instances mais aussi représenter les ressources en générant les métadonnées correspondantes.

Cette approche vise ainsi à représenter l'ensemble des métadonnées qui peuvent être associées aux textes du Journal Officiel. Cette ontologie est formelle et permet de mettre en place des inférences à partir de leurs axiomes. Elle contient des éléments décrivant les ministères, les publications ainsi que la nature de l'acquisition de connaissance et les domaines connexes.

2.6.2.1. Identification des concepts et des instances existant dans l'ontologie :

La première étape de l'indexation conceptuelle consiste à identifier les concepts et/ou instances de l'ontologie apparaissant dans les granules.

Une approche consiste à identifier ces éléments de l'ontologie manuellement dans les textes du Journal officiel. Cette approche a pour intérêt d'être fiable pour l'interprétation de la sémantique associée aux concepts dans l'ontologie et choisit le concept représentant au mieux la notion abordée.

2.6.2.2. Extraction des termes du granule :

L'approche suivie consiste à extraire des textes du Journal Officiel l'ensemble des termes y apparaissant et d'y rechercher les labels contenus dans l'ontologie. Les expressions sont extraites soit statistiquement, soit syntaxiquement. L'extraction d'expressions est quasiment obligatoire car les labels des concepts sont souvent composés de ce type d'éléments.

2.6.2.3. Recherche des labels correspondant à des concepts ou instances de l'ontologie :

Les labels sont recherchés dans l'ensemble des termes extraits en favorisant la prise en compte des labels les plus longs et donc des concepts les plus spécifiques. Par exemple, dans le cas où les labels « مؤسسة », « عمومية », et « مؤسسة عمومية » apparaissent dans le document, le label retenu - et donc le concept correspondant - sera مؤسسة عمومية car l'expression formée de deux termes est plus précise que le ou les termes seuls.

2.6.2.4. Désambiguïsation des labels :

Les labels peuvent se rapporter à plusieurs concepts. Afin d'identifier quel est le concept abordé dans le texte, on utilise la stratégie du « tout » correspond au cas dans lequel tous les concepts sont considérés. La stratégie du « premier » consiste à restituer le concept le plus fréquent dans le document ou bien dans la collection. La stratégie du « contexte » base la désambiguïsation sur la proximité sémantique des concepts candidats et du contexte dans lequel ils apparaissent dans les documents.

2.6.2.5. Extraction de nouvelles instances :

L'extraction d'instances de concepts a pour but d'extraire les métadonnées qui permettront de représenter les ressources dans le cadre du Web Sémantique. L'extraction d'instances repose sur des techniques du domaine de l'extraction d'information.

L'extraction d'instances de concepts peut se faire à partir de techniques d'extraction d'entités nommées. Une entité nommée est un nom ou syntagme nominal se rapportant à une entité comme, par exemple, une personne, une organisation ou une localisation. Les entités sont extraites à partir d'une base de connaissance qui, à partir de ressources lexicales, permet la détection automatique des entités. Les ressources lexicales décrivent par exemple les formes pouvant permettre la détection de noms d'entreprises ou de noms de villes ou de personnes. La base de connaissances contient un ensemble d'instances prédéfinies et décrites à partir d'axiomes. Un mécanisme d'inférence définit des règles permettant d'extraire de nouvelles instances.

2.7. Accès aux Textes du Journal Officiel à partir de l'ontologie :

La plupart des SRI fonctionnent avec une interface qui permet à l'utilisateur de formuler son besoin en informations à partir d'une requête. Le système lui présente le résultat de sa recherche sous forme d'une liste de références vers les textes retrouvés. Une alternative au principe de recherche d'information consiste à fournir des outils qui permettent à l'utilisateur d'explorer la collection de documents pour trouver les documents pertinents sans avoir à exprimer son besoin en informations sous forme d'une requête conventionnelle. L'utilisateur a en effet du mal à spécifier son besoin et à l'exprimer, surtout s'il ne connaît pas le contenu de la collection à sa disposition.

2.7.1. Langage d'interrogation, requête et appariement :

Afin de communiquer son besoin au système, l'utilisateur doit le formuler dans un langage interprétable par le système.

2.7.1.1. Interrogation en langage libre :

Dans le cadre de la recherche d'information, ce besoin est formulé sous forme de requêtes. La formulation de la requête est un problème crucial car de sa qualité dépend la qualité des documents restitués. Le format de la requête dépend du SRI. Les requêtes booléennes sont composées de termes et d'opérateurs booléens (ET, OU, SAUF).

Un autre type de requête consiste à formuler les requêtes en langage libre. Aucune syntaxe particulière n'est alors définie. L'appellation langage libre est préférée à langage naturel car généralement, les requêtes formulées par l'utilisateur ne constituent pas des phrases grammaticales correctes mais des listes de mots ou d'expressions.

Le modèle booléen est basé sur l'algèbre de Boole et repose sur une représentation booléenne des requêtes. Dans ce modèle, les documents restitués à l'utilisateur sont ceux contenant exactement les termes de la requête. Il repose donc sur l'absence ou la présence des termes retenus pour indexer les documents et les termes de la requête. Dans l'ensemble de ces modèles, les documents sont restitués à l'utilisateur par ordre de pertinence supposée décroissante.

L'utilisation d'ontologies dans les SRI permet de définir d'autres types d'interrogation qui s'appuient sur les langages du Web Sémantique.

2.7.1.2. Appariement à partir d'ontologies :

Les ontologies peuvent servir à calculer la similarité entre la représentation de la requête et la représentation des textes dans le cas où les deux représentations sont faites à partir des concepts d'une même ontologie.

Les documents et requêtes sont représentés à partir du langage et de l'ontologie. Cette ontologie contient l'ensemble de concepts et de relations entre concepts, dont la relation de subsomption. Elle est considérée comme un graphe orienté. L'avantage du calcul de la similarité est de classer les documents restitués par rapport à leur similarité à la requête, cette similarité reposant sur l'organisation des concepts dans l'ontologie.

2.7.1.3. Reformulation de requête à partir des termes de l'ontologie :

L'objectif de la reformulation est soit de limiter le silence (le silence fait référence aux documents pertinents mais qui ne sont pas retrouvés par le système) soit de réduire les risques de bruit (le bruit fait référence aux documents non pertinents retrouvés par le système). Dans le premier cas, la requête est étendue à partir de termes similaires à ceux de la requête initiale. Dans le second cas la requête initiale est étendue ou modifiée à partir de termes qui ajoutent de l'information complémentaire à la représentation du besoin. Il existe une approche qui permet l'expansion des requêtes. Elle consiste à utiliser des ressources, comme par exemple un dictionnaire, en étendant les requêtes à partir de nouveaux termes en relation avec les termes de la requête.

2.7.1.4. Exploration à partir de hiérarchie de concepts :

La catégorisation suivant une hiérarchie de concepts vise à aider l'utilisateur dans la spécification de son besoin en lui donnant une vue d'ensemble sur la collection, puis en lui permettant de spécifier les vues en fonction des informations qui l'intéressent.

3. Conclusion

Nous avons procédé à la construction d'une ontologie de domaine pour le Journal Officiel en utilisant le langage OWL. Ce choix est justifié par les avantages très intéressants de ce langage. Et l'analyse menée par protégé nous a conduit à admettre qu'il n y a pas de classification typique pour chaque domaine. L'ajout de paramètres est possible et très bien supporté. Aussi les ontologies ne peuvent être pensées comme une conceptualisation finie d'un domaine de connaissances délimité et stable.

Conclusion générale

Notre contribution consista en la construction d'une ontologie de domaine pour le Journal Officiel en langue arabe. Pour ce faire, nous avons eu recours à un processus basé sur certains travaux intéressants, trouvés dans la littérature.

Cette recherche consacrée à l'élaboration de cette ontologie de domaine, nous a permis de mesurer l'importance de la connaissance dans l'ingénierie linguistique, quelques soit les voies qu'elle prenne.

Il est nécessaire de pouvoir disposer de systèmes de connaissances sous forme informatique et manipulables dans des ontologies utilisant des outils de raisonnement et de calcul. La nécessité s'impose d'extraire, de structurer et de pouvoir réutiliser les connaissances. Les raisons majeures qui ont poussé la recherche sur les ontologies ces dernières années étaient de permettre la réutilisation du savoir sur un domaine. En effet, lorsqu'un groupe de chercheurs développe une telle ontologie en détail, les autres groupes peuvent simplement la réutiliser pour leurs propres domaines d'ontologie développée, et si besoin, construire une ontologie plus large. Il serait possible d'intégrer plusieurs ontologies existantes décrivant des portions d'un domaine. C'est tout le problème de la réutilisabilité.

En ce qui concerne notre étude, nous réalisons après avoir donné tous les résultats de notre analyse que les spécifications explicites du savoir dans le domaine du Journal Officiel sont utiles pour les chercheurs qui s'intéressent à compléter et enrichir cette ontologie.

Nous reconnaissons l'existence d'une difficulté. Il nous est arrivé recours à des notions plus ou moins liées aux sciences juridiques. Cette tache a nécessité un élargissement du champ d'étude et à faire appel à d'autres domaines. Il y a là une extension dans notre représentation du domaine.

L'analyse menée par protégé nous a conduit à admettre qu'il n y a pas de classification typique pour tel ou tel domaine ou sous domaine. Le noyau reste stable, et l'ajout de paramètres liés à des domaines est possible et très bien supporté par OWL.

Un autre point sur lequel insister est l'aspect multilingue, une ontologie est nécessaire pour le chercheur pour la compréhension des concepts et leur représentation dans un langage formel. L'intérêt de cette application, est que bien que la conception et la réalisation sont faites en langue arabe, dans le système de la traduction automatique elle permet de produire à partir de ces représentations, des concepts ou des textes dans une ou plusieurs langues telles que le français ou l'anglais.

Les perspectives de ce travail sont à la fois nombreuses et prometteuses. Tout d'abord étendre le contenu de l'ontologie en ajoutant de nouveaux concepts et relations, qui sont liés aux textes du Journal Officiel. Nous pouvons par la suite travailler sur un domaine plus élargi qui est le domaine du langage juridique en langue arabe.

Une autre perspective qui, s'offre à nous, concerne la dimension multilingue, en utilisant l'équivalent de chaque concept dans la langue voulue. Il faut prendre en compte les différentes formes linguistiques sous lesquelles peut apparaître un concept (les synonymes) en utilisant la relation « same-as » entre les concepts. Le but aussi d'acquérir des expériences récentes menées pour les différentes langues et les investir dans une conception des ressources linguistiques et terminologique en langue arabe.

Bibliographie

- [Abduljaleel 2003]Abduljaleel N., Larkey L., Statistical transliteration for English-Arabic Cross Language Information Retrieval. In Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management, New Orleans, Louisiana, 2003, disponible sur le lien: http://pdf.aminer.org/000/095/075/statistical_transliteration_for_english_arabic_cross_language_information_retrieval.pdf
- [Agirre 2000]E. Agirre, O. Ansa, E. Hovy, D. Martinez, Enriching very large ontologies using the WWW, In Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI-00), 2000.
- [aldebaran]http://aldebaran.revues.org/1592
- [Alfonseca 2002]E. Alfonseca, S. Manandhar, Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures, EKAW-2002, Lecture Notes in Artificial Intelligence 2473, Springer Verlag, 2002.
- [Aliane 2010]Hassina Aliane, Zaia Alimazighi, Al –Khalil: The Arabic Linguistic Ontology Project, Semantic web and Arabic Language Team, Research Center on Scientific and technical Information, Alger, 2010.
- [Al-Khalifa 2009]H.S. Al-Khalifa, M.M. Al-Yahya, A. Bahanshal, I Al-Odah, SemQ: A Proposed Framework for Representing Semantic Opposition in the Holy Quran using Semantic Web Technologies, CTIT, 2009.
- [Allan 2003a]J. Allan (Ed.), Challenges in information retrieval and language modeling, SIGIR Forum, 37(1), 2003.
- [Allan 2003b]J. Allan, Hard Track Overview in TREC 2003: High Accuracy Retrieval from Documents, Text Retrieval Conference, 2003, disponible sur le lien: http://trec.nist.gov/pubs/trec12/papers/HARD.OVERVIEW.pdf.
- [Andreasen 2003]T. Andreasen, H. Bulskov, R. Knappe, Similarity for Conceptual Querying, In Proceedings for the 18th International Symposium on Computer and Information Sciences, 2003.
- [Assadi 1999]H Assadi, Construction of a regional ontology from text and its use within a documentary system. In Proceedings of the 2nd Formal Ontology in Information Systems Conference, N. Guarino (Ed.), 1999.
- [ASSTICCOT 2003]Rapport de l'action spécifique ASSTICCOT, Action Spécifique STIC «Corpus et Terminologie» (AS 34), Rattachée au RTP-DOC (RTP 33), Rapport internet IRIT/2003-23-R, 2003.
- [Aussenac-Gilles 2000a] N. Aussenac-Gilles, B. Biébow, S. Szulman, Modélisation du domaine par une méthode fondée sur l'analyse de corpus, In Actes de la Conférence en Ingénierie des Connaissances (IC'2000), 2000.
- [Aussenac-Gilles 2000b]N. Aussenac-Gilles, B. Biébow, S. Szulman, Revisiting ontology design - a method based on corpus analysis, In Proceedings of the 12th European Knowledge Acquisition Workshop (EKAW'00), R Dieng, O. Corby (Eds.), 2000.
- [Aussenac-Gilles 2004]N. Aussenac-Gilles, J. Mothe, Ontologies as Background Knowledge to Explore Document Collections, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), 2004.
- [Baader 1991]F. Baader, B. Hollunder, A Terminological Knowledge Representation System with Complete Inference Algorithms, In Proceedings of the Workshop on Processing Declarative Knowledge, 1991.
- [Bachimont 1996]B. Bachimont, Herméneutique matérielle et Artéfacture - des machines qui pensent aux machines qui donnent à penser, Thèse d'épistémologie, Ecole Polytechnique, Paris, 1996.

- [Bachimont 1999]B. Bachimont, L'intelligence artificielle comme écriture dynamique - de la raison graphique à la raison computationnelle, Grasset, Paris, 1999.
- [Bachimont 2000]B. Bachimont, Engagement sémantique et engagement ontologique - conception et réalisation d'ontologies en ingénierie des connaissances, In Ingénierie des connaissances - évolutions récentes et nouveaux défis, Eyrolles, 2000.
- [Bachimont 2004]B. Bachimont, Arts et sciences du numérique - Ingénierie des connaissances et critique de la raison computationnelle, Mémoire d'Habilitation à Diriger des Recherches, Université de Technologie de Compiègne, 2004.
- [Baeza-Yates 1999]R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, New York (NY), 1999.
- [Banerjee 2002]S. Banerjee, T. Pedersen, An adapted Lesk algorithm for word sense disambiguation using Word-Net, In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, 2002.
- [Banyex 2007]A. Baneyx, J. Charlet, Évaluation, évolution et maintenance d'une ontologie en médecine : état des lieux et expérimentation, Revue I3 – Information, Interaction, Intelligence, numéro spécial « Corpus et ontologies », 2007.
- [Baziz 2005]M. Baziz, M. Boughanem, N. Aussenac-Gilles, C. Chrisment. Semantic Cores for Representing Documents in IR, In Proceedings of the 20th ACM Symposium on Applied Computing, pp. 1020-1026, ACM Press, 2005.
- [Bechhofer 2001] S. Bechhofer, I. Horrocks, C. Goble, R. Stevens, OilEd: a Reasonable Ontology Editor for the Semantic Web, In Proceedings of the Joint German/Austrian Conference on Artificial Intelligence (KI'2001), volume 2174, Springer-Verlag LNAI, 2001.
- [Belkin 2004]N.J. Belkin, G. Muresan, X.M. Zhang, Using User's Context for IR Personalization, In Proceedings of the ACM/SIGIR Workshop on Information Retrieval in Context, 2004.
- [Benjamins 1999]R. Benjamins, D. Fensel, D. Decker, A. Gomez Perez, (KA)2 - building ontologies for the internet - a mid-term report, In Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure, 1999.
- [Berners-Lee 2001]Berners-Lee T., Hendler J., Lassila O, The semantic web - a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 2001, disponible sur le lien: <http://www.med.nyu.edu/research/pdf/mainim01-1484312.pdf>.
- [Bernstein 2005]A. Bernstein, E. Kaufmann, C. Buerki, M. Klein, How Similar Is It? Towards Personalized Similarity Measures in Ontologies, In Proceedings of the 7 Internationale Tagung Wirtschaftsinformatik, 2005.
- [Beseiso 2010]M. Beseiso, A. R. Ahmad, I. Roslan, A Survey of Arabic Language Support in Semantic Web, International Journal of Computer Applications, Volume 9 – No.1, 2010.
- [Borst 1997] P. Borst, Construction of Engineering Ontologies for Knowledge Sharing and Reuse, Ph.D Dissertation, Tweente University, 1997.
- [Bourigault 1996]D. Bourigault, LEXTER, a Natural Language Processing Tool for Terminology Extraction, In Proceedings of 7th EURALEX International Congress, 1996.
- [Bourigault 2000]D. Bourigault, C Fabre, Approche linguistique pour l'analyse syntaxique de corpus, Cahiers de Grammaire, 25, Université Toulouse le Mirail, 2000.
- [Bourigault 2002a]D. Bourigault, G. Lame, Analyse distributionnelle et structuration de terminologie documentaire du droit, Journal TAL, 43-1, 2002 ;
- [Bourigault 2002b]D. Bourigault, UPERY - un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, In Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), 2002.
- [Brachman 1977]R.J. Brachman, What's in a concept - structured foundation for semantic networks, International Journal of Man-Machine Studies 9, 1977.

- [Brachman 1985]R.J. Brachman, J. Schmolze, An overview of the KL-One knowledge representation system, Cognitive Science, 9(2), 1985.
- [Bradley 2001]N. Bradley, The {XML} Companion, Addison-Wesley Professional Publisher, 2001.
- [Brewster 2004]C. Brewster, H. Alani, S. Dasmahapatra, Y. Wilks, Data driven ontology evaluation, In Proceedings of 4th International Conference on Language Resources and Evaluation, 2004.
- [Brini 2005]A. Brini, M. Boughanen, D. Dubois, A Model for Information Retrieval based on Possibilistic Networks, In Proceedings of the 12th Symposium on String Processing and Information Retrieval (SPIRE), à paraître, 2005.
- [Bruandet 1983]M.F. Bruandet, Y. Chiaramella, D. Kerkouba, Méthodes empiriques de construction de thésaurus: expérimentation, Revue de CID, 1983.
- [Budanitsky 2001]A. Budanitsky, G. Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, In Proceedings of the Workshop on WordNet and Other Lexical Resources, ACL, 2001.
- [Caropreso 2000]M-F. Caropreso, S. Matwin, F. Sebastiani, A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization, In Text Databases and Document Management: Theory and Practice, A.G. Chin (Ed.), Idea Group Publishing, Hershey, US, 2000.
- [Charlet 2000]J. Charlet, G. Kassel, M. Zacklad, D. Borigault, Ingénierie des connaissances - recherches et perspectives, In Ingénierie des connaissances, Évolutions récentes et nouveaux défis, Eyrolles, Paris, 2000.
- [Charlet 2002]J. Charlet, L'ingénierie des connaissances, développements, résultats et perspectives pour la gestion des connaissances médicales, Mémoire d'Habilitation à Diriger des Recherches, Université Pierre et Marie Curie, Paris, 2002.
- [Chevalier 2002]M. Chevalier, Interface adaptative pour l'aide à la recherche d'information sur le web, Thèse de doctorat, Université Paul Sabatier, Toulouse, 2002.
- [Chinchor 1998]Chinchor, P. Robinson, Hub-4 Named Entity Task Definition (version 3.5), In Proceedings of the MUC-7, 1998.
- [Chrisment 2006]C. Chrisment, B. Dousset, T. Dkaki, S. Karouach, J. Mothe, Combining Mining and Visualization Tools to Discover the Geographic Structure of a Domain, Computers, Environment and Urban Systems Journal, 2006.
- [Cimiano 2005]Cimiano P., Hotho A., Staab S. Learning concept hierarchies from text corpora using formal concept anaylsis. Journal of Artificial Intelligence Research, JAIR,2005.
- [Condamines 2005]A. Condamines, Sémantique et Corpus, Hermès Science Publications, 2005.
- [Cucchiarelli 2004]R. Cucchiarelli, R. Navigli, F. Neri, P. Velardi, Extending and Enriching WordNet with OntoLearn, In Proceedings of the 2nd Global WordNet Conference, 2004.
- [Cunningham 2002]H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.
- [Dachlet 1990]R. Dachlet, Etat de l'Art de la recherche en informatique documentaire - la représentation des documents et l'accès à l'information, Rapport de recherche de l'INRIA-Rocquencourt, 1990, disponible sur le lien: http://www.inria.fr/rrrt/rr-1201.html, 1990.
- [David 1990]S. David, P. Plante, Termino version 1.0, Report, Centre d'Analyse de Textes par Ordinateur, Université du Québec, 1990.
- [Davis 1993]R. Davis, H. Sorbe, P. Szolovits, What is a Knowledge Representation?, AI Magazine, 1993.

- [Deerwester 1990]S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41(6), 1990.
- [Denjean 1989]P. Denjean, Interrogation d'un systeme videotex - l'indexation automatique des textes, memoire de doctorat, Université Paul Sabatier, Toulouse, 1989.
- [Dervin 1992]B. Dervin, From the mind's eye of the user: the sense-making qualitative-quantitative methodology, In Qualitative Research in Information Management, J.Glazier, R. Powell (Eds.), Englewood, Libraries Unlimited, 1992.
- [Ding 2002]Y. Ding, S. Foo, Ontology Research and Development: Part 1 – A Review of Ontology Generation, Journal of Information Science 28(2), 2002.
- [Domingue 1998]J. Domingue, Tadzebao and WebOnto: Discussing, Browsing and Editing Ontologies on the Web, In Proceedings of the 11th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'98), 1998.
- [Dublincore] http://www.dublincore.org/.
- [Dumais 1994]S. Dumais, Latent Semantic Indexing (LSI) and TREC-2, In D. Harman (Ed.), The Second Text Retrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215, 1994.
- [Dumais 1995]S. Dumais, Using LSI for information filtering: TREC-3 experiments, In D. Harman (Ed.), The Third Text Retrieval Conference (TREC3), National Institute of Standards and Technology Special Publication, 1995.
- [Dumais 2003] S. Dumais, E. Cutrel, J. Cadiz, G. Jancke, R. Sarin, D. Robbins, Stuff I've Seen: A system for personal information retrieval and re-use, In Proceedings of the 26th ACM, Conference on Research and Development in Information Retrieval (SIGIR'03), 2003.
- [Ehrig 2005]M. Ehrig, P. Haase, N. Stojanovic, M. Hefke, Similarity for Ontologies - A Comprehensive Framework, In Proceedings of the 13th European Conference on Information Systems, 2005.
- [Englmeier 2003]K. Englmeier, J. Mothe, IRAIA: A portal technology with a semantic layer coordinating multimedia retrieval and cross-owner content building, In Proceedings of the International Conference on Cross Media Service Delivery, Cross-Media Service Delivery Series, The International Series in Engineering and Computer Science, V. 740, 2003.
- [Erdmann 2000]M. Erdmann, A. Maedche, H. Schnurr, S. Staab, From manual to semi-automatic semantic annotation: About ontology-based text annotation tools, In Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, P. Buitelaar, K. Hasida (Eds.) 2000.
- [esperonto]http://www.esperonto.net
- [Euzenat 2002]J. Euzenat, Eight questions about semantic Web annotations, IEEE Intelligent systems 17(2), 2002.
- [Faatz 2002]A. Faatz, R. Steinmetz, Ontology enrichment with texts from the WWW, In Proceedings of the 2nd Semantic Web Mining Workshop at ECMLI/PKDD, 2002.
- [Fallside 2001]D.C. Fallside, XMLSchema, World Wide Web Consortium (W3C), W3C Recommendation, disponible sur le lien: http://www.w3.org/XML/Schema, 2001.
- [Falquet 2003]G. Falquet and J.-C. Ziswiler - A Virtual Hyperbooks Model to Support Collaborative Learning. AIED 2003 Supplemental Proceedings. Sydney, Australia, July 2003. Disponible sur le lien: < http://sydney.edu.au/engineering/it/~aied/vol10/vol10_FalquetZiswiler.pdf>
- [Farquhar 1997]A. Farquhar, R. Fikes, J. Rice, The Ontolingua Server: a tool for collaborative ontology construction, International Journal of Human-Computer Studies, 1997.
- [Faure 1998]D. Faure, C. Nedellec, A corpus-based conceptual clustering method for verb frames and ontology acquisition, In Proceedings of the LREC workshop on Adapting lexical and corpus resources to sublanguages and applications, 1998.

- [Fernandez 1997]M. Fernandez, A. Gómez-Pérez, N. Juristo, METHONTOLOGY: from ontological art towards ontological engineering, In Proceedings of the Spring Symposium Series on Ontological Engineering (AAAI'97), 1997.
- [Fikes 1985]R. Fikes, T. Kehler, The Role of Frame-Based Representation in Reasoning, Communications of the ACM (CACM), 28(9), 1985.
- [Fischer 1998]D. H. Fischer, From Thesauri towards Ontologies?, In Structures and Relations in Knowledge Organization - Proceedings of the 5th International ISKO Conference, W.M. Hadi, J. Maniez, S. Pollitt (Eds.), 1998.
- [Foskett 1977]D.J. Foskett, Thesaurus, Reproduced in Readings in Information Retrieval, P. Willett, K Sparck-Jones (Eds.), 1977.
- [Foskett 1980]D.J. Foskett, Thesaurus, In Encyclopedia of Library and Information Science, A. Kent, H. Lancour (Eds), 1980.
- [Frakes 1992]W.B. Frakes, R Baeza Yates (Eds.), Information Retrieval Data Structures and Algorithms, Prentice Hall, Englewood Cliffs, New Jersey, 1992.
- [Freund 2005]L. Freund, E.G. Toms, Using contextual factors to match intent, In Proceedings of the ACM SIGIR Workshop on Information Retrieval in Context (IriX), 2005.
- [Furst 2004]F. Furst, Contribution à l'ingénierie des ontologies - une méthode et un outil d'opérationalisation, Thèse de doctorat, Université de Nantes, 2004.
- [Genesereth 1994]M.R. Genesereth, R.E. Fikes, Knowledge interchange format version 3.0 reference manual, 1994, disponible sur le lien: http://logic.stanford.edu/kif/Hypertext/kif-manual.html.
- [getty]http://www.getty.edu/research/conducting_research/vocabularies/aat/.
- [Gómez-Pérez 1996]A. Gómez-Pérez, M. Fernandez, A.J. de Vicente, Towards a Method to Conceptualize Domain Ontologies, In Proceedings of the European Conference on Artificial Intelligence (ECAI'96), 1996.
- [Gómez-Pérez 1999]A. Gómez-Pérez, Evaluation of taxonomic knowledge in ontologies and knowledge bases, In Proceedings of the 12th Knowledge Acquisition for Knowledge-Based Systems Workshop, 1999.
- [Gómez-Pérez 2001]A. Gómez-Pérez, A. Moreno, J. Pazos, A. Sierra-Alonso, Knowledge Maps: An essential technique for conceptualisation, In Data & Knowledge Engineering, 33(2), S. Hyon Myseng (Eds), Kluwer, 2001.
- [Gonzalo 1998]J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarrán, Indexing with WordNet synsets can improve text retrieval, In Proceedings of the COLING/ACL Workshop on Usage of WordNet for Natural Language Processing, 1998.
- [Grefenstette 1992]G. Grefenstette, Use of syntactic context to produce term association lists for text retrieval, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), 1992.
- [Grefenstette 2005]Grefenstette, G., Semmar, N., and Elkateb-Gara, F. 2005. Modifying a natural language processing system for European languages to treat Arabic in information processing and information retrieval applications. In Proceedings of the ACL Workshop on Computational Approaches To Semitic Languages.
- [Gruber 1993a]T.R. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition, 5 (2), 1993, disponible sur le lien: http://tomgruber.org/writing/ontolingua-kaj-1993.htm.
- [Gruber 1993b]T.R. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Stanford Knowledge Systems Laboratory, 1993.
- [Gruber 2007]T.R. Gruber, Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009, disponible sur le lien : http://tomgruber.org/writing/ontology-definition-2007.htm.

- [Gruninger 1995a]M. Gruninger, M. Fox, The logic of enterprise modelling. In Reengineering the Enterprise, J. Brown, D. O'Sullivan (Eds.), Chapman and Hall, 1995.
- [Grunninger 1995b]M. Grunninger, M. Fox, Methodology for the design and evaluation of ontologies, In Proceedings of the IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.
- [Guarino 1994]N. Guarino, M. Carrara, P. Giaretta, Formalizing ontological commitments, In Proceedings of the AAAI conference, 1994.
- [Guarino 1996]Guarino, Understanding, building and using ontologies, In Workshop on Knowledge Acquisition for Knowledge-Based Systems, 1996, disponible sur le lien: http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html.
- [Guarino 1998]N. Guarino, Formal Ontology and Information Systems, In Formal Ontology in Information Systems, N Guarino (Ed.), IOS Press, 1998. Disponible sur le lien: < http://www.loa.istc.cnr.it/Papers/FOIS98.pdf>
- [Guarino 1999]N. Guarino, C. Masolo, G.Vetere, OntoSeek: Content-Based Access to the Web, IEEE Intelligent Systems, 14 (3), 1999.
- [Guarino 2000]N. Guarino et C. Welty, Identity, Unity, and Individuality: Towards a Formal Toolkit forOntological Analysis, In Proceedings of the European Conference on Artificial Intelligence (ECAI), 2000.
- [Guarino 2001]N. Guarino, C. Welty, Identity and Subsumption, In The Semantics of Relationships: an Interdisciplinary Perspective, R. Green, C.A. Bean, S. Hyon Myseng (Eds), Kluwer, 2001.
- [Guarino 2002]N. Guarino, C. Welty, Evaluating Ontological Decisions with OntoClean, In Communication of the ACM, 45(2), 2002.
- [Guha 2003]R.V. Guha, R. McCool, E. Miller, Semantic search, In Proceedings of the 12th International World Wide Web Conference, 2003.
- [Guo 2009]Guo, Ren, Towards the Relationship Between Semantic Web and NLP, 2009.
- [Haav 2001]H.M. Haav, T.L. Lubi, A Survey of Concept-based Information Retrieval Tools on the Web, In Proceedings of the 5th East-European Conference ADBIS, Vol 2, 2001.
- [Hammo 2005]Hammo, Abu-Salem & Lytinten. QARAB: A Question Answering System to Support the Arabic Language, 2005.
- [Hammo 2009]Hammo B., Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents, 2009.
- [Harman 1992]D. Harman, The DARPA TIPSTER project, In SIGIR Forum, volume 26(2), 1992.
- [Harper 1978]D.J. Harper, C.J. van Rijsbergen, An Evaluation of Feedback in Document Retrieval Using Co-Occurrence Data, Journal of Documentation, 34(3), 1978.
- [Harris 1968]Z. Harris, Mathematical Structures of Language, New-York, John Wiley & Sons, 1968.
- [Hawking 1999]D. Hawking, N. Craswell, P. Thistlewaite, D. Harman, Results and challenges in Web search evaluation, In Proceeding of the 8th International Conference on World Wide Web, 1999.
- [Hearst 1992]M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, In Proceedings of the 14th International Conference on Computational Linguistics, 1992.
- [Hearst 1997]M.A. Hearst, C. Karadi, Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, In Proceedings of the 20th International conference on Research and Development in Information Retrieval, SIGIR, 1997.
- [Heijst 1997]G. van Heijst, G. Schreiber, B. Wielinga, Using explicit ontologies for KBS development, International Journal of Human-Computer Studies, 42(2/3), 1997.

- [Hernandez 2003]N. Hernandez, Etude de l'utilisation de syntagmes nominaux pour la catégorisation automatique de documents, In Actes de la conférence INFORSID, 2003.
- [Hernandez 2005]N Hernandez, Ontologies de domaine pour la medélisation du contexte en RI, Thèse Doctorat, Université Paul Sabatier-Toulouse, France 2005.
- [Hersh 2004] W.R. Hersh et al., TREC 2004 Genomics Track Overview, disponible sur le lien: http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf, 2004.
- [Horrocks 2001]I. Horrocks, F. van Harmelen, P.F. Patel-Schneider, Reference description of the DAML+OIL (March2001) ontology markup language, disponible sur le lien: http://www.daml.org/2001/03/ reference.html, 2001.
- [Jacquemin 1999]C. Jacquemin, E. Tzoukermann, NLP for term variant extraction_ A synergy of morphology lexicon and syntax, In Natural Language Information Retrieval, T. Strzalkowski (Ed.), 1999.
- [Jain 1999]A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No 3, 1999;
- [Jarvelin 1996]K. Jarvelin, J. Kristensen, T. Niemi, E. Sormunen, H. Keskustalo, Expansion Tool: a deductive data model for thesauri and query expansion, Finnish information studies FIS-1996-5, Department of Information Studies,University of Tampere, 1996.
- [Jarvelin 2004]K. Jarvelin. Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance, 2004.
- [JFIC2009]Actes des 20es Journées Francophones d'Ingénierie des Connaissances « Connaissance et communautés en ligne » du 25 au 29 mai 2009 à Hammamet, Tunisie.
- [Jiang 1997]J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical terminology, In Proceedings of the International Conference on Computational Linguistics, (RoclingX), 1997.
- [Johnson 2003]J.D Johnson, On contexts in information seeking, Journal of the American Society for Information Science, 39 (5), 2003.
- [Jones 2000]G.J.F. Jones, New Challenges for Cross-Language Information Retrieval: Multimedia Data and the User Experience, Lecture Notes In Computer Science; Vol. 2069, Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation, 2000.
- [joradp]http://www.joradp.dz
- [Kahan 2001]J. Kahan, M. Koivunen, E. Prud'Hommeaux, R. Swick, Annotea: An Open RDF Infrastructure for Shared Web Annotations, In Proceedings of the 10th International World Wide Web Conference, 2001.
- [Karp 1999]R. Karp, V. Chaudhri, J. Thomer, Xol: An xml-based ontology exchange language, disponible sur le lien: http://www.ai.sri.com/ pkarp/xol, 1999.
- [Kassel 1999]G. Kassel, S. Perpette, Cooperative ontology construction needs to carefully articulate terms, notions and objects, In Proceedings of the International Workshop on Ontology Engineering on the Global Information Infrastructure, 1999.
- [Kassel 2002]G. Kassel, OntoSpec - une méthode de spécification semi-informelle d'ontologies, In Actes des 13èmes journées francophones d'Ingénierie des Connaissances (IC), 2002.
- [Kavalec 2004]M. Kavalec, A. Maedche, V. Svátek, Discovery of Lexical Entries for Non-taxonomic Relations in Ontology Learning, In Proceedings of SOFSEM, 2004.
- [Kaveh 2004]Kaveh B., Gilles F., Management et Technologies des Systèmes d'Information (MATIS), Groupe Interface des Systèmes d'Information (ISI) Centre Universitaire d'Informatique (CUI) Université de Genève; Suisse, Juin 2004, disponible sur le lien: http://cui.unige.ch/~bazargan/PDF/Rapport-KB-DEA-MATIS.pdf.
- [Kayser 1997]D. Kayser, La représentation des connaissances, Hermès, Paris, 1997.

- [Khan 2002]L. Khan, F. Luo, Ontology Construction for Information Selection, In Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence, 2002.
- [Kifer 1995]M. Kifer, G. Lausen, J. Wu, Logical Foundations of Object-Oriented and Frame-Based Languages, Journal of the ACM, 42(4), 1995.
- [Kiryakov 2004]A. Kiryakov, B. Popov, I. Terziev, D. Manov, D. Ognyanoff, Semantic annotation, indexing, and retrieval, Journal of Web Semantics, 2(1), 2004.
- [Klein 2000] M. Klein, D. Fensel, F. van Harmelen and I. Horrocks. The Relation between Ontologies and Schema-Languages: Translating OIL-Specifications to XMLSchema. In Proceedings of the Workshop on Applications of Ontologies and Problem-solving Methods, 14th European Conference on Artificial Intelligence ECAI-00, Berlin, Germany, August 20-25th 2000.Disponible sur le lien: <http://www.cs.vu.nl/~frankh/postscript/ECAI00-WS2.pdf>
- [Koo 2003]S. Koo, S.Y. Lim, S.J. Lee, Building an Ontology based on Hub Words for Informational Retrieval, In Proceedings of the IEEE/WIC International Conference on Web Intelligence, 2003.
- [Lame 2002]G. Lame, Construction d'ontologie à partir de texte, une ontologie du droit dédiée à la recherche d'information sur le Web, Thèse de doctorat, Ecole des Mines de Paris, 2002.
- [Lassila 1999]O. Lassila, R. R. Swick, Resource description framework (rdf) model and syntax speci_cation w3c recommendation, 1999, disponible sur le lien: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.
- [Lassila 2001]O. Lassila, D. McGuiness, The role of frame-based representation on the semantic Web, Rapport technique KSL-01-02, Knowledge Systems Laboratory, Stanford University, 2001.
- [Laublet 2003]P. Laublet, Chantal Reynaud, Jean Charlet, Sur quelques aspects du Web sémantique, Actes des deuxièmes assises nationales du GdR I3, Paris 2003.
- [Lausen 2004]H. Lausen, M. Stollberg, R. Lara, Y. Ding, S.-K Han, D. Fensel, Semantic Web Portals: State of the Art Survey, Technical Report DERI-TR-2004-04-03, 2004.
- [Lawrie 2000]D. Lawrie, W. B. Croft, Discovering and Comparing Topic Hierarchies, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), 2000.
- [Le Moigno 2002]S. Le Moigno, J. Charlet, D. Bourigault, M.C. Jaulent, Construction d'une ontologie à partir de corpus - expérimentation et validation dans le domaine de la réanimation chirurgicale, In Actes des 6 es Journées Ingénierie des Connaissances, 2002.
- [Leacock 1998]C. Leacock, M. Chodorow, Combining local context and Wordnet similarity for word sense identification, in WordNet: an electronic lexical database, C. Felbaum (Ed), Cambridge, MA, The MIT Press, 1998.
- [Lee 1995]J. Lee, G. Yost, P.W. Group, The PIF process interchange format and framework, Technical Report 180, MIT Center for Coordination Science, 1995.
- [Lesk 1988]M. Lesk, "They said true things, but called them by wrong names" – vocabulary problems in retrieval systems, In Proceedings of the 4th Annual Conference of the University of Waterloo Centre for the New OED, 1988.
- [Liebowitz 1998]J. Liebowitz, T. Beckman, Knowledge Organizations: What Every Manager Should Know, St. Lucie Press, 1998.
- [Lin 1998]D. Lin, An information-theoric definition of similarity, In Proceedings of the 15th international conference on Machine Learning, 1998.
- [Lindberg 1993]D.A. Lindberg, B.L. Humphreys, A.T. McCray, The Unified Medical Language System, Methods Inf Med, 32(4), 1993, disponible sur le lien: http://www.openclinical.org/medTermUmls.html.
- [Lord 2003] P.W. Lord, R.D. Stevens, A. Brass, C.A. Goble, Semantic similarity measures as tools for exploring the Gene Ontology, In Proceedings of the Pacific Symposium on Biocomputing, 2003.

- [Lovins 1968]J.B. Lovins, Development of a stemming algorithm, Mechanical translation and computational linguistics, Vo11, 1968.
- [Lozano-Tello 2004]A. Lozano-Tello, A. Gómez-Pérez, ONTOMETRIC: A Method to Choose the Appropriate Ontology, Journal of Database Management, 15(2), 2004.
- [Lri-annaba]http://lri-annaba.org/ds_lri/sites/default/files/article%20ged.pdf
- [Luke 2000]S. Luke, J. Hein, Shoe 1.01 proposed specification, SHOE project, 2000, disponible sur le lien: http://www.cs.umd.edu/projects/plus/SHOE/spec.html, 2000.
- [MacGregor 1991]R. MacGregor, Using a description classifier to enhance deductive inference, In Proceedings of the 7th IEEE Conference on AI Application, pp 93-97, 1991.
- [Maedche 2000]A. Maedche, S. Staab, Mining ontologies from text, In Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management, Springer Lecture Notes in Artificial Intelligence, 2000.
- [Maedche 2001]A. Maedche, S. Staab, Ontology Learning for the Semantic Web, IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2), 2001
- [Maedche 2002]A. Maedche et S. Staab, Measuring similarity between ontologies, In Proceedings of the 13th International Conference EKAW, 2002.
- [Maedche 2002]A. Maedche, V. Pekar, S. Staab, Ontology learning part one – on discovering taxonomic relations from the web, In Web Intelligence, Z. Ning et al (Eds.), Spinger, 2002.
- [Maedche 2003]A. Maedche et S. Staab, N. Stojanovic, R. Studer, Y. Sure, SEmantic portAL: The SEAL Approach, In Spinning the Semantic Web, D. Fensel, J.A. Hendler, H. Lieberman, W. Wahlster, (Eds.), MIT Press, Cambridge London, 2003.
- [Maedche 2004]A. Maedche, S. Staab, Ontology Learning, Handbook on Ontologies, S Staab, R. Stubers (Eds.), 2004.
- [Mahiou 1984]A. Mahiou, Etudes de droit public algérien, OPU, Alger, 1984.
- [Mandala 1999] R. Mandala, T. Tokunaga, H. Tanaka, Combining multiple evidence from different types of thesaurus for query expansion, In Proceedings of the 22nd International ACM SIGIR conference on Research and Development in Information Retrieval, 1999.
- [Manning 1999]C.D. Manning, H. Schuetze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, Massachusetts, 1999.
- [Martin 1995]P. Martin. Using the WordNet Concept Catalogue and a Relation Hierarchy for Knowledge Acquisition. Proc. of Peirce'95, 4th, International Workshop on Peirce, University of California, Santa Cruz, USA, 1995, disponible sur le lien: <http://www.phmartin.info/webKB/doc/papers/peirce95/peirce95.pdf>.
- [Mc Hale 1998]M. Mc Hale, A comparison of Wordnet and Roget's taxonomy for measuring semantic similarity, In Proceedings of the COLING/ACL Workshop on Usage of Wordnet in Natural Langage Processing Systems, 1998.
- [McGuinness 2004]D.L McGuinness, F. van Harmelen, OWL Web Ontology Language Overview, W3C Recommendation http://www.w3.org/TR/owl-features/, 10 February 2004.
- [Mesfar 2008]S. Mesfar, Analyse morpho-syntaxique automatique et reconnaissance des entités nomées en arabe standard, these de doctorat en informatique, universite de Franche-Compte Besançon, France, 2008,
- [Mihalcea 2000]R. Mihalcea, D.I. Moldovan, Semantic Indexing using WordNet Senses, In Proceedings of ACL Workshop on IR & NLP, 2000.
- [Miles 2005]A. Miles, D. Brichley, SKOS Core GuideW3C Working Draft 10 May 2005, disponible sur le lien: http://www.w3.org/TR/swbp-skos-core-guide/.
- [Milks 2002]Y. Milks, Ontotherapy or how to stop worrying about what there is, In Proceedings of the Workshop on Ontologies and Lexical Knowledge Bases, 2002.
- [Miller 1988]G.A. Miller, Nouns in WordNet, In WordNet, An Electronic Lexical Database C. Fellbaum (Ed), MIT Press, 1988.

- [Miller 1991]G. Miller,W.G. Charles, Contextual Correlates of Semantic Similarity, Language and Cognitive Processes, 6(1), 1991.
- [Miller 1993]G. Miller, C. Leacock, R. Tengi, R.T. Bunker, A Semantic Concordance, In Proceedings of ARPA Workshop on Human Language Technology, 1993.
- [Miller 2002]L. Miller, A. Seaborne, A. Reggiori, Three implementations of squishql, a simple rdf query language, In Proceedings of the International Semantic Web Conference, 2002.
- [Minsky 1975]M. Minsky, A framework for representing knowledge, In Psychology of Computer Vision, P.H. Winston (Ed), 1975.
- [Mitra 1997] M. Mitra, C. Buckley, A. Singhal, C. Cardie, An analysis of Statistical and Syntactic Phrases, In Actes de la conférence Recherche d'Information Assistee par Ordinateur (RIAO), 1997.
- [Moldovan 1999]D. Moldovan, S. Harabagiu, M. Pasca, R. Mihal-cea, R. Goodrum, R. Girju, V. Rus, LASSO: A tool for surfing the answer net, Proceedings of the 8th Text Retrieval Conference (TREU-8), 1999.
- [Montaner 2003]M. Montaner, B. Lopez, J.L. De La Rosa, A taxonomy of recommender agents on the Internet. Artificial Intelligence Review, 19, 2003.
- [Morin 1999]E. Morin, Using Lexico-Syntactic Patterns to Extract Semantic Relations between terms from Technical Corpus'', In Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE'99), 1999.
- [Murtagh 1998]F. Murtagh, Clustering and Classification, The Computer Journal, 41, 1998.
- [Nonaka 1995]I. Nonaka, H. Takeuchi, The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation, Oxford University Press, 1995.
- [Noy 2000]N. Noy, R.W. Ferguson, M.A. Musen, The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility, In Proceedings of the 12th European Knowledge Acquisition Workshop (EKAW'00), 2000.
- [ontoWeb]http://www.ontoWeb.org
- [owl-guide]http://www.w3.org/TR/owl-guide/
- [PAPINI 2010]O. PAPINI, Introduction au WEB Sémantique, ESIL Université de la méditerranée, 2010, disponible sur le lien: <http://odile.papini.perso.esil.univmed.fr/sources /WEBSEM/cours-WEBSEM-1.pdf>.
- [Paralic 2003]J.Paralic, I.Kostial, Ontology-based Information Retrieval, In Proceedings of the 14th International Conference on Information and Intelligent Systems, 2003.
- [Patwardhan 2003] S. Patwardhan, S. Banerjee, T. Pedersen, Using Measures of Semantic Relatedness for Word Sense Disambiguation, In Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics, 2003.
- [Pinto 2001]H.S. Pinto, J.P. Martins, A methodology for ontology integration, In Proceedings of the International Conference on Knowledge Capture, ACM Press, 2001.
- [Pohlmann 1997] R. Pohlmann, W. Kraaij, The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), L. Devroye, C. Chrisment (Ed.), 1997.
- [Ponte 1998]M. Ponte, W. B. Croft, A Language Modeling Approach to Information Retrieval, Research and Development in Information Retrieval, In Proceeding of the 21st International ACM-SIGIR conference on Research and Development in Information Retrieval, 1998.
- [Porter 1980]M. Porter, An algorithm for suffix stripping Program, 14(3), 1980.
- [Porzel 2004]R. Porzel, R. Malaka, A Task-based Approach for Ontology Evaluation, In Proceedings of the ECAI Workshop on Ontology Learning and Population, 2004.
- [Quillian 1968]M.R. Quillian, Semantic memory, In Semantic Information Processing, MIT press, Cambridge, 1968.

- [Rada 1989]R. Rada, H. Mili, E. Bicknell, M.Blettner, Development and application of a metric on semantic nets, IEEE Transaction on Systems, Man and Cybernetics, 19(1), 1989.
- [Rauber 2001]A. Rauber, A. Muller-Kogler, Integrating automatic genre analysis into digital libraries, In Proceedings of the 1st ACM-IEEE-CD Joint Conference on Digital Libray (JCDL), 2001.
- [Resnik 1995]P. Resnik, Using information content to evaluate similarity in a taxonomy, In Proceedings of the 14th joint conference in Artificial Intelligence, 1995.
- [Resnik 1999]P. Resnik, Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural langage, Journal of Artificial Intelligence Research, volume 11, 1999.
- [Richardson 1995]R. Richardson, A.F. Smeaton, Using WordNet in a Knowledge-Based Approach to Information Retrieval, Working Paper, CA-0395, School of Computer Applications, Dublin City University, Ireland, 1995.
- [Rieu 1999]D. Rieu, Ingénierie des systèmes d'information - bases de données, bases de connaissances, et méthodes de conception, Mémoire d'Habilitation à Diriger des Recherches, INP de Grénoble, 1999.
- [Riloff 1996]E. Riloff, Automatically generating extraction patterns from untagged text, In Proceedings of the 13th National Conference on Artificial Intelligence, 1996.
- [Rivier 1990]A. Rivier, Construction des langages d'indexation Aspects théoriques, Documentaliste, vol. 27 (6), 1990.
- [Roberston 2002] S. Roberston, I Soboroff, The TREC 2002 Filtering Track Report, disponible sur le lien: http://trec.nist.gov/pubs/trec11/papers/OVER.FILTERING.pdf, 2002.
- [Robertson 1976]S. E. Robertson, K. Sparck Jones, Relevance weighting of search terms, Journal of the American Society for Information Sciences, 27 (3), 1976.
- [Rocchio 1971]J. Rocchio, Relevance Feedback in Information Retrieval, The SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton (Ed), 1971.
- [Rocha 2004]C. Rocha, D. Schwabe, M.P. Aragão, A Hybrid Approach for Searching in the Semantic Web, In Proceedings of the 13th International World Wide Web Conference, 2004.
- [Saias 2003]J. Saias, P. Quaresma, A Methodology to Create Ontology-Based Information Retrieval Systems, In Proceedings of the EPIA Conference, 2003.
- [Salton 1971] G. Salton, The Smart Retrieval System, Prentice Hall, USA, 1971.
- [Salton 1990]G. Salton, C. Buckley, Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science, 44 (4), 1990.
- [Sanderson 1999]M. Sanderson, W.B. Croft, Deriving concept hierarchies from text, In Proceedings of the 22nd International ACM SIGIR Conference, 1999.
- [Sanderson 2000] M. Sanderson, Retrieving with good sense, In Information Retrieval Vol. 2 No. 1, 2000.
- [Savoy 1993]J. Savoy, Stemming of French Words Based on Grammatical Categories, Journal of the American Society for Information Science, 44(1), 1993.
- [Schmid 1994]H. Schmid, Probabilistic Part-of-Speech Tagging Using Decision Trees, In Proceedings of the International Conference on New Methods in Language Processing, 1994.
- [Sebastiani 2006]F. Sebastiani, Classification of text, automatic, In The Encyclopedia of Language and Linguistics, K. Brown (Ed.), Volume 14, 2nd Edition, Elsevier Science Publishers, Amsterdam, NL, 2006, disponible sur le lien: http://www.math.unipd.it/~fabseb60/Publications/ELL06.pdf.
- [Seeling 2003]C. Seeling, A. Becks, Exploiting Metadata for Ontology-Based Visual Exploration of Weakly Structured Text Documents, In Proceedings of the 7th International Conference on Information Visualisation (IV03), IEEE Press, 2003.

- [Séguéla 1999]P. Séguéla, N. Aussenac-Gilles, Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine, In Actes de la Conférence Ingénierie des Connaissances, 1999.
- [Semmar 2007]Semmar, N., Fluhr, C., Arabic to French Sentence Alignment: Exploration of A Crosslanguage Information Retrieval Approach, 5th Workshop on Important Unresolved Matters, Prague, Czech Republic, 2007.
- [Shadbolt 1993]N. Shadbolt, E. Motta, A. Rouge, Constructing knowledge based systems. IEEE Software, 10(6), 1993.
- [Small 1982]S. Small, C. Rieger, Parsing and comprehending with word experts (a theory and its realisation), in Strategies for Natural Language Processing, W.G. Lehnert & M. H. Ringle (Eds.), 1982.
- [Soboroff 2003]I. Soboroff, D. Harman, Overview of the TREC 2003 Novelty Track, disponible sur le lien: http://trec.nist.gov/pubs/trec12/papers/NOVELTY.OVERVIEW.pdf
- [Soergel 1974]D. Soergel, Indexing Languages and Thesauri: Construction and Maintenance, Los Angeles, Melville Publ. Company, 1974.
- [Soergel 2004]D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer, S. Katz, Reengineering Thesauri for New Applications: the AGROVOC Example, Journal of Digital Information, Volume 4 Issue 4, 2004.
- [Sowa 1984]J.F. Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley Publishing Company, USA, 1984.
- [Spiteri 1999]L. Spiteri, The essential elements of faceted thesauri, Cataloging & Classification Quarterly, 28(4), 1999.
- [Srikant 1995]R. Srikant, R. Agrawal, Mining generalized association rules, In Proceedings of the 21st Conference on Very Large DataBases (VLDB'95), 1995.
- [Staab 2000]S. Staab, A. Maedche, Axioms are objects too: Ontology engineering beyond the modeling of concepts and relations, Research report 399, Institute AIFB, Karlsruhe, 2000.
- [Stuckenschmidt 2004]H. Stuckenschmidt, F. van Harmelen, A. de Waard, T. Scerri, R. Bhogal, J. van Buel, I.Crowlesmith, C. Fluit, A. Kampman, J. Broekstra, E. van Mulligen, Exploring large document repositories with RDF technology: the DOPE project, Intelligent system, IEEE, Vol. 19, No. 3, 2004.
- [Studer 1998]R. Studer, R. Benjamins, D. Fensel, Knowledge Engineering: Principles and Methods, Data and Knowledge Engineering, 25(1-2), 1998.
- [Sugiura 2004]Sugiura, Y. Shigeta, N. Fukuta, N. Izumi, T. Yamaguchi, Towards On-the-Fly Ontology Construction – Focusing on Ontology Quality Improvement. In Proceedings of the 1st European Semantic Web Symposium (ESWS), 2004.
- [Sure 2002]Y. Sure, J. Angele, S. Staab, OntoEdit: Guiding Ontology Development by Methodology and Inferencing, In Proceedings of the Confederated International Conferences CoopIS, DOA and ODBASE 2002, volume 2519, Springer-Verlag LNCS, 2002.
- [Teimzit 2010]A. Teimzit, T.E. Belhaoues, T. Bensebaa, Construction d'ontologie d'algorithmique et son utilisation dans un EIAH, Laboratoire de Recherche en Informatique (LRI), Université Badji Mokhtar Annaba, Algérie, 2010.
- [Tfidf] : http://www.tfidf.com/
- [Thieu 2004]M. Thieu, O. Steichen, Ch. Le Bozec, E. Zapletal, M.-Ch. Jaulent, Mesures de similarité pour l'aide au consensus en anatomie pathologique, In Proceedings of the 5th International Conference on Internet Computing, 2004.
- [topicmaps] http://www.topicmaps.org/
- [Tudhope 2001]D. Tudhope, H. Alani, C. Jones, Augmenting Thesaurus Relationships: Possibilities for Retrieval, Journal of Digital Information, 1-8(41), 2001.
- [Turtle 1991]H.R. Turtle, Inference Networks for Document Retrieval, PhD Thesis, University of Massachusetts, 1991.

- [Uschold 1995]M. Uschold, M. King, Towards a Methodology for Building Ontologies. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at the International Joint Conference on Artificial Intelligence (IJCAI'1995), 1995.
- [Uschold 1996]M. Uschold, M. Gruninger, Ontologies: principles, methods, and applications, Knowledge Engineering Review, 11(2), 1996.
- [Uschold 1998]M. Uschold, M. Healy, K.Williamson, P. Clark, S.Woods, Ontology Reuse and Application, In Proceedings of the International Conference on Formal Ontology and Information Systems, 1998.
- [Uschold 2003] M. Uschold, Where are the semantics in the semantic Web?, AI Magazine, 24(3), 2003.
- [Vakkari 2003]P. Vakkari, Task-based information searching, Annual Review of Information Science and Technology, 37, 2003.
- [Vallet 2005]D. Vallet, M. Fernández, P. Castells, An Ontology-Based Information Retrieval Model, In Proceedings of the 2nd European Semantic Web Conference, 2005.
- [Velardi 2002]P. Velardi, P. Fabriani, M. Missikoff, Using text processing techniques to automatically enrich a domain ontology, In Proceedings of the ACM Conference on Formal Ontologies and Information Systems, 2002.
- [Véronis 1989]J. Véronis, N. Ide, N. Wurbel, Extraction d'informations sémantiques dans les dictionnaires courants, In Actes du 7ème congrès Reconnaissance des Formes et Intelligence Artificielle, 1989.
- [Voorhees 1993]E.M. Voorhees, Using WordNet to disambiguate word sense for text retrieval, In Proceedings of the 13th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993.
- [Voorhees 2004]E.M. Voorhees, D.M. Tice, The TREC-8 Question Answering Track Evaluation, 2004, sur: http://trec.nist.gov/pubs/trec13/papers/QA.OVERVIEW .pdf.
- [W3]: http://www.w3.org/.
- [WELTY 2001]WELTY C. & GUARINO N., Supporting ontological analysis of taxonomic relationships, Data et Knowledge Engineering (39), 2001.
- [Woods 1975]W.A. Woods, What's in a link: Foundation for Semantic Networks, In Representation and Understanding; Studies in Cognitive Science, D.G. Bobrow, A. Collins (Eds.), Academic Press, 1975.
- [Woolf 1990]H. Woolf (Ed.), Websters New World Dictionary of the American Language, G. & C.Merrian, 1990.
- [Wu 1994]Z. Wu, M. Palmer, Verb semantics and lexical selection, In Proceedings of the 32nd annual meeting of the Association for Computational Linguistics, 1994.
- [Xu 2000]J. Xu, W.B. Croft, Improving the Effectiveness of Information Retrieval with Local Context Analysis, ACM Transactions of Information Systems, 18(1), 2000.
- [Yan 2005]Yan Qu, Gregory Grefenstette, David A. Evans: The Use of Monolingual Context Vectors for Missing Translations in Cross-Language Information Retrieval, Conférence internationale conjointe sur Natural Language Processing, 2005.
- [Zaidi 2008]S. Zaidi, Les ontologies, cours de mastère, LRI 2008.
- [Zhang 2004]S. Zhang, O. Bodenreider, Comparing Associative Relationships among Equivalent Concepts across Ontologies, In Proceedings of MEDINFO 2004, 2004.
- [Zipf 1949]G. Zipf, Human Behaviour and the Principle of Least Effort, Addison-Wesley, 1949.
- [Zweigenbaum 1993]P. Zweigenbaum et al., Linguistic and medical knowledge bases: An access system for medical records using natural language, Technical report, MENELAS: deliverable 9, AIM Project A2023, 1993.