Gradients de prototypicalité, mesures de similarité et de proximité sémantique : une contribution à l'Ingénierie des Ontologies.

Xavier Aimé

UNIVERSITÉ DE NANTES

Ecole polytechnique de l'Université de Nantes

———

ÉCOLE DOCTORALE

« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE MATHEMATIQUES »

Année 2011

N° attribué par la bibliothèque

Gradients de prototypicalité, mesures de similarité et de proximité sémantique : une contribution à l'Ingénierie des Ontologies

———

THÈSE DE DOCTORAT

Discipline : Informatique

Spécialité : Ingénierie des Connaissances

Présentée
et soutenue publiquement par

Xavier AIME

Le 8 avril 2011, devant le jury ci-dessous

Président	Serge GARLATTI, Professeur, Telecom Bretagne, Brest
Rapporteurs	Jean CHARLET, Maître de conférences, Faculté de médecine, Paris V Gilles KASSEL, Professeur, Université Picardie Jules Verne, Amiens
Examinateurs	Frédéric FÜRST, Maître de conférences, Université Picardie Jules Verne, Amiens Pascale KUNTZ, Professeur, École Polytechnique de l'Université de Nantes Philippe LAUBLET, Maître de conférences, Université Paris-Sorbonne, Paris IV Francky TRICHET, Maître de conférences, Université de Nantes
Invité	Bernard FORT, Société Tennaxia, Paris

Directeur de thèse : Pascale KUNTZ

ED : 503-122

A ma femme et mes enfants

Remerciements

Mes pensées et ma reconnaissance s'adressent en premier lieu à tous mes encadrants, Francky, Fred et Pascale. Ils m'ont guidé, inspiré, supporté (dans tous les sens du terme) durant ces années. Un immense merci pour cette autre vision du monde de la recherche. Qu'Abélard et Bacon se rassurent, leur combat n'aura pas été vain, leurs héritiers continuent l'ouvrage.

Mes pensées et ma reconnaissance s'adressent aussi à Tennaxia, et à toutes les personnes qui m'ont accompagné dans ce projet. Merci à Bernard, pour son soutien et son aide. Merci à toute l'équipe de Paris avec qui j'ai partagé bien plus qu'un bureau : Vincent, Laurent et Mariusz (mais aussi Nicolas et Christophe de Lyon), Max, Anne, Françoise, Laurence, Olivier, Raphaël et Rossella, sans oublier Antoine, Cyrille, Karine et Xavier.

Mes pensées et ma reconnaissance s'adressent également à tous ceux qui m'ont suivi de près ou de loin dans cette aventure, et dont l'influence est loin d'être négligeable. Je pense en particulier aux professeurs Ganasacia et Rastier. Je les remercie infiniment pour l'attention qu'ils ont bien voulu apporter à mes travaux.

Mes pensées et ma reconnaissance s'adressent enfin à toutes ces personnes rencontrées qui font qu'une thèse est également une aventure humaine. Toutes et tous ont contribué à leur manière, non seulement à ce travail mais aussi à ce que je suis devenu. Merci à Sylvie et Anthony pour leur petit coup de pouce durant la première année. Merci à Laetitia pour m'avoir fait découvrir Georges Kleiber. Merci à M. Tanara pour son intérêt pour ce travail, ces entretiens des plus enrichissants et pour m'avoir fait découvrir le Japon et sa merveilleuse langue. Merci à Mme Li pour son intérêt pour ce travail et ces discussions passionnantes sur l'Empire du Milieu. Merci à Yohan et Frédéric pour Osiris et Morris (vive les acronymes). Merci à Hélène pour sa gentillesse et son accueil inoubliable à Athènes lors de ma première conférence internationale. Merci à Juan José et Concuello pour la découverte d'une cuisine mexicaine toute aussi chaleureuse que leur cœur. Merci à Lise, hôtesse sur American Airlines, pour sa gentillesse et son très joli sourire durant les 22 heures de vol (aller et retour). Merci à Cacho et Roseo pour m'avoir donné envie de découvrir l'Argentine. Merci à Fiona, Dany et toutes ces personnes rencontrées en Chine qui m'auront fait aimé encore d'avantage ce pays fabuleux. Merci à Richard, pour ces nombreuses discussions mémorables sur les signes et les symboles. Merci à Jeremy, pour toutes ces idées échangées et qui - une fois semées - devraient donner de belles plantes. Merci à tout ceux que j'oublie dans cet inventaire à la Prévert, mais qui d'une manière ou d'une autre sont bien présents à travers les mots de cette thèse.

Mais cette thèse n'est pas qu'une histoire d'hommes et de femmes, elle a été aussi une histoire de lieux (avec une pensée particulière en ces lignes pour Amblard de Guerry). Tout autant que les Hommes, ils m'ont accueilli, inspiré, transporté du-

rant ces 1200 jours. Certains d'entre eux sont associés très étroitement à quelques parties de ma thèse, il aurait été ingrat de ne pas les citer dans les remerciements. Je pense au restaurant universitaire "Le Rubis", où la première formule du gradient de prototypicalité conceptuelle a été griffonnée sur un coin de serviette. Je pense au "Café de la Mairie", sur la place St-Sulpice à Paris, où la première formule du gradient de prototypicalité lexicale a été trouvée. Je pense au "Café des Editeurs", près de l'Odéon à Paris, où l'idée des mesures de similarité et proximité a germé. Je pense au "Bistrot des Frangines", près de Montparnasse, où les dernières lignes de cette thèse furent écrites. Je pense aux vieilles pierres vendéennes de ma maison, où j'ai préparé ma soutenance. Sans oublier, mon premier survol du Groenland, les 14h de train de nuit debout entre Pingyao et Beijing, la balade en vélo sur les murailles de Xi'an, les jardins de Faro...

Pour terminer cette page, toujours un peu spéciale, je voudrais également témoigner de ma profonde pensée pour tous ceux avec qui je ne pourrai pas partager cet instant, mais qui - de là où ils sont - ne seraient pas mécontents d'un tel résultat.

Je pense très fort à mes parents - mes auteurs favoris, ma femme Léa - ma plus belle trouvaille, et à mes enfants Apolline et petit bébé - mes plus belles créations. Je tiens à leur dédier à tous cette thèse, en guise de remerciements éternels pour m'avoir aidé, porté et supporté durant cette longue période.

Table des matières

Introduction

Du besoin particulier de Tennaxia...

Cette thèse a été réalisée au sein de la société TENNAXIA [1], dans le cadre d'un contrat CIFRE entre cette dernière et l'Université de Nantes. Tennaxia est une société de services, créée en 2001, qui associe conseils et logiciels en ligne dans les domaines de l'Hygiène, de la Sécurité et de l'Environnement (HSE). Elle s'adresse principalement aux grands groupes industriels afin de leur permettre de (1) maîtriser leurs risques et sécuriser leur gestion opérationnelle, (2) gagner en efficacité au niveau de leur fonction HSE en externalisant les tâches à faible valeur ajoutée ou les tâches faisant appel à des compétences difficiles à acquérir et maintenir en interne, et (3) anticiper les évolutions réglementaires qui peuvent avoir une influence importante sur leur avenir.

Afin de répondre à ces problématiques, Tennaxia propose à ses clients une suite logicielle regroupant les fonctionnalités nécessaires à la gestion HSE et Développement Durable. Ces fonctionnalités servent dans trois domaines : (1) *réglementaire* comme la veille et les audits de conformité, (2) *opérationnel* comme le suivi des rejets et des consommations de ressources, et (3) *management* comme la stratégie développement durable et performance environnementale.

Le projet porté par ce contrat CIFRE concerne le module *Veille & Conformité* de cette suite logicielle. Ce module permet d'assurer une traçabilité complète depuis la veille réglementaire jusqu'aux audits de conformité réglementaire. Il s'agit en premier lieu d'une offre de conseils qui consiste à identifier les textes et les exigences applicables afin d'anticiper les évolutions induites par certaines nouveautés réglementaires. Ces conseils s'appliquent de manière graduée en fonction des besoins des industriels. Ils reposent sur un ensemble de textes réglementaires, allant de l'arrêté municipal à la Directive européenne, textes analysés manuellement par les consultants et ventilés en *exigences réglementaires* puis complétés par des notes d'analyse.

Le module *Veille & Conformité* dispose d'un moteur de recherche sur la base de textes réglementaires, laquelle est entièrement indexée au moyen du moteur d'indexation *Lucène* [2]. Il s'agit essentiellement d'un filtre sur cette base de textes : ce module sélectionne les documents suivant des critères saisis (mots/termes contenus dans le titre, le texte, le résumé, etc.) par l'utilisateur, sans fournir en retour un tri des documents par pertinence par rapport à la requête.

1. Société Anonyme avec Directoire et Conseil de Surveillance au capital de 77 498 euros. Siège social : 3 rue Léonard de Vinci, 53000 Changé. http ://www.tennaxia.com
2. http ://lucene.apache.org/java/docs/index.html

L'objectif applicatif de cette thèse est de proposer, dans le cadre de ce module, un prototype de moteur de recherche sémantique fondé sur une ontologie du domaine HSE. Ce moteur de recherche doit, à partir d'un terme saisi par un consultant, (1) rechercher sa correspondance conceptuelle dans l'ontologie de domaine, et (2) étendre la requête au moyen des termes dénotant une série de concepts issue du parcours de l'ontologie à partir du terme saisi et identifié. Par exemple, une recherche sur le terme *carton* peut être étendue aux concepts *Emballage* (un carton est un emballage), *Boîte en carton*, *Collecteur en carton* et *Caisse en carton* (tous trois sont des types de cartons). Il est également envisagé d'étendre ces requêtes en utilisant les relations du domaine. Par exemple, une recherche sur le concept *Peinture* peut porter également sur les *Composés Organiques Volatils (COV)*, car les peintures *rejettent* des COV.

Mais, cette extension se doit d'être guidée par les usages. En effet, deux consultants de spécialités différentes peuvent obtenir des résultats différents pour une même recherche. Par exemple, pour une même recherche sur le concept *Peinture*, un consultant spécialisé dans les déchets est intéressé par une extension sur le concept *Solvant*, alors qu'un consultant spécialisé en sécurité est plus intéressé par une extension sur le concept *COV*.

L'idée sous-jacente est donc de pouvoir tenir compte d'un contexte de recherche dans le parcours de l'ontologie. L'objet de notre recherche porte donc sur la pragmatisation d'une ontologie de domaine et son insertion dans le cadre d'une recherche d'informations sémantiques.

... à l'introduction de la prototypicalité en Ingénierie des Connaissances (IC)

En 1999, A. Léger dans la préface de [Charlet 2000], définissait l'Ingénierie des Connaissances comme une "science" jeune et pluridisciplinaire s'inscrivant dans la quête humaine de la maîtrise calculable du sens. Les auteurs complétaient cette définition par : *domaine correspondant à l'étude des concepts, méthodes et techniques permettant de modéliser et/ou d'acquérir les connaissances pour des systèmes réalisant ou aidant des humains à réaliser des tâches se formalisant a priori peu ou pas.* Historiquement située au sein de l'Intelligence Artificielle, l'IC est également au carrefour de plusieurs domaines de recherche telles que la linguistique, la psychologie cognitive, la sociologie ou encore la logique formelle. Points communs à l'ensemble de ces recherches : les **Connaissances**, leur stockage, leur consultation et leur maintenance, le raisonnement automatique, le partage, etc. Avec l'essor du Web sémantique, ces connaissances sont aujourd'hui principalement modélisées sous la forme d'ontologies. Ces ontologies sont des représentations consensuelles et formelles de connaissances (selon la définition de [Gruber 1993a]). Elles sont modélisées

sous la forme de hiérarchies (arbres, treillis...) de concepts définis au moyen de propriétés et de termes, et dotés d'instances, et de relations.

T. Berners-Lee, dans [Berners-Lee 2001], pose en 2001 la pierre fondationnelle de ce qui va devenir le Web Sémantique. Il s'agir d'établir *une extension du Web actuel dans lequel on donne à une information un sens bien défini pour permettre aux ordinateurs et aux individus de travailler en coopération.* Cela a conduit le World Wide Web Consortium (W3C) à mettre en place les recommandations suivantes. Tout d'abord, le Web devra pouvoir être décrit par des classifications précises, à l'aide d'ontologies exploitables par les machines et compréhensibles par les humains. Ces ontologies devront être exprimées au moyen d'un langage commun. Enfin, des moteurs de raisonnement devront permettre d'inférer sur les annotations d'après les axiomes déclarés dans ces ontologies.

Théorie du prototype et Gradients de Prototypicalité

À la base de la "théorie des prototypes", [Barsalou 1983, Rosch 1975c] formulent certaines hypothèses quant aux catégories (*i.e.* ensembles d'éléments de même nature). Tout d'abord, certaines catégories sont de meilleurs représentants de leur catégorie mère que d'autres (par exemple, *moineau* est plus représentatif de la catégorie des *oiseaux* que *poule*). Ensuite, l'appartenance de certaines catégories à une catégorie mère est incertaine (par exemple, une *télévision* fait-elle partie de la catégorie des *meubles* ?). De plus, la similarité au prototype peut varier pour les "non-membres" (par exemple, une *télévision* n'appartient pas à la catégorie *outil électro-portatif* mais elle est plus similaire au prototype de cette catégorie qu'un *chien*). Enfin, toutes les catégories d'une catégorie mère ne sont pas analogues. Selon [Rosch 1973], le prototype pour chacune des catégories est le meilleur exemplaire (le plus représentatif) selon un consensus social de l'ensemble d'individus (l'endogroupe) concerné.

Suivant cette proposition, nous avons dans un premier temps établi une nouvelle typologie des ontologies de domaine en distinguant deux niveaux : (1) les *Ontologies Vernaculaires de Domaine* (OVD) - les ontologies de domaine dont la conceptualisation est propre à un endogroupe, et (2) les *Ontologies Pragmatisées Vernaculaires de Domaine* - les OVD placées dans un contexte donné et sur laquelle nous appliquons la théorie des prototypes.

Afin de moduler cette représentativité au sein d'une hiérarchie conceptuelle, nous calculons un *gradient de prototypicalité conceptuelle* entre chaque concept et chacun de ses sous-concepts sur les trois dimensions sémiotiques : l'intension (suivant les propriétés), l'expression (suivant les termes dénotant le concept concerné), et l'extension (suivant les instances du concept concerné).

Nous évaluons également la différence de typicalité entre les termes dénotant un même concept au moyen du *gradient de prototypicalité lexicale* (par exemple, pour

un individu donné, le terme *clébard* est plus typique pour dénoter le concept *Chien* que le terme *cabot*).

Nous évaluons enfin la différence de typicalité entre les instances rattachées à un même concept au moyen du *gradient de prototypicalité extensionnelle*. Par exemple, pour un individu donné, l'instance *Jolly Jumper* est une instance plus typique du concept *Cheval de BD* que l'instance *Petit Tonnerre*.

Mesures sémantiques, mesures de similarité et de proximité

Selon de nombreux auteurs (*e.g.* [Blanchard 2008]), la structuration des concepts au sein d'une ontologie est associée aux relations sémantiques entre les concepts ; de nombreuses mesures, dites "sémantiques", tentent d'exploiter cette structure. Ces mesures permettent d'évaluer une liaison entre les concepts d'une même ontologie sur la base des relations qu'ils entretiennent. Aussi, les principales mesures sémantiques peuvent-être en quatre groupes :

1. les *mesures de type structurel*, telles celles de Rada [Rada 1989], de Resnik [Resnik 1995], de Leacock [Leacock 1998], ou encore de Wu & Palmer [Wu 1994] ;
2. les *mesures de type intensionnel*, c'est à dire reposant sur les propriétés des concepts, telle la mesure de Tversky [Tversky 1977] ;
3. les *mesures de type extensionnel*, c'est à dire reposant sur les instances rattachés aux concepts, telle la mesure de Jaccard [Jaccard 1901] ;
4. les *mesures de type expressionnel*, c'est à dire reposant sur les termes dénotant les concepts, telles les mesures de Lin [Lin 1998] ou de Jian & Conrath [Jiang 1997].

Cependant, comme le souligne [Blanchard 2008], l'interprétation de ces mesures est variable et rarement formellement explicitée. En particulier, il n'est pas rare de voir apparaître dans la littérature, dès lors qu'il s'agit de traiter des mesures sémantiques, les termes *proximité* et *similarité*, sans que le sens en soit réellement précisé. Pour établir cette distinction, on peut s'appuyer sur la théorie de la Gestalt [Koffka 1935]. Cette théorie de la perception comporte six "lois", dont une de similarité et une de proximité. La "loi de **similarité**" permet de regrouper les éléments qui nous paraissent semblables. Il peut s'agir de similitudes descriptives (perceptibles) ou fonctionnelles. La "loi de **proximité**" permet de regrouper des éléments qui apparaissent souvent ensemble, qui sont proches dans une même zone perceptive. C'est le cas des lettres qui forment un mot, des mots qui forment un syntagme. Il s'agit d'un regroupement présentant une certaine cohérence, et la plupart du temps inconscient.

À partir de ces deux "lois", nous avons développé deux mesures sémiotiques distinctes : une mesure de similarité SemioSem et une mesure de proximité ProxSem. La première se fonde uniquement sur les caractéristiques des concepts (propriétés, termes, instances) indépendamment de la structure de l'ontologie et du corpus de

textes. La seconde se fonde sur la représentation simultanée des deux concepts (relations, présence de termes dans un même document, présence simultanée des instances). Afin de valider ces deux mesures, nous avons effectué une comparaison avec les réponses données à un questionnaire proposé à un ensemble d'individus sur la similarité et la proximité d'un ensemble de couples de concepts.

TooPrag et Theseus

Nous avons été amenés au cours de cette thèse à développer deux prototypes qui permettent d'exploiter opérationnellement nos contributions scientifiques :

(1) Un **outil**, TooPrag (*A Tool dedicated to the Pragmatics of Ontology*), dont l'objectif est le calcul automatique des gradients de prototypicalité et l'administration/gestion de l'ontologie HSE-Tennaxia. Cette application, développée entièrement en Java, prend en entrée un index de corpus généré par *Lucène* et une ontologie de domaine implémentée en OWL.

(2) Un **moteur de recherche sémantique** utilisant les gradients de prototypicalité en effectuant une extension de requête à partir d'un terme saisi par l'utilisateur, et en parcourant l'ontologie en tenant compte des valeurs de gradients. Theseus est un moteur de recherche sémantique sur la base de textes réglementaires de la société Tennaxia. Il est fondé sur l'ontologie HSE-Tennaxia sur laquelle ont été calculés les différents gradients de prototypicalité (conceptuel et lexical).

Organisation de la thèse

Cette thèse se décompose en trois parties.

La **première partie**, généraliste, est constituée des chapitres 1, 2, 3 et 4. Elle présente l'état de l'art relatif aux domaines en lien avec nos travaux.
Le *chapitre 1* introduit le domaine du Web sémantique selon trois points de vue : (1) un point de vue global (architecture et principes généraux), (2) un point de vue industriel (brevets et projets), et (3) un point de vue applicatif (langages du Web sémantique).
Le *chapitre 2* est consacré aux ontologies et à la modélisation des connaissances. Nous détaillons en particulier les différentes étapes pour passer des "données" aux "connaissances" et construire ainsi une ontologie de domaine.
Le *chapitre 3* aborde la problématique de la catégorisation et de la prototypicalité, problématique introduite par la théorie des prototypes et notamment les travaux d'E. Rosch en psychologie cognitive.
Le *chapitre 4* constitue un état de l'art sur les mesures sémantiques dont nous présentons les principales en les regroupant suivant trois dimensions : intensionnelle, extensionnelle et expressionnelle.

La **deuxième partie** est constituée des chapitres 5, 6, 7 et 8. Elle est consacrée à la présentation de nos contributions scientifiques.
Le *chapitre 5* pose les bases de notre approche avec (1) la proposition d'une nouvelle typologie des ontologies de domaines, et (2) la présentation du cadre sémiotique dans lequel s'insèrent nos travaux.
Le *chapitre 6* présente trois mesures visant à évaluer la typicalité d'un élément par rapport à un autre : (1) un gradient de prototypicalité conceptuel (entre un concept et ses sous-concepts), (2) un gradient de prototypicalité lexical (entre un concept et les termes le dénotant), et (3) un gradient de prototypicalité extensionnel (entre un concept et ses instances).
Le *chapitre 7* expose deux nouvelles mesures sémantiques et sémiotiques : (1) une mesure de similarité, SemioSem, et (2) une mesure de proximité, ProxSem.
Le *chapitre 8* présente les expérimentations mises en place pour évaluer nos différentes propositions.

La **troisième partie** est constituée des chapitres 9, 10 et 11. Elle est consacrée à la présentation de nos contributions applicatives.
Le *chapitre 9* présente la suite logicielle de la société Tennaxia, notamment la partie *Veille & Conformité* et son moteur de recherche sur la base des textes réglementaires.
Le *chapitre 10* expose les différentes composantes de l'ontologie HSE-Tennaxia, ainsi que les ressources employées pour son élaboration.
Le *chapitre 11* décrit d'une part le prototype élaboré pour mettre en œuvre le calcul de nos mesures TooPrag, et d'autre part l'application réalisée en Recherche d'Information Sémantique fondée sur des ontologies de domaine et nos travaux sur la prototypicalité : Theseus.

Partie I

Etat des lieux

L'objet de cette thèse est (1) la réalisation d'une ontologie du domaine Hygiène-Sécurité-Environnement (HSE), et (2) son exploitation dans le cadre de la recherche d'information au sein d'une base de textes réglementaires. Nous y mettons en œuvre des outils du Web Sémantique, au travers des ontologies, de leur personnalisation en s'inspirant des travaux de psychologie cognitive relatifs à la catégorisation, et des mesures sémantiques. Le Web sémantique, en coopération avec le Web social sur la Toile, a pour objectif d'enrichir les données par des représentations sémantiques. Il permet, outre l'émergence d'une forme d'intelligence collective, de formaliser et d'assembler des informations jusque-là disparates et hétérogènes, et ce notamment au moyen d'ontologies. Ces dernières sont des spécifications formelles de conceptualisations partagées [Gruber 1993a] ; en d'autres termes une formalisation consensuelle d'une catégorisation de l'univers cognitif d'un ensemble de personnes. Cette formalisation se fonde sur les trois dimensions du triangle sémiotique : les propriétés, les termes et les instances. Les psychologues ont montré que ce processus de catégorisation est naturellement présent chez l'humain dès son plus jeune âge. Il commence par rassembler les objets de son univers par forme, puis par couleur, puis par fonction... Ses connaissances s'organisent principalement en catégories. Ces mêmes connaissances s'articulent autour d'un prototype, l'élément le plus représentatif de sa catégorie. Il est possible d'intégrer cette particularité au sein d'ontologies afin de les personnaliser, et obtenir ainsi une modélisation de la connaissance plus proche de celles des utilisateurs. Enfin, le Web des données, ou Web 3.0, avec sa croissance quasi exponentielle des données et les nouveaux usages qu'il engendre, nécessite la création de nouveaux outils. Les mesures sémantiques occupent aujourd'hui une place prépondérante dans ces nouveaux usages, telles les problématiques d'alignement et d'extraction de concepts/relations en Ingénierie des connaissances, ou encore les problématiques d'indexation de documents en recherche d'information.

—

Chapitre 1

Web sémantique

1.1 Introduction

Les premiers moteurs de recherche sont apparus dans les années 1980 au sein d'Intranet afin de pouvoir traiter des documents en provenance de systèmes d'information militaires, universitaires ou privés. Depuis cette époque, ces moteurs de recherche qualifiés de syntaxiques n'effectuent que ce que l'on pourrait assimiler à une sorte de reconnaissance de formes. Un utilisateur saisit, par exemple, le mot *chats* (au pluriel), et le système va - en réalité - à la recherche, non pas de tous les documents parlant de chats, mais de ceux contenant la suite de symboles formés par les lettres *c, h, a, t* et *s*. Point de documents contenant le mot *chat* (au singulier), ce n'est pas la même forme, et encore moins de réponse contenant le mot *minou* (synonyme) ou *siamois* (hyponyme). Cette recherche pouvait se faire au moyen de nombreuses options avancées, telle une recherche dans le titre d'un document, ou à l'aide d'opérateurs booléens, etc.

Quelques méthodes lexicales ont ensuite fait leur apparition avec l'extraction de la racine des mots recherchés. Il est alors possible d'obtenir des documents parlant de *chat*, de *chatte* et de leurs formes au pluriel pour une même recherche sur le mot *chat*. Mais toujours point de documents contenant le mot *minou* ou *siamois* dans les réponses.

Ce n'est que dans les années 1990 que les moteurs de recherche commencèrent à indexer les documents présents sur la Toile. Ce changement de paradigme s'accompagne également d'un changement d'échelle en terme de volume de données à traiter. Il devient rapidement nécessaire de les classer et les catégoriser en ajoutant de l'information sur l'information au moyen de méta-données. Elles permettent de décrire les données dans les données. ; grâce à elles, la suite de caractères c.h.a.t se rapproche enfin d'un "chat". Le Web va non seulement contenir des données mais aussi certaines informations sur le sens qu'il faut leur donner. Ainsi naquit le *Web Sémantique* ; tout au moins dans la tête de Tim Berners Lee, inventeur du World Wide Web et président aujourd'hui du World Wide Web Consortium (W3C) - organisme de standardisation des technologies du Web. Le but ultime du Web Sémantique

est affiché par ses créateurs : permettre aux ordinateurs de comprendre les données contenues dans les documents. Le Web Sémantique, qui se transformera au fil des années, en *Web des données* consiste donc à ajouter une couche de connaissances aux symboles contenus dans les documents circulant sur le Web. Cette connaissance va pouvoir s'exprimer sous forme de méta-données (structure du document, auteur et informations sur celui-ci, éditeur et informations sur celui-ci, contexte de publication, etc.). Des conventions sont définies pour pouvoir structurer ces méta-données. Pour donner sens aux mots, d'autres mots seront ajoutés suivant d'autres conventions. Si on se limite au système de tags, il s'agit alors de compléter des suites de symboles par d'autres suites de symboles - lesquelles pourraient également nécessiter des suites de symboles supplémentaires pour en compléter le sens et ainsi de suite dans une chaîne sans fin.

Il est cependant possible de parfaire ce système avec de l'intelligence de synthèse, de la logique, du raisonnement, en ajoutant une organisation des connaissances en amont sous forme d'ontologies. Ces dernières, lointaines cousines de leur homologue philosophique aristotélicienne, sont des représentations consensuelles et formelles de connaissances (selon la définition de [Gruber 1993a]). Formelles pour pouvoir être comprises et utilisées tant par des hommes que par des machines dans la lignée de l'IA classique, consensuelles car elles sont le fruit d'une intelligence collective et d'un certain consensus sur un domaine. Avec un tel outil intégré dans un moteur de recherche, nous pouvons obtenir non seulement des documents contenant le mot *chat* (sous toutes ses déclinaisons syntaxiques), mais également des réponses contenant le mot *minou* (synonyme) ou *siamois* (hyponyme), ou encore *félin* (hyperonyme), sans oublier les documents parlant de *souris* et de *croquettes* (car un chat mange des souris et des croquettes). Cette fois, ce ne sont plus des formes qui complètent d'autres formes comme dans le cas précédent, mais des connaissances intégrées à un système de recherche d'informations.

Il devient dès lors pleinement légitime de faire la remarque suivante : "MES" réponses obtenues dans un tel système sont donc tributaires de "LEUR" vision du monde. En toile de fond, se posent les questions d'une connaissance normative et du choix des experts définissant cette norme. Aujourd'hui, avec RDF et ses milliards de triplets, chacun est en effet libre de fixer sa grammaire et de définir ses propres relations. OWL tente de remédier à cette multiplicité sémantique en organisant quelques correspondances. L'adjonction de thésaurus permet de manipuler différentes dénominations de mêmes éléments. Mais si une entreprise publie sur son site Internet la liste de son personnel en distinguant noms de famille, prénoms et initiales, tandis qu'une autre décide de rassembler toutes ces informations sous l'appellation *noms*, les données de l'une seront difficilement exploitables par l'autre. Actuellement, il existe une multitude de langages pour décrire et organiser des données, de RDF à OWL-Lite, DL ou Full, en passant par les micro-formats RDFa et autres thésaurus en SKOS. Face à cette absence d'organisation des ressources sur la toile, il est devenu nécessaire de définir un nouveau standard. Mi-2010, Tim Berners-Lee et le

W3C annoncent le lancement d'un nouveau standard baptisé *Rule Interchange Format* (RIF) [1]. Le but de cette nouvelle norme est d'augmenter l'interopérabilité des données entre différents sites. Selon les chercheurs du MIT, *dans sa version actuelle, le web est un fichier texte géant dans lequel on peut rechercher différents mots. Le Web sémantique se rapproche plus d'une base de données où chaque information est catégorisée et où de nouvelles requêtes peuvent combiner ces catégories de toutes les manières possibles. Le problème est qu'il appartient à chaque personne publiant sur la Toile d'organiser et d'étiqueter ces informations* [2]. Avec l'ajout d'une couche *Rule Interchange Format*, il sera possible de définir une règle stipulant, par exemple, que toutes les informations précédant le premier espace dans un répertoire d'employés appartiennent à la catégorie *prénom*, et toutes celles situées après le dernier espace correspondent au *nom de famille*. Ce type de règles devraient également permettre aux utilisateurs d'exploiter un ensemble de données complexes sans pour autant utiliser de bases de données.

1.2 Architecture du Web sémantique

Le *Web sémantique* est une idée relativement récente qui remonte à 2001. T. Berners-Lee caractérise le Web sémantique comme étant *une extension du Web actuel dans lequel on donne à une information un sens bien défini pour permettre aux ordinateurs et aux individus de travailler en coopération* [Berners-Lee 2001]. Le W3C met en place les recommandations suivantes :

- description du web par des classifications précises, à l'aide d'ontologies exploitables par les machines et compréhensibles par les humains ;
- utilisation d'un langage commun pour exprimer les ontologies et décrire des annotations utilisant leurs termes ;
- création de moteurs de raisonnement permettant d'inférer sur les annotations d'après les axiomes déclarés dans les ontologies.

En conséquence, le Web sémantique peut se considérer suivant deux facettes : premièrement comme solution à l'insuffisance du Web originel et à l'énergie déployée autour du référencement des sites, et deuxièmement comme axe de développement pour passer du Web originel vers le Web des connaissances. Avec ces recommandations, le W3C propose une représentation pyramidale des langages associés aux ontologies. Elle est basée sur une infrastructure en couches, dont seules les premières peuvent être actuellement considérées comme relativement stabilisées.
Ces sept couches peuvent être décrites succinctement comme suit :

- **couche 1** : couche de base assurant l'interopérabilité syntaxique. *Unicode* est une norme qui permet de coder les caractères de n'importe quelle langue naturelle. *Uniform Resource Identifier* (URI) permet d'identifier de façon unique toutes les ressources par un adressage unique, standard et universel ;

1. http ://www.w3.org/TR/2010/REC-rif-core-20100622/
2. Article du 22 juin 2010 sur http ://www.atelier.net/

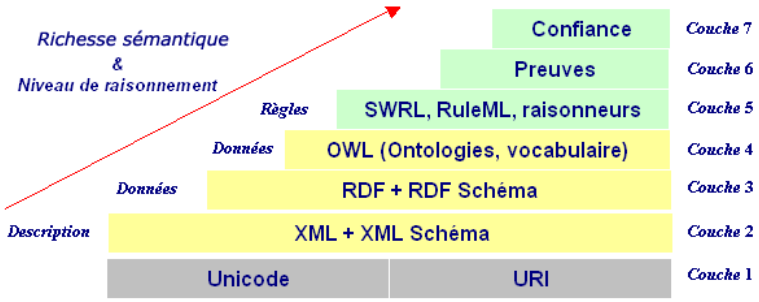


FIGURE 1.1 – Représentation pyramidale du web sémantique [Berners-Lee 2001].

- **couche 2** : *eXtended Markup Language* (XML) est un langage de balisage extensible qui a la possibilité, grâce aux espaces de noms (Namespaces, NS) et aux définitions de schémas (XML Schéma), d'assurer l'interopérabilité avec les bases XML existantes.
- **couche 3** : *Resource Description Framework* (RDF) et *Resource Description Framework Schema* (RDFS) sont des vocabulaires XML qui ont pour rôle d'affecter des méta-données à la description des ressources, les URI.
- **couche 4** : une formalisation consensuelle des connaissances ;
- **couche 5** : écriture des règles de déduction entre ontologies ;
- **couche 6 et 7** : production de preuves des déductions réalisées précédemment par la couche 5 dans le but d'augmenter le niveau de confiance pour les utilisateurs. Celle-ci ne pourra être obtenue qu'avec l'utilisation de moyens permettant d'authentifier les différentes données utilisées (ontologies, méta-données, annotations, etc.). Aujourd'hui, les mécanismes d'authentification sont connus (signature électronique) ; en revanche, la production de preuves par déduction relève de domaines complexes.

Bien que présentes dans ce schéma, les couches 5, 6 et 7 font actuellement l'objet de recherches actives et ne sont pas insérées dans les recommandations du W3C.

1.3 Web 2.0 et plus

Web 1.0, 2.0, 3.0... Voilà autant de dénominations qui fleurissent depuis quelques années, versionnage non-officiel censé représenter les différentes évolutions du Web. La figure 1.2 représente de manière synthétique ces versions multiples. Fred Cavazza livre au sein de son blog [3] une analyse très pertinente de cette évolution.
L'informatique est née solitaire avec l'avènement de l'ordinateur individuel, du Personal Computer (PC). Elle ne devient communautaire qu'à partir des années 1970 et les premiers balbutiements des réseaux informatiques.
Le Web premier, que certains qualifient de 1.0, était avant tout une plate-forme de

3. http ://www.fredcavazza.net

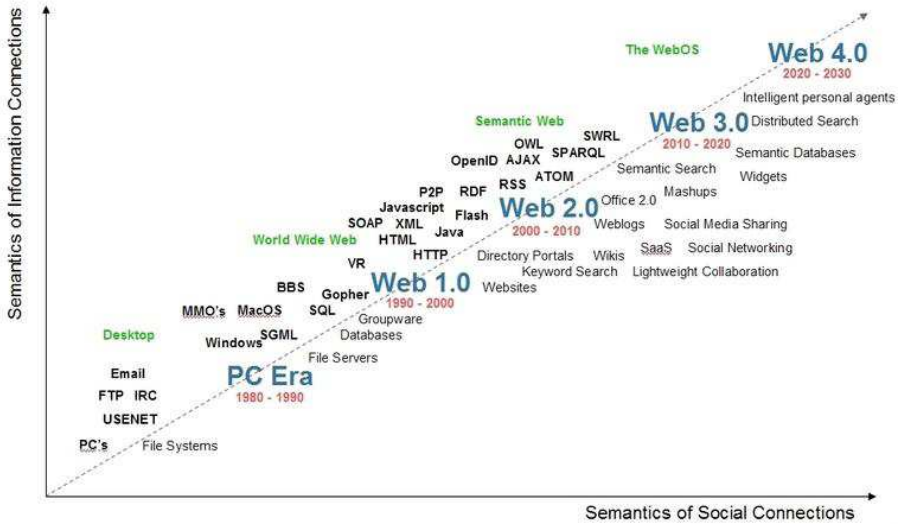


FIGURE 1.2 – Résumé de l'histoire du Web, selon Radar Networks and Nova Spivack.

documents. Le contenu était placé sur la Toile par des individus, ou de gros acteurs du domaine, à destination du reste du monde. Placé dans une relation de type *1-n*, chaque acteur maîtrisait l'ensemble de la chaîne d'information.
Le Web 2.0 est le Web pour les individus avec le Web des réseaux sociaux. Il dispose désormais d'une palette bien plus large de sources d'informations et de services marchands. Cette chaine est désormais éclatée et déstructurée : certains acteurs s'occupent de la valorisation, d'autres du paiement en ligne, d'autres encore de la livraison, etc.
Le Web 3.0, plus connu sous le terme de *Web des données* ou encore *Linked data*, est avant tout une plate-forme pour les données. Il repose sur quelques principes simples : utilisation des URIs et non des URLs pour un accès universel (aussi bien pour les utilisateurs que pour les machines), structuration de l'information sur la ressource à l'aide de métalangages comme RDF, enfin liens entre ces informations via les URIs. L'objectif *in fine* est de créer une méta-base de données qui regrouperait toutes ces bases de données, comme le montre la figure 1.3. Parmi ces immenses base de données, citons DBPedia, une base de connaissance contenant plus d'un milliard de triplet RDF, décrivant plus de 3,4 millions d'éléments (personnes, lieux, albums de musiques, films, jeux vidéos, organisations, espèces, maladies).

Les inventeurs du Web sémantique auraient bien aimé le voir comme le Web 2.0. Mais sémantiser le Web est une problématique complexe encore très ouverte. Aussi, le Web sémantique est bien plus qu'une affaire de version. Présent par l'intermédiaire des méta-données dès le Web 1.0, il n'en finira pas d'accompagner cette évolution de la Toile, que ce soit dans sa version 2.0, 3.0 et plus encore.

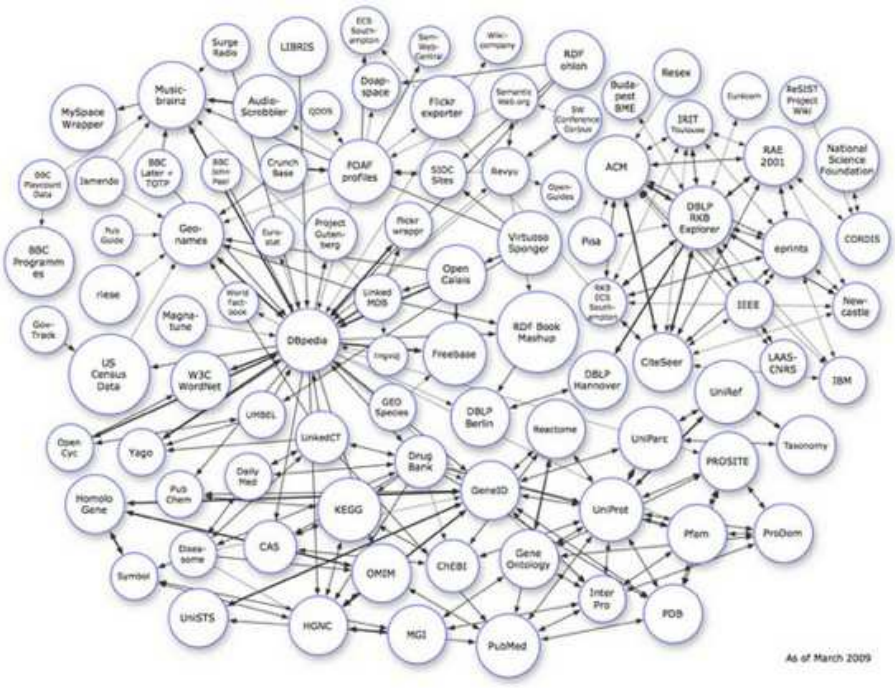


FIGURE 1.3 – Data Ecosystem (2009).

1.4 D'un point de vue industriel

Si conférences et articles offrent une certaine visibilité en matière de productions scientifiques, il n'en est pas de même pour tout ce qui est du ressort de la recherche privée. Il ne s'agit pas d'établir dans la présente section un état exhaustif de tout ce qui se fait dans le domaine du Web sémantique d'un point de vue industriel, mais plutôt de présenter les résultats d'une veille technologique réalisée dans le cadre de cette thèse portée par le projet industriel de la société Tennaxia. Cette veille s'est orientée principalement vers les domaines des moteurs de recherche sémantique et des ontologies. Elle a été réalisée sur la base des brevets d'une part, mais aussi sur les différents produits logiciels mettant en place des systèmes de recherche d'information à visée sémantique.

1.4.1 Une utilisation progressive des technologies

On peut distinguer au niveau industriel cinq niveaux d'appropriation du Web Sémantique [Simon 2007]. Ces cinq niveaux peuvent être décrits de la manière suivante :

1. Niveau *Social - Tagging - Folk*. Ce niveau repose sur la production de mots clés par les utilisateurs afin de qualifier des contenus. Ce genre de système est facile à créer, ne nécessite pas d'algorithmes ou d'ontologies à maintenir. Cependant, il très basique et comporte trop d'approximations. De plus, il manque d'outils

de normalisation statistiques et linguistiques.

2. Niveau *Statistiques*. Ce niveau repose sur des calculs d'occurrences et de co-occurrences pour définir les mots clés qualifiant des contenus non structurés. Ce genre de système utilise des algorithmes purement mathématiques et statistiques, capables de fonctionner sur de larges échelles, indépendants du langage, et produisant rapidement des agrégats et des indicateurs... Cependant, il n'offre pas de compréhension du contenu, et se trouve tributaire des volumes de données sans permettre de trouver finement ce qui est recherché.
3. Niveau *Linguistiques*. Ce niveau repose sur l'extraction d'entités nommées la plus fine possible. Ce genre de système utilise la détection de la langue, l'extraction d'entités, la mise en correspondance à travers des tables de synonymes... Cependant, il demande des ressources machines plus importantes, ainsi qu'un effort et une maintenance lourde pour chaque langue traitée.
4. Niveau *Web sémantique*. Ce niveau repose sur la mise en relation de contenus à travers des descripteurs et des usages convergents. Ce genre de système génère des requêtes plus précises. Ce système ne requiert pas trop de ressources machines, et fonctionne tant pour les données structurées que non structurées. En revanche, l'utilisation d'ontologies pose le problème de leur construction d'une part, puis de leur maintenance d'autre part.
5. Niveau *Intelligence Artificielle*. Ce niveau réutilise l'ensemble des approches précédentes pour que l'application interagisse intelligemment et de façon évolutive avec ses utilisateurs. Ce genre de système fonctionne bien dans des domaines restreints. Il répond correctement aux questions, raisonne et apprend ... Cependant, il demande des ressources machines importantes, et reste encore difficile à programmer et à généraliser.

1.4.2 Les brevets

Les brevets ont été recherchés sur la base *espacenet*[4]. Contrairement aux articles scientifiques, le dépôt d'un brevet ne requiert aucune validation de la part d'un collège d'experts dûment agréés par une quelconque organisation indépendante. Aussi, il nous faut distinguer au moins trois types de brevets : les brevets "fantaisistes", les brevets de barrage (la technologie n'est pas encore développée, mais le jour où elle le sera l'entreprise déposante aura pris date), et enfin les brevets de protection pour des applications aujourd'hui bien en place. Nous avons orienté nos recherches en direction du domaine des ontologies et des technologies gravitant autour de leur utilisation telle que personnalisation des ontologies, extension de requêtes par ontologie, moteur de recherche sémantique.

Il ressort de cette veille que :

- il existe une très forte activité industrielle en terme de brevets autour des ontologies, du sémantique... le tout lié - en grande partie - au développement

4. http ://fr.espacenet.com/

de l'informatique mobile et de ses outils ;
- la qualité des dossiers déposés est très variable d'une entreprise à l'autre, avec cependant des rapports très complets (*IBM, France Telecom, Nokia, Motorola*) qui donnent une assez bonne idée de la direction prise par ces sociétés et des moyens parfois considérables mis en œuvre ;
- l'Asie offre une production impressionnante (de manière générale, propriétaire de 3/4 des brevets déposés au monde), les Etats-Unis se positionnent après, et l'Europe non loin derrière. D'un point de vue quantitatif, on dénombre environ 1600 brevets dans le monde ayant traits aux technologies du Web sémantique (dont environ 17% de français).

Parmi l'ensemble de ces brevets, il est possible d'en retenir au moins deux qui nous paraissent paradigmatiques des recherches industrielles actuelles (le premier prend acte assez tôt, le second par sa provenance marque l'émergence de la Chine dans ce domaine) :

- *Système et procédés d'indexation et de recherche à extension de requêtes, moteurs d'indexation et de recherche.* Brevet n°FR2835334, publié le 01/08/2003 par France Telecom. Ce système d'indexation et de recherche comporte des moyens de stockage d'une base d'indexation, des moyens d'indexation de ressources pour la création et la mise à jour de la base d'indexation, des moyens de recherche de ressources adaptés pour interroger la base d'indexation à partir d'une requête, et des moyens d'extension de requêtes pour l'obtention d'une requête étendue, à partir d'une requête initiale formulée par un utilisateur et comportant des termes initiaux, par l'ajout à cette requête initiale de termes voisins des termes initiaux. Les moyens d'extension comportent en outre des moyens de limitation de l'extension de la requête initiale par l'ajout à celle-ci uniquement de termes voisins des termes initiaux non généraux, c'est-à-dire ne comportant pas un trop grand nombre de termes voisins. Des moyens de généralisation de l'indexation peuvent également être mis en œuvre selon l'invention.
- *Intelligent retrieval system and method based on domain ontology.* Brevet n°CN101582073, publié le 18/11/2009 par Beijing Zhongjikehai Technolog, une société chinoise spécialisée dans les systèmes de recherche d'information. L'invention concerne le champ de la recherche documentaire en chinois, en particulier une méthode de recherche d'information fondée sur les ontologies de domaine et un système de recherche intelligent utilisant la méthode. Le système comporte : (1) un module d'inférence d'ontologie utilisé pour analyser une requête en langage naturel saisie par des utilisateurs, (2) un module d'indexation, (3) un module pour traiter la question posée, et (4) un module de linéarisation de résultats. En outre, le système comporte également une bibliothèque de domaine, un entrepôt de données, et l'index de la base de données. Le système de recherche d'information basée sur les ontologies de domaine utilisent pleinement les concepts de l'ontologie et leurs interdépendances, analysent les demandes de l'utilisateur, optimisent les résultats de la

recherche. La qualité de l'information retournée est ainsi largement améliorée par rapport aux systèmes existants.

En matière d'utilisation d'ontologies au sein de systèmes de recherche d'information, nous pouvons également citer :

- Querying data and an associated ontology in a database management system (2008, IBM, protection dans environ 140 pays) - WO 2008/088722A2
- Conceptual reverse query expander (2009, Raytheon Company, protection dans environ 140 pays) - WO 2009/108587A1
- System for extending data query using ontology and method therefore (2007, Blakely sokoloff taylor & zafman, US) - US2007/0150458A1
- Information retrieval (2008, Nixon, US) - US2008/0235203A1

1.4.3 Les Projets industriels

Des organismes, privés comme publics, réfléchissent au moyen d'accéder à l'ensemble des données liées en un unique point. Citons par exemple le cas du projet *Uberclic*[5] qui établit périodiquement une consolidation des données présentes dans des entrepôts de données tels que DBpedia, Geonames, Musicbrainz ou encore TheMovieDB. Il offre ainsi (1) l'avantage d'un mapping entre ressources hétérogènes (par exemple : je cherche quelque chose sur les Beatles, et je peux aussi avoir des informations sur Liverpool, etc.), mais également (2) le bénéfice de la mise à jour quasiment en temps réel de l'une de ces bases, avec ses répercussions sur ma recherche d'information. Cela débouche sur une sorte de base de données universelle et non structurée, avec un seul point d'entrée. Reste alors le problème de la pertinence des données enregistrées d'une part, et du taux de subjectivité des informations contenues dans cette base.

Dans cette myriade de projets s'inscrivant dans des projets de Web sémantique, il en est un qui a retenu tout particulièrement notre attention : GoPubMed[6] conçu par Transinsight[7]. Ce projet, propriétaire, très proche de celui développé dans le cadre de cette thèse, est un moteur de recherche sémantique fondé sur l'ontologie *Gene Ontology*[8], la terminologie Medical Subject Headings (MeSH)[9] et la base de connaissances des protéines Universal Protein Resource (UniProt). Ce moteur de recherche fonctionne sur la base de données bibliographiques MEDLINE, rassemblant des citations et des résumés d'articles de recherche biomédicale publiés dans près de 6 000 revues. Il est également proposé une recherche globale sur le Net avec le projet GoWeb[10], avec la même base sémantique.

5. http ://uberblic.org/blog/
6. http ://www.gopubmed.org
7. http ://www.transinsight.com/
8. http ://www.geneontology.org/
9. http ://www.ncbi.nlm.nih.gov/mesh
10. http ://www.gopubmed.org/web/gopubmed/1 ?WEB0h55rld7tq8trIiI1I00h001000j100203010

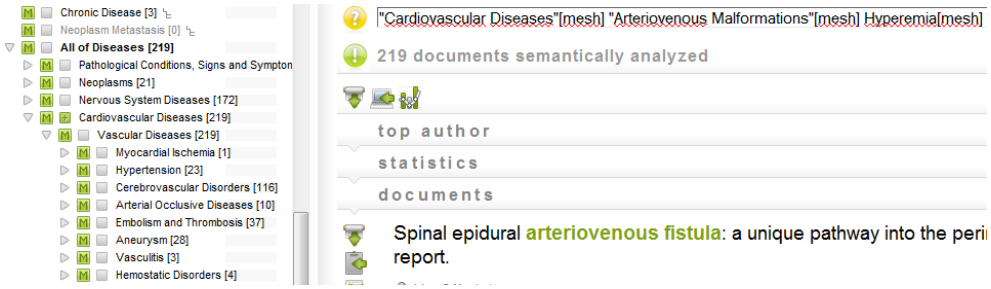


FIGURE 1.4 – Interface d'interrogation de GoPubMed.

La recherche s'effectue sur quatre critères possibles : *what, who, where, when.*
La figure 1.4 illustre les possibilités offertes par le critère de recherche *What ?*. L'utilisateur navigue au sein d'une hiérarchie de concepts réalisée à partir des bases de connaissances citées plus haut. Pour chaque concept, il est possible de :

- étendre la requête pour la recherche de documents avec ou sans ce concept ;
- obtenir une description du concept (par exemple pour le concept *Serum* : The clear portion of blood that is left after blood coagulation to remove blood cells and clotting proteins) ;
- connaître la liste des termes synonymes avec lesquels la recherche sera étendue (par exemple pour le concept *Serum* : Blood Serum et Serums) ;
- visualiser le chemin dans la base de connaissances (MeSH...) de départ ;
- d'accéder à un article Wikipédia correspondant ;
- voir l'évolution en temps réel du résultat, et ce au fur et à mesure de la sélection ou non d'un concept.

Le critère de recherche *Who* concerne les auteurs des documents de la base PubMed sur lesquels porte la recherche (environ 3 millions d'auteurs recensés). Pour chaque auteur, il est rappelé le nombre d'articles et dans certains cas la bibliographie complète. Le critère de recherche *Where* concerne aussi bien la publication où apparaît l'article (avec une catégorie spéciale pour les *High impact factors journals*), que le lieu d'affectation du ou des auteurs. Enfin, le critère de recherche *When* concerne les dates de publication avec une granularité qui va de l'année jusqu'au jour.

D'un point de vue sémantique, les requêtes sont, par exemple, de la forme : *Proteins[mesh] (France[mesh] OR Spain[mesh]) NOT (2002[time] OR 2003[time])*. Cette dernière retourne tous les articles dont le titre ou le résumé contient le terme exact (*Proteins* ou l'un des ses synonymes) ou l'un des concepts de sa descendance (ou leurs synonymes), et les concepts *France* ou *Espagne*, et qui n'ont pas été publiés dans les années 2002 et 2003.

Enfin, notons que ce prestataire fournit également un éditeur d'ontologie muni de nombreux outils de fouille de textes afin de permettre aux utilisateurs de maintenir,

aisément et au sein d'un même univers cognitif, leur ontologie.

1.5 Les langages du Web sémantique

Les langages de formalisation sont les éléments centraux sur lesquels repose le Web sémantique. Dans les sections suivantes, nous développons les différents types de formalisation.

1.5.1 Resource Description Framework (RDF)

Resource Description Framework (RDF) est un dialecte XML - ce qui signifie que RDF peut s'écrire en XML, mais aussi avec d'autres syntaxes comme *N3*, *N-Triples* ou encore *Turtle*. RDF est un modèle conceptuel, normalisé par le W3C permettant de décrire des ressources, simplement et sans ambiguïté sous la forme de déclarations <*sujet*> <*prédicat*> <*objet*> (proximité avec le langage naturel et le triplet sujet - verbe - complément) [ENS 2005]. Prenons l'exemple d'une photo de *Chiang Wei* présente sur le site web *china.org*. La figure 1.5 représente les deux manières de la décrire avec RDF.

RDF / langage naturel

```
<La photo> <a> <un auteur>
<La photo> <est publiée sur> <China>
<Un auteur> <est> <Chiang Wei>
<Chiang Wei> <est> <photographe>
<Chiang Wei> <travaille pour> <China>
<China><est à l'adresse><www.china.org>
```

RDF / XML

```
<?xml version="1.0" encoding="iso-8859-1"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
xmlns:dc="http://purl.org/metadata/dublin_core#">

 <rdf:Description about="www.china.org"/>
    <dc:title>Photo</dc:title>
    <dc:contributor>Chiang Wei</dc:contributor>
 </rdf:Description>

</rdf:RDF>
```

FIGURE 1.5 – exemples avec RDF / langage naturel et RDF / XML.

RDF et XML sont complémentaires. RDF, comme méta-modèle de données basées sur XML, spécifie leur sémantique d'une manière standardisée et inter-opérable. La syntaxe XML permet l'encodage, le transport et le stockage de fichiers RDF (avec l'avantage d'un format entièrement normalisé et non-propriétaire, des outils très disponibles, etc.) Ce système repose donc sur trois piliers :

- **RDF**, description de ressources Web (méta-données) ;
- puis **RDF Schéma**, vocabulaires de description (ontologies) ;
- une **syntaxe**, ici XML, pour l'échange des méta-données et des schémas.

FIGURE 1.6 – Triplet RDF.

RDF procède par descriptions des savoirs (données comme méta-données) à l'aide d'expressions dont la structure est fixée [Lacot 2005]. Comme le montre la figure

1.6, la composition fondamentale de toute expression en RDF est une collection de triplets sous la forme <*sujet*> <*prédicat*> <*objet*>. Chaque triplet est représenté par un arc prédicat orienté du nœud source sujet vers le nœud destination objet.

L'ensemble de ces triplets forme un graphe orienté, appelé *graphe RDF*. RDF travaille avec des données élémentaires : les **ressources**, les **propriétés** et les **déclarations** [Dubost 1999]. Ces trois axes (*cf.* figure 1.7) sont définis comme suit :

- **les ressources** : toute entité d'information pouvant être référencée en un bloc, par un nom symbolique (littéral) ou un identificateur. L'identificateur est obligatoirement un *Uniform Resource Identifier* (URI).
- **les propriétés** : un aspect, une caractéristique, un attribut, ou une relation spécifique utilisée pour décrire une ressource. Chaque propriété possède une signification spécifique, définit ses valeurs permises, les types de ressources qu'elle peut décrire, et les relations qu'elle entretient avec les autres propriétés.
- **les déclarations** : une ressource spécifique associée à une propriété définie ainsi que la valeur de cette propriété pour cette ressource est une déclaration RDF. Ces trois parties individuelles d'une déclaration sont appelées, respectivement, le sujet, le prédicat, et l'objet. Cet objet peut tout aussi bien être une autre déclaration RDF (nous parlons alors de réification) qu'un URI ou une valeur typée (littéral).

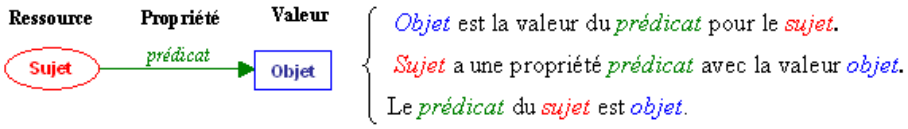


FIGURE 1.7 – Dénominations et rôles.

Un URI permet d'identifier de façon unique toutes les ressources, qu'elles proviennent du web (de type texte, image, vidéo, etc.) ou non (des objets, des personnes, etc.) [Lemire 2006]. Un URI contient :

- un protocole (*http, mailto, ftp, etc.*) ;
- un domaine (par exemple *chat.tv*) ;
- et un chemin (par exemple */mesfichiers/index.html*).

Un fichier RDF/XML possède un élément racine : **rdf :RDF** En règle générale, les documents XML représentant du contenu RDF n'ont pas de DTD mais utilisent les espaces de nommage. D'une manière générale, pour toute déclaration RDF avec utilisation d'URI, nous avons le tableau décrit dans la figure 1.8.
Pour la propriété, la première partie de l'URI (avant le #) devient un *espace de nommage*, et la deuxième (le fragment) devient un nom d'élément [11]. À l'intérieur

11. Nous supposons jusque-là que l'URI du prédicat contient un symbole #, ce qui nous permet de la décomposer en deux parties : le préfixe et le fragment. Mais que faire si l'URI prend plutôt la forme protocole + domaine + chemin simple (sans ancrage) ? Dans un tel cas, nous pouvons

FIGURE 1.8 – RDF et URI.

de l'élément racine vont s'imbriquer ou se juxtaposer des déclarations, représentées généralement par les éléments *rdf :Description* avec l'attribut *rdf :about* dont la valeur sera le sujet. Prenons comme exemple la phrase : *ce maillon a l'air coûteux* [Martin 2006]. Nous pouvons la décomposer suivant les valeurs données par la table 1.5.1.

Sujet	Sujet	Ressource	*ce maillon*
Verbe	Prédicat	Propriété	*a l'air*
Complément	Objet	Littéral	*coûteux*

TABLE 1.1 – Décomposition en sujet-prédicat-objet.

Avec une traduction en syntaxe RDF/XML :

```
<?xml version="1.0" encoding="UTF-8?>
<rdf:RDF xmlns:rdf= "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:fs="http://ma_description.fr/schema/">
       <rdf:Description rdf:about="http://www.objet.com/Maillon">
            <fs:a_l_air> couteux </fs:a_l_air>
       </rdf:Description>
  </rdf:RDF>
```

Ici la propriété est une chaîne de caractères, mais les propriétés peuvent aussi être des ressources, des valeurs d'attributs, ou une autre description, et dans ce cas un objet devient sujet.

[Lemire 2006] synthétise RDF par les règles suivantes :
- RDF peut être utilisé pour représenter tout objet ;
- RDF peut être traité par une machine ;
- RDF est composé de triplets (sujet, prédicat, objet) ;

choisir de retenir protocole + domaine comme préfixe et chemin comme fragment.

- le sujet est toujours identifié par un URI (factice ou réel) ;
- le prédicat est toujours identifié par un URI, sans aucune exception ;
- l'objet est soit un URI, soit une valeur explicite (une chaîne de caractères, par exemple) ;
- RDF peut être représenté en XML.

1.5.2 Resource Description Framework Schema (RDFS)

Resource Description Framework Schema propose un modèle de description de vocabulaires RDF, à base de classes et de propriétés. Parmi ces classes et propriétés, nous trouvons :

- la classe *Class* - un ensemble de plusieurs objets (par exemple, la classe des téléphones) ;
- la propriété *subClassOf* qui permet de définir qu'une classe est un sous-ensemble d'une autre classe (par exemple la classe *voiture* est sous-classe de la classe *Véhicule*) ;
- la classe *Ressource* qui est la classe parente de toute chose, en sachant que tout est ressource sauf la classe *Literal* (valeur typée $\neq$ concept) et que toute classe est une sous-classe de la classe ressource ;
- la propriété *range* permettant d'indiquer le champ d'application d'une propriété (par exemple la propriété *sont gardés par* s'applique aux classes *berger* et *nounou*) ;
- la propriété *domain* permettant de spécifier quelles sont les classes auxquelles nous pouvons affecter telle ou telle propriété (par exemple la classe *mouton* peut être l'objet de la propriété *sont gardés par*).

A titre d'exemple, une ressource *chat* sous-classe de la ressource *félin* se traduit par le code suivant :

```
<?xml version="1.0" encoding="UTF-8?>
<rdf:RDF xmlns:rdf= "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
       <rdf:Description rdf:ID="félin">
            <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
       </rdf:Description>
       <rdf:Description rdf:ID="chat">
            <rdf:type   rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
            <rdfs:subClassOf rdf:resource="#félin"/>
       </rdf:Description>
</rdf:RDF>
```

Nous pouvons écrire ce schéma d'une manière plus abrégée, à savoir :

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf= "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
```

```
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
   <rdfs:Class rdf:ID="félin" />
   <rdfs:Class rdf:ID="chat">
   <rdfs:subClassOf rdf:resource="#félin"/>
</rdfs:Class>
</rdf:RDF>
```

Une collection de classes (écrite typiquement pour un domaine ou un but spécifique) est appelée *schéma*. Les classes sont organisées en hiérarchie, et offrent une extensibilité grâce aux sous-classes. Les schémas peuvent s'intégrer au sein d'un même fichier RDF et ne s'excluent pas mutuellement grâce à l'utilisation des *espaces de nommage* [Dubost 1999].

RDFS décrit donc les ressources en termes de *classes*, *propriétés* et *valeurs*. Ce système est très similaire aux classes des langages de programmation orientée objet ; ce qui permet aux ressources d'être définies comme des instances. Il est intéressant de noter qu'en 2000, une représentation de UML en RDF a été faite [12], et ce en s'inspirant de XMI, langage standard d'encodage d'UML en XML. La représentation d'un modèle conceptuel d'UML en RDF y est jugée similaire à la définition d'une spécification en RDFS.

1.5.3 Web Ontology Language (OWL)

RDF est un modèle puissant largement utilisé, mais trop limité pour formuler des contraintes sémantiques. RDFS permet d'écrire un vocabulaire organisé avec une taxinomie associée, mais n'est pas suffisant pour décrire une ontologie. Afin de palier à ce problème des couches ont été rajoutées au-dessus de RDF, comme *Ontology Inference Layer* (OIL [13]) - pour définir les ontologies - et *DARPA Agent Markup Language* (DAML [14]) pour rendre plus facile la définition de nouveaux langages permettant la communication entre agents intelligents. Ces deux spécifications ont été regroupées par le W3C en 2004 sous le nom d'OWL. Nous parlerons ici uniquement de la version 1.0 de ce langage, cette dernière étant utilisée pour l'élaboration du projet depuis 2007. Néanmoins, signalons que la version 2.0 [15] est sortie en octobre 2009. Outre une syntaxe légèrement différente par rapport à la version antérieure, OWL 2.0 définit une nouvelle politique d'usage des termes : comment peuvent-ils être utilisés et comment peuvent-ils être inférés. La nouveauté de OWL 2.0 réside dans l'instauration de trois *profils*, sous-ensembles du langage facilitant l'utilisation et l'implémentation : (1) *OWL-EL* pour un temps de raisonnement de type polynomial, (2) *OWL-QL* pour une implémentation au sein d'une base de données

12. http ://infolab.stanford.edu/ melnik/rdf/uml/
13. Il s'agit ici d'une extension d'XML Schéma proche de RDFS et offrant des primitives inspirées des logiques de descriptions.
14. Il s'agit là d'un langage fondé sur RDF, et plus proche des langages objets.
15. http ://www.w3.org/TR/2009/REC-owl2-syntax-20091027/

relationnelles, et (3) *OWL-RL* pour une implémentation au sein d'un moteur de règles.
[Euzenat 2004] présente au travers d'une ontologie de la bibliographie les possibilités offertes par RDFS pour décrire un vocabulaire.

```
<rdfs:Class rdf:ID="Référence" />
  <rdfs:Class rdf:ID="Livre">
    <rdfs:subClassOf rdf:resource="#Référence" />
  </rdfs:Class>
  <rdfs:Class rdf:ID="Biographie">
    <rdfs:subClassOf rdf:resource="#Livre" />
  </rdfs:Class>
  <rdfs:Class rdf:ID=" Autobiographie ">
    <rdfs:subClassOf rdf:resource="#Biographie" />
  </rdfs:Class>
  <rdf:Property rdf:ID="bib:auteur">
    <rdfs:domain rdf:resource="#Référence"/>
    <rdfs:range  rdf:resource="&foaf;Personne"/>
  </rdf:Property>
```

Ces possibilités s'organisent selon une hiérarchie de classes, avec comme classe mère *Référence* qui est plus générale que *Livre*, laquelle est plus générale que *Biographie* qui l'est plus qu'*Autobiographie*. La propriété *auteur* de type *Personne* est également définie avec le vocabulaire RDF : Friend of a Friend (FOAF)[16]. Il est également possible de contraindre plus précisément la description :

- des *classes*, en les décrivant comme union, intersection, complémentaire d'autres descriptions ou comme l'ensemble d'un certain nombre d'individus ;
- des *domaines de relations*, en spécifiant le type de toutes leurs valeurs, ou d'un certain nombre de leurs valeurs ;
- des *relations*, en les déclarant transitives, symétriques ou en spécifiant leur inverse.

Deux classes ou ressources peuvent être équivalentes ou, au contraire, différentes. RDFS ne le permet pas, mais OWL l'implémente de la manière suivante :

```
<owl:Class rdf:ID="Livre">
 <owl:intersectionOf rdf:parseType="Collection">
    <owl:Class rdf:resource="#Référence" />
    <owl:Restriction>
         <owl:onProperty rdf:resource="#titre" />
         <owl:minCardinality rdf:datatype="&xsd;Integer">
              1
         </owl:minCardinality>
      </owl:Restriction>
```

16. http ://www.foaf-project.org/

```
      <owl:Restriction>
          <owl:onProperty rdf:resource="#éditeur" />
          <owl:allValuesFrom rdf:resource="#Publisher" />
      </owl:Restriction>
   </owl:intersectionOf>
</owl:Class>
<owl:Class rdf:ID="Biographie">
  <owl:intersectionOf rdf:parseType="Collection">
      <owl:Class rdf:resource="#Livre"/>
      <owl:Restriction>
            <owl:onProperty rdf:resource="#object"/>
            <owl:allValuesFrom rdf:resource="&foaf;Person"/>
      </owl:Restriction>
    </owl:intersectionOf>
  </owl:Class>
```

La première expression OWL signifie : *la classe Livre est l'intersection de la classe Référence et des choses dont la propriété titre a au moins une valeur et dont la propriété éditeur a pour valeur un Publisher.* La seconde se lit : *la classe biographie est l'intersection de la classe Livre et des choses dont la propriété objet est une personne..*

1.5.3.1 La famille OWL

Il existe trois OWL correspondant à trois sous-langages et trois types d'utilisation.

Sous-langage : *OWL LITE*	Complexité : *faible*

- *Utilisation :* à destination des utilisateurs qui ont essentiellement besoin d'une hiérarchie de concepts, de classifications, et d'une expressivité limitée.
- *Cas d'utilisation :* migrations rapides depuis d'anciens thésaurus ou taxinomies vers les ontologies.

Sous-langage : *OWL Description Logic (DL)*	Complexité : *moyenne*

- *Utilisation* : à destination des utilisateurs qui souhaitent une expressivité maximum sans perdre la complétude du calcul (toutes les inférences sont assurées) et la décidabilité des systèmes de raisonnement (tous les calculs seront terminés dans un intervalle de temps fini) [W3C 2004]. OWL DL est fondé sur les logiques de description, un champ de la recherche ayant étudié un fragment décidable particulier de la logique de premier ordre.
- *Cas d'utilisation :* système de raisonnement automatisé.
- *NB :* le langage OWL DL inclut toutes les structures de langage de OWL, avec des restrictions comme la séparation des types (une classe ne peut pas

être un individu ou une propriété, une propriété un individu ou une classe) [W3C 2004].

Sous-langage : *OWL FULL*	Complexité : *forte*

- *Utilisation* : à destination des utilisateurs qui souhaitent une expressivité maximale et la liberté syntaxique de RDF sans garantie de calcul (*i.e.* complétude comme décidabilité).
- *Cas d'utilisation :* besoin d'un haut niveau de capacité de description, extension du vocabulaire par défaut d'OWL.
- *NB :* dans OWL Full, une classe peut se traiter simultanément comme une collection d'individus et comme un individu à part entière.

Comme [W3C 2004] le souligne, chacun de ces sous-langages représente une extension plus simple par rapport à son prédécesseur (pour ce qu'il est possible d'exprimer et pour ce qu'il est possible de conclure de manière valide). Les affirmations suivantes sont vraies, mais leurs symétriques ne le sont pas :

- *toute ontologie OWL Lite conforme est une ontologie OWL DL conforme ;*
- *toute ontologie OWL DL conforme est une ontologie OWL Full conforme ;*
- *toute inférence OWL Lite valide est une inférence OWL DL valide ;*
- *toute inférence OWL DL valide est une inférence OWL Full valide.*

Dès lors comment choisir le type d'OWL à utiliser ? [Horridge 2004] donne deux règles basées sur des questions très simples :

1. choix entre OWL Lite et OWL DL : est-ce que les constructions simples d'OWL Lite suffisent ?
2. choix entre OWL DL et OWL Full : est-il important de pouvoir effectuer des raisonnements automatiques sur des ontologies ? Est-il important de pouvoir utiliser une expressivité maximale et une puissance de modélisation avec par exemple des méta-classes ?

Une analyse précise des besoins est donc primordiale avant tout développement d'une ontologie informatisée afin de choisir le bon sous-langage OWL. C'est un langage très ouvert puisqu'il permet non seulement d'écrire de nouvelles ontologies mais également d'en employer ou d'en compléter.

1.5.4 Structure d'un document OWL

Une ontologie OWL se présente sous la forme d'un fichier texte avec l'extension *owl* ou parfois *rdf*. Avant de commencer l'écriture d'une ontologie, il faut signaler au début du fichier quels vocabulaires sont utilisés. Les espaces de nommage remplissent ce rôle au sein de tout fichier écrit avec la syntaxe XML. OWL dépend de structures définies par RDF / RDFS et des types de données du schéma XML (*cf.* les préfixes *rdf :* , *rdfs :* et *xsd :*). En détail, selon [W3C 2004] :

- **xmlns** indique à quelle ontologie se rapporter en cas d'utilisation de noms sans préfixe dans la suite du fichier ; cet espace de nommage est qualifié d'implicite.
- **xml :base** identifie l'adresse URI de base de l'ontologie courante.
- **xmlns :owl** indique que dans ce document les éléments préfixés par *owl :* devraient se comprendre comme se référant aux choses issues de l'espace de nommage appelé *http :...* (déclaration conventionnelle introduisant le vocabulaire OWL.)

De façon assez similaire aux fichiers HTML, les en-têtes OWL comportent des méta-données sur l'ontologie créée. [W3C 2004] explicite les éléments cités ci-dessus :

- **rdf :about**="" : cet attribut fournit un nom ou une référence pour l'ontologie (si sa valeur est "" - cas habituel - alors le nom de l'ontologie vaut l'adresse URI de base défini par l'attribut *xml :base*) ;
- **rdfs :comment** : cet élément apporte une annotation à l'ontologie ;
- **owl :priorVersion** : cet élément fournit des crochets pour les systèmes de contrôle de version fonctionnant avec les ontologies. Plusieurs étiquettes permettent d'agrémenter les informations relatives aux différentes versions en matière de compatibilité (ou non) entre elles, de compatibilité (ou non) entre les classes, etc.
- **owl :imports** : cet élément (qui n'admet qu'un seul argument) fournit un mécanisme de type *include* pour l'importation d'une autre ontologie (*i.e.* de l'ensemble de ses assertions) en coordination avec l'espace de nommage.
- **rdfs :label** : cet élément fournit une étiquette en langage naturel à l'ontologie.

Les méta-données disponibles avec OWL peuvent être complétées par des éléments de la norme de *Dublin Core*, par exemple, à condition bien entendu qu'elle soit présente dans l'espace de nommage.

1.5.5 Éléments de langage

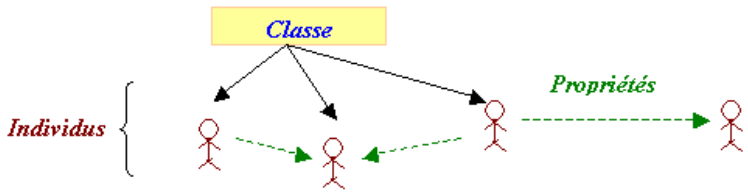


FIGURE 1.9 – Éléments de base d'OWL.

OWL (quelque soit sa déclinaison) est constitué de trois parties : les *classes*, les *individus* et les *propriétés* (*cf.* figure 1.9). Des règles peuvent également être définies et appliquées aux classes comme aux propriétés : symétrie, inverse, égalité, etc. Nous explorerons respectivement ces trois parties en détail au travers de leurs spécifications et de quelques exemples.

1.5.5.1 Éléments de langage - les classes

Les trois déclinaisons de OWL comportent les mêmes mécanismes de classe, à la petite différence près que pour OWL-Full une classe peut être l'instance d'une autre classe (d'une méta-classe) alors que pour les deux autres toute instance d'une classe ne peut être qu'un individu. Une classe se définit par sa description, et il y a six manières de le faire :

1. par appel d'adresse URI, il s'agit uniquement de nommer une classe ;
2. par énumération de ses individus, une classe est décrite par la liste de ses instances à l'aide de la propriété *rdf :oneOf* (n'existe pas en OWL Lite) ;
3. par restriction de propriétés, avec contraintes sur les valeurs ou sur les cardinalités ;
4. par intersection de descriptions (équivalent à une conjonction logique) ;
5. par union de descriptions (équivalent à une disjonction logique) ;
6. par le complémentaire d'une description (équivalent à la négation logique).

Il est possible de compléter cette définition par trois types de liens : l'*héritage*, l'*équivalence* et la *différence*. Selon [W3C 2004], l'héritage (*rdfs :subClassOf*) permet de faire valoir que l'extension de classe d'une description de classe est un sous-ensemble de l'extension de classe d'une autre description de classe. L'équivalence (*owl :equivalentClass*) permet de faire valoir qu'une description de classe a exactement la même extension de classe qu'une autre description de classe. Enfin la différence (*owl :disjointWith*) permet de faire valoir que l'extension de classe d'une description de classe n'a aucun membre commun avec l'extension de classe d'une autre description de classe.

1.5.5.2 Éléments de langage - les propriétés

Il existe deux types de propriétés :
- *Object Property* pour relier des individus entre eux ;
- *Datatype Property* pour relier des individus à des valeurs de données.

Avec OWL, les propriétés peuvent être organisées en hiérarchie. Il peut également être pertinent de caractériser une relation, en la définissant par exemple comme *symétrique*[17] ou *inverse*[18]. Une propriété peut également être *transitive*[19] , *fonctionnelle*[20] ou *fonctionnelle inverse*[21]. Il est enfin possible de définir des restrictions sur les propriétés. Cette opération sur les relations peut s'opérer sur deux plans : restriction sur les valeurs, ou restriction sur les cardinalités (un peu comme dans les modèles entités-association). Ces restrictions se font à l'intérieur d'un élément *owl :Restriction*.

17. $\forall x, y, p(x, y) \Leftrightarrow p(y, x)$
18. $\forall x, y, p_1(x, y) \Leftrightarrow p_2(y, x)$
19. $\forall x, y, z, p(x, y) \wedge p(y, z) \Rightarrow P(x, z)$
20. $\forall x, y, z, p(x, y) \wedge p(x, z) \Rightarrow y = z$
21. $\forall x, y, z, p(y, x) \wedge p(z, x) \Rightarrow y = z$

1.5.5.3 Éléments de langage - les individus

Un individu représente des objets dans le domaine concerné par l'ontologie. Un individu existe s'il est relié à au moins une classe. Il faut cependant faire attention au fait qu'OWL n'utilise pas l'hypothèse du nom unique (*Unique Name Assumption - UNA*). Cela signifie que plusieurs noms différents peuvent désigner la même chose [22]. Il faut donc signifier explicitement si un individu est identique ou non à un autre.

1.5.6 SPARQL Protocol And RDF Quey Language (SPARQL)

Le Web des données fournit aujourd'hui un nombre très important d'informations sous la forme, essentiellement, de triplets RDF. SPARQL est un langage de requête faisant l'objet d'une recommandation de la part du W3C depuis janvier 2008 [23]. Inspiré de la syntaxe SQL, SPARQL effectue une recherche de triplets au sein du graphe RDF. Ainsi, les moteurs de recherche SPARQL analysent les correspondances des patrons de graphe, le plus simple étant le triplet. Les requêtes SPARQL intègrent également des fonctions spécifiques comme l'union ou l'intersection de patrons, des filtrages et des opérateurs de comparaisons sur les valeurs.

Aujourd'hui, la plupart des entrepôts de données RDF offrent un point d'accès SPARQL où l'utilisateur peut saisir sa requête. A la différence tout de même, et non négligeable, que contrairement à une base de données où la structure est fixe et connue, ici nous travaillons sur un graphe en perpétuelle évolution donc dotée d'une structure dynamique.

L'exemple donnée est pris sur l'entrepôt de données DBPedia [24], doté de plusieurs points d'accès [25] dont les résultats pour une même requête peuvent parfois diverger. Supposons que nous cherchions toutes les personnalités présentes dans la base et née à Paris avant 1945, nous écrivons alors la requête suivante :

```
<!-- Liste des préfixes -->
SELECT ?name ?birth ?death ?person
WHERE {
    ?person dbpedia2:birthPlace <http://dbpedia.org/resource/Paris>.
    ?person dbo:birthDate ?birth .
    ?person foaf:name ?name .
    ?person dbo:deathDate ?death
    FILTER (?birth < "1945-01-01"^^xsd:date) .
}
ORDER BY ?name
```

22. Ce qui est le cas des termes associés à un concept.
23. http ://www.w3.org/TR/rdf-sparql-query/
24. http ://dbpedia.org
25. http ://dbpedia.org/sparql, http ://querybuilder.dbpedia.org/ , http ://dbpedia.org/snorql

Il s'agit de rechercher dans le graphe tous les motifs correspondant au patron représenté sur la figure 1.10, avec comme contraintes sur valeur le fait que la propriété *dbo :birthDate* soit inférieure au 1er janvier 1945 et que la propriété *dbpedia2 :birthPlace* soit la ressource DBPedia correspondant à la ville de Paris.

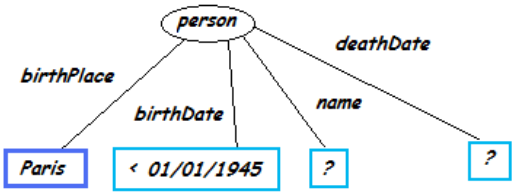


FIGURE 1.10 – Patron de requête SPARQL.

Les résultats d'une requête peuvent nous parvenir soit sous forme d'une liste de valeurs (comme le montre le tableau 1.2), ou sous la forme de graphes RDF.

TABLE 1.2 – Résultat requête DBPedia

name	birth	death	person
A. E. Becquerel	24/03/1820	11/05/1891	A._E._Becquerel
Adrien-Marie Legendre	18/09/1752	10/01/1833	Adrien-Marie_Legendre
Alain Bombard	27/10/1924	19/07/2005	Alain_Bombard
Alfonso Bertillon	24/04/1853	13/02/1914	Alphonse_Bertillon
Antoine Henri Becquerel	15/12/1852	25/08/1908	Henri_Becquerel
Armand Jean du Plessis,	09/09/1585	04/12/1642	Cardinal_Richelieu

SPARQL possède différents opérateurs, tel l'*union* qui permet de donner des patrons alternatifs. Mais il est également possible de construire un résultat, c'est à dire ajouter au graphe résultat fourni par la requête ses propres données. L'exemple suivant crée une relation entre *Substance dangereuse* et *Explosif* pour toutes les substances ayant un risque d'explosion.

```
<!-- Liste des préfixes -->
CONSTRUCT
{ ?substance rdf:type txa:Explosifs}
WHERE {
  ?substance rdf:subClassOf txa:Substance_dangereuse.
  ?substance txa:risque txa:R22
}
```

1.5.7 RDFa

Le langage RDFa, objet d'une recommandation W3C depuis octobre 2008 [26], fournit une syntaxe et un ensemble de balises XML pour décrire les données struc-

26. http ://www.w3.org/TR/rdfa-syntax/

turées en (X)HTML. D'usage fort simple, il permet d'injecter directement de la sémantique dans le code d'une page web à la manière d'un système d'annotation de texte. Pour ce faire, RDFa propose d'utiliser des vocabulaires RDF (tel FOAF ou Dublin Core, par exemple), puis d'étiqueter le contenu de la page web au moyen d'éléments contenus dans ces vocabulaires. Par exemple, dans le code qui suit, RDFa permet de distinguer dans le texte : l'auteur, le titre, la date de création, une vidéo, etc.

```
<div about="http://uri.to.newsitem">
<i>Publié le
   <span property="dc:date">08/07/2010 à 10:13</span>
</i><br>
ÉNERGIE
   <span property="dc:title">
   L'avion solaire Solar Impulse a atterri après un vol de 26 heures
   </span>
<br>
Par <span property="dc:creator">Steve Bird</span>
<br>
L'avion expérimental Solar Impulse a atterri jeudi matin
...
Voir <a href="http://www.youtube.com/watch?v=T0wsoERR0-M&feature=fvst"
      rel="dcmtype:MovingImage">
```

RDFa utilise en partie la syntaxe HTML avec les attributs *class*, *id*, *rel*, *rev* et *href*. Il ajoute ses propres éléments, avec les attributs :

- *about* et un URI afin de spécifier la ressource décrite par des méta-données ;
- *property* afin de spécifier une propriété pour le contenu d'un élément ;
- *content* afin de remplacer le contenu d'un élément quand on utilise l'attribut de propriété (optionnel) ;
- *datatype* afin de spécifier le type de donnée du contenu (optionnel).

1.6 Conclusion

Le Web, à son origine, avait pour vocation de relier des documents statiques en HTML au moyen de liens hyper-textes. Le Web Sémantique propose un ensemble de méthodes et de technologies fondées sur des langages XML et visant à rendre compréhensible la sémantique de ces documents par les ordinateurs. Cependant, si les méta-données permettent de réaliser des recherches grâce à un thésaurus qui met en correspondance certains mots ou listes de mots, ces recherches restent limitées et binaires (*e.g.* [Chausson 2007]). En effet, si le terme n'est pas présent le système retourne une réponse vide ou inadéquate. Les ontologies, et le langage OWL, permettent de palier cet inconvénient grâce aux liens entre les informations. Ainsi, au moyen de cette technologie, il est possible de faire ressortir des documents sur les

joints ou le caoutchouc alors que ces pages traitent d'autres termes.
Cette dernière décennie, RDF et OWL sont devenus la véritable ossature du Web Sémantique. RDF est aujourd'hui le standard utilisé pour le Web des Données, il est utilisé dans ces grands entrepôts que sont DBPedia, Freebase ou Geonames. OWL est le langage du Web Sémantique et des ontologies. SPARQL, langage de type SQL, permet de créer des requêtes tant sur les données RDF que OWL. Si le modèle économique sous-jacent au Web Sémantique est encore balbutiant, les technologies afférentes quittent progressivement les laboratoires de recherche pour se démocratiser notamment en se greffant sur les applications de web social.

Chapitre 2

Ontologie

2.1 Introduction

Définitions [Ontologie]

- du latin *ontologia* (1646). Philo. Partie de la métaphysique qui s'applique à l'être en tant qu'être, indépendamment de ses déterminations particulières [Robert 1984]
- du préfixe *onto*, être, et du grec *logos*, discours. (...) Philo. Science de l'être en tant qu'être, de ses diverses espèces, de ses propriétés et relations. Elle discute sur le réel, le possible, l'impossible, le potentiel, l'actuel, etc. avec une partie essentielle sur les causes (*cf. métaphysique d'*Aristote). Kant la subordonne à la théorie de la connaissance, en étudiant le rapport entre sujet et objet, entre la pensée et l'être. C'est pour lui le système de la raison pure. La question est dès lors : comment passer de sujet à objet, comment la pensée peut-elle s'accorder avec la réalité. Évoluant avec le temps, aujourd'hui l'ontologie cherche à saisir sous les apparences les choses en soi [Larousse 1932].

Comme précisé dans l'une de ces définitions, il nous faut remonter à Aristote (384-322 av. JC) et plus précisément à l'un de ses écrits : *les Catégories*. Dans cette œuvre (première partie de l'*Organon*[1]), le philosophe établit les différentes descriptions associées aux manifestations de l'être dans le monde. Aristote donne ainsi une liste de dix catégories sur la base des catégories grammaticales du grec : la substance (**quoi ?**), la quantité (**combien ?**), la qualité (**quels attributs ?**), la relation (**plus X que Y, etc.**), le lieu (**où ?**), le temps (**quand ?**), la posture (**comment ?**), la possession (**avec quoi ?**), l'action (**en faisant quoi ?**) et le pâtir (**affecté par ?**) [Eco 1994]. Selon ce philosophe, un individu donné est identique dans le temps en essence (structure stable - la substance), mais subit de multiples changements dans ses catégories dites accidentelles ou transitoires (comme sa qualité/apparence, sa posture, ses actions, etc.). Il a ainsi pu établir la première ontologie de l'être.

L'histoire de l'Ontologie est étroitement liée à celle de la métaphysique, la *philosophie première*. Tant est si bien que le terme même d'Ontologie n'apparaîtra pas

1. Qui signifie instrument, la logique est l'instrument du savoir.

avant le milieu du $XVII^e$ siècle, puisque jusqu'à cette époque la recherche sur l'être en tant qu'être faisait partie intégrante de la métaphysique. À partir de 1613, le mot latin *ontologia* est usité pour désigner l'étude de l'être en général, dans l'ouvrage *Lexicum philosophicum* du philosophe allemand R. Göckel. Au $XVIII^e$ siècle, le vocable Ontologie désigne la métaphysique générale, par distinction des métaphysiques spéciales que sont la *psychologie rationnelle* (l'âme de l'être), *la cosmologie rationnelle* (l'être dans le monde) et *la théologie rationnelle* (l'être, créature de Dieu).

Cette catégorisation est l'ancêtre des catégories de l'entendement (concepts purs) de Kant. Ce dernier pense que la logique objective prend la place de la métaphysique d'autrefois. L'Ontologie va ainsi subir maintes évolutions, d'abord entre les mains de la phénoménologie (ontologie régionale ou science idéale de genre d'être, avec Husserl, Heidegger et Hartmann) puis entre celles de la philosophe analytique (l'ontologie est déterminée par la sémantique de son langage - naturel ou non, avec W. Quine). Cette vision s'approche de celle des ontologies formelles telles qu'elles peuvent être élaborées dans le domaine de l'Intelligence Artificielle.

En résumé, l'Ontologie, en tant que concept philosophique, s'intéresse à la qualité de l'être, à la notion d'existence et ses catégories fondamentales - soit les propriétés constitutives de l'être. L'Intelligence Artificielle et les chercheurs du Web sémantique ont adopté une signification spécifique : une ontologie est un modèle qui définit de façon formelle les relations entre les concepts [Lacombe 2006].

De ces données philosophiques peuvent être tirés trois constats :

1. le premier est la mise en avant au travers de ces quelques définitions du fort caractère classificateur et descriptif de l'Ontologie ;
2. le deuxième est la définition philosophique donnée par [Larousse 1932] où sont évoquées les notions d'espèces, de propriétés et de relations, notions présentes tant dans l'*approche orientée objet* que dans OWL ;
3. la troisième est l'évolution de l'Ontologie dans le temps, avec Kant et son extension à la théorie des connaissances.

La suite de ce chapitre présente les principaux degrés de modélisation des connaissances : organisation, vocabulaire, thésaurus, taxinomie et ontologie. Les différentes composantes des ontologies, à savoir les concepts (termes, propriétés et instances) et les relations, sont explicitées dans la section 2.6. Enfin, les méthodologies et outils permettant de construire ces dernières sont abordés dans la section 2.7.

2.2 L'organisation

Définition [Organisation]
– mise en place réfléchie de dispositions en vue d'un résultat déterminé.

– *Fig.* Constitution intellectuelle d'une manière donnée. *Biol.* Juxtaposition d'un certain nombre de parties, semblables ou différentes, fonctionnant synergiquement. [Larousse 1932]

L'idée première de ce degré de modélisation est d'établir une certaine forme d'organisation des objets et de la pensée. Le plus souvent, ce regroupement est issu d'un processus de catégorisation consistant à rassembler des éléments suivant une similarité de propriétés fonctionnelles ou descriptives. Ce processus est détaillé dans le chapitre 3.

2.3 Vocabulaire

Définition [Vocabulaire]
– ensemble de mots ayant une valeur de dénomination [de Poche 1983].

La notion de vocabulaire, qu'il soit composé de symboles ou de termes, est liée à l'action de communiquer, de transmettre de l'information soit instantanément, soit dans le temps. Vocabulaire, communication et indirectement écriture sont fortement liés. L'écriture, au sens de représentation graphique d'une langue au moyen de signes inscrits ou dessinés sur un support, c'est d'abord le pouvoir [Eco 1993]. Le pouvoir de se souvenir. Les premières écritures furent avant tout comptables : il fallait pouvoir marquer dans l'argile, donc se souvenir, qu'à telle date untel avait troqué avec tel autre tant de bêtes contre tant de fruits. Cette trace avait valeur de pacte social entre les individus. Il ne s'agissait pas encore de lettres ni de mots, mais bien de symboles parfaitement identifiables et dont le sens répondait à des règles propres à chaque communauté (à un symbole/signe correspond un signifié et un signifiant, tels sont les trois piliers de la Sémiotique) [Eco 1994]. Des siècles plus tard, sous les couleurs des vitraux, le bien et le mal furent inculqués aux peuples illettrés par les pouvoirs religieux et politiques. Voilà, autant de médias simples véhiculant des symboles dont l'interprétation intuitive ou dirigée (par l'éducation et la culture, la société, la religion...) transforme la donnée brute en information puis en connaissance et savoir. L'écriture n'est ainsi qu'une suite de symboles (la syntaxe) regroupés mentalement en ensembles de formes plus ou moins connues appelées mots (le vocabulaire) et auxquels l'apprentissage (avec l'influence de paramètres comme la culture, le groupe, etc.) associe un sens (la sémantique). La signification du vocabulaire n'est donc pas indépendante du groupe qui l'utilise (*e.g.* [Dubuc 2005]). De par le fait que ces mots ont une valeur de dénomination et non de définition, il est possible de dire qu'ils forment un ensemble de termes premiers. Notons enfin qu'il serait plus juste de parler de dialecte plutôt que de vocabulaire. En effet, le dialecte est défini comme étant une variété d'une langue dont les caractéristiques dominantes sont sensibles aux utilisateurs, alors que le vocabulaire est un ensemble de termes dont le sens est expliqué.

Les ressources termino-ontologiques (RTO) permettent d'associer un vocabulaire indépendant à une ontologie. Elles contiennent les concepts du domaine, mais également les termes les dénotant. Au sein d'une RTO, le vocabulaire doit répondre à au moins deux contraintes :

1. si un même terme est utilisé pour signifier deux concepts dans deux contextes différents, alors il faut modifier le nom du concept pour lever l'ambiguïté ;
2. si plusieurs termes sont utilisés pour désigner la même chose, alors il faut en choisir un et placer les autres dans la catégorie des synonymes.

2.4 Thésaurus

Définition [Thésaurus]

– mot latin signifiant trésor. Répertoire de termes normalisés pour l'analyse de contenu et le classement de documentation dans un domaine [Robert 1984].

Le thésaurus - par l'apport de nouveaux liens - ajoute des informations sur les relations linguistiques qu'entretiennent les mots du vocabulaire, et permet ainsi d'élargir ou de restreindre le champ des connaissances d'un simple vocabulaire. Il s'agit d'un dictionnaire dont les liens entre termes sont uniquement de nature linguistique (*i.e.* synonymie [2], antonymie [3], homonymie [4], etc.).

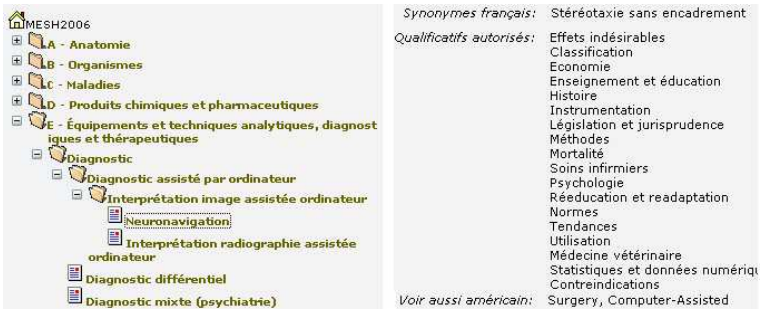


FIGURE 2.1 – Exemple de thésaurus médical (MESH [Dailland 2005]).

La figure 2.1 représente un extrait de l'arborescence de termes du thésaurus médical bilingue Medical Subject Heading (MESH) [Dailland 2005], avec pour chacun des relations de synonymie, des liens avec des thèmes similaires ou en rapport avec le sujet.

2.5 Taxinomie

Définition [Taxinomie]

2. Relation entre deux mots de sens très voisins ou identiques.
3. Relation entre des mots de sens contraires.
4. Qui a la même prononciation ou la même orthographe qu'un autre, mais un sens différent.

– (gr. *taxis* classement, et *nomos* administrer) classification d'éléments, d'espèces. [Robert 1984]

Les éléments d'un environnement donné peuvent être appréhendés au moyen d'une organisation par catégories et d'un vocabulaire organisé pour dénoter ces éléments. Il est possible de compléter ce système par l'apport d'une structure hiérarchique et sémantique. La taxinomie, à la base réservée aux sciences du vivant avant de se généraliser, a largement été utilisée par les naturalistes au $XVIII^e$ siècle, la connaissance des espèces se trouvant alors modélisée sous une forme structurée et hiérarchisée de concepts.

Nous entendons par hiérarchisation le fait de classer par spécialisation, en plaçant des liens précis "$parent \rightarrow enfant$" d'*héritage*[5] entre les termes. Selon [Brunner 2003, Dubost 2004], ce lien donne un sens supplémentaire, une signification.

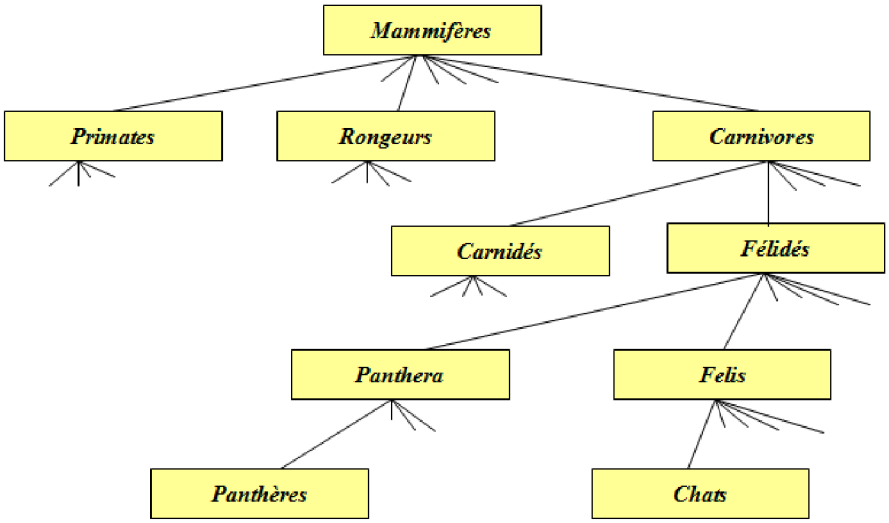


FIGURE 2.2 – Extrait d'une taxinomie des êtres vivants.

Prenons à titre d'exemple la taxinomie décrivant les organismes vivants (*cf.* figure 2.2). Elle se développe principalement sur neuf niveaux hiérarchiques (avec parfois des niveaux intermédiaires tels que sous-famille, sous-ordre, etc.), soit du plus général au plus précis : *Monde vivant* → *Domaine* → *Règne* → *Embranchement* → *Classe* → *Ordre* → *Famille* → *Genre* → *Espèce*. Ainsi, il est possible de rattacher l'espèce *Chat* au genre felis, lequel est de la famille des félidés, qui est de l'ordre des carnivores, lequel est de la classe des mammifères, etc.

5. Appelée également relation de spécialisation, il s'agit d'une relation d'hyponymie, liant un hyperonyme (concept plus général) à son hyponyme (concept spécialisé). Les taxinomies au sens biologique répondent à cette définition.

Ainsi, grâce à cette taxinomie, il est possible de savoir qu'un chat siamois possède (hérite) de toutes les caractéristiques d'un chat, mais également d'un félidé et par conséquent d'un carnivore. Et grâce aussi à cette représentation, il est possible de savoir ce qui différencie un chat d'une panthère, pourtant tous les deux classées comme des félidés. Enfin, il est également possible de déterminer une relation d'instantiation car le chat du voisin n'est pas n'importe quel chat, il s'agit d'un élément physique, réel, appartenant à l'ensemble des chats.

2.6 Ontologie en Ingénierie des Connaissances

Définitions [ontologie]
- théorie qui tente d'expliquer les concepts qui existent dans le monde et comment ces concepts s'imbriquent et s'organisent [Eppstein 2005] ;
- spécification explicite d'une conceptualisation partagée pour un domaine de connaissance [Gruber 1993a].

La littérature offre de nombreuses définitions du vocable *ontologie* car elles dépendent pour une large part du domaine d'appartenance de leurs auteurs, mais également des objectifs qu'ils désirent atteindre par l'intermédiaire de cet outil et du niveau de formalisation nécessaire à la réalisation de leurs objectifs. L'ontologie informatique a en commun avec son homologue philosophique le fait de **conceptualiser, de décrire l'existant au travers de catégories**.

La définition la plus populaire de [Gruber 1993a] [6] s'appuie sur deux points. Une ontologie est :
- la ***conceptualisation*** d'un domaine, *i.e.* un choix quant à la manière de décrire un domaine au travers des éléments qui le composent ;
- et la ***spécification*** de cette conceptualisation, *i.e.* sa description formelle (*i.e.* un ensemble de concepts, de propriétés, d'axiomes, de fonctions et de contraintes pouvant être comprises par un ordinateur).

[Guarino 1995] illustre la définition de Grüber par l'exemple ci-après (*cf.* figure 2.3). Prenons une conceptualisation selon la structure suivante : *{{Achille, Bertrand, Charles, David, Édouard} , {Dessus, Dessous, Rien, Sol}}* où *{Achille, Bertrand, Charles, David, Édouard}* représente l'univers du discours - les habitants des immeubles, et *{Dessus, Dessous, Rien, Sol}* l'ensemble des relations entre ces habitants. L'arrangement "*Charles habite en dessous de David ; Achille et Bertrand habitent au-dessus du sol*" peut se traduire par les deux conceptualisations suivantes :

6. [Fensel 1998] précise la définition par *spécification formelle et* ***explicite*** *d'une conceptualisation partagée.* Pour [Guarino 1995], il s'agit plus d'une *spécification* ***partielle*** *et formelle d'une conceptualisation.*

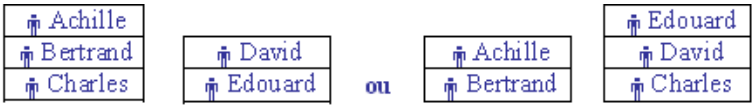


FIGURE 2.3 – Conceptualisations.

Ces deux vues différentes sont correctes, mais elles représentent la conceptualisation énoncée dans deux mondes possibles. Une ontologie est donc une conceptualisation qui, une fois formalisée, va s'appliquer à un seul monde. Deux phases dans la construction d'une ontologie peuvent être distinguées : la *conceptualisation* puis l'*ontologisation* (formalisation de la conceptualisation).

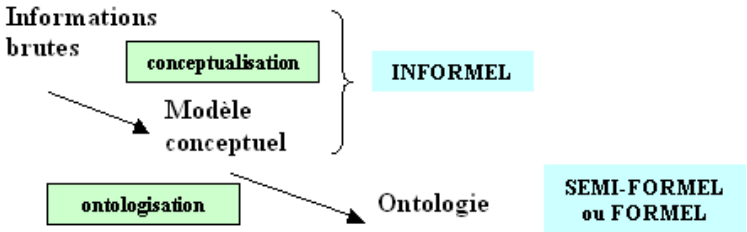


FIGURE 2.4 – Construction d'une ontologie [Fürst 2002].

2.6.1 Ontologie, pour faire quoi ?

L'explosion des données et connaissances accessibles durant les deux dernières décennies nécessite des formes d'organisation des "savoirs" qui facilitent l'interopérabilité pour l'échange et le partage des connaissances. Pour [Eco 1994], "*un grand nombre de théories (depuis la taxinomie jusqu'à la linguistique, depuis les langages formalisés jusqu'aux projets d'intelligence artificielle et aux recherches des sciences cognitives) sont nées comme autant d'effets collatéraux d'une recherche sur la langue parfaite*" (selon cet auteur, une sorte de langue universelle idéalement constituée de toutes les langues).

[Noy 2000] donne au moins cinq raisons de construire des ontologies :

1. *partager la compréhension commune de la structure de l'information* : des agents travaillant sur des sites distincts mais sur des thèmes identiques ou proches peuvent ainsi communiquer ;
2. *permettre la réutilisation du savoir d'un domaine* : si une ontologie est aisément réutilisable, elle peut alors être intégrée dans une seconde plus large et ainsi de suite afin de pouvoir décrire d'autres domaines, la diffusion et l'exploitation du savoir s'en trouvent améliorées ;

3. *expliciter ce qui est considéré comme implicite sur un domaine* : l'ontologie permet de rendre compréhensible des données propres à un domaine et jusque-là exprimées sous forme de dialectes, elle facilitera là encore la communication entre humains, machines et humains et machines entre elles ;
4. *distinguer le savoir sur un domaine du savoir opérationnel* : il s'agit d'une démarche indépendante d'une quelconque implémentation sur machine et sans mécanismes de raisonnement - ce qui rend l'ontologie à la fois réutilisable et adaptable ;
5. *analyser le savoir sur un domaine* : analyse formelle des termes employés et standardisation du vocabulaire afin de former une ontologie avec une possibilité de réutilisation et d'extension.

Comme le montre la figure 2.5, les ontologies, et bases de connaissances élaborées à partir d'elles, peuvent être utilisées pour la résolution de problèmes (notamment en recherche médicale [7]), pour le Web sémantique (ressources terminologiques [8]), etc.

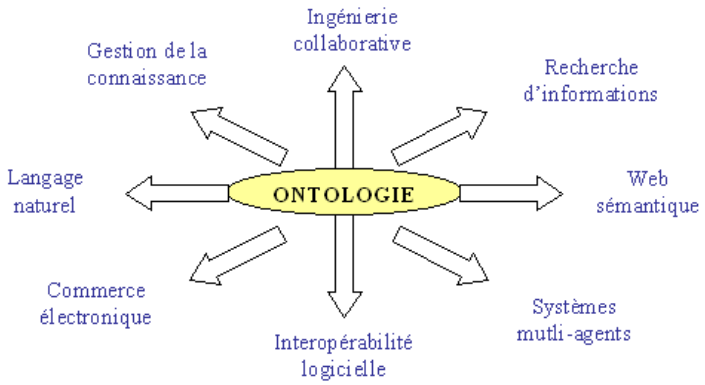


FIGURE 2.5 – L'ontologie au centre de multiples domaines.

2.6.2 Une ontologie, des ontologies

Selon [Uschold 1996], les ontologies peuvent être regroupées en plusieurs types, en prenant en compte au moins trois critères :

1. le **degré de formalisme** de la représentation suivant le contexte d'usage : informelles (en langage naturel), semi-formelles (langage artificiel ou forme restreinte et structurée d'un langage naturel), et hautement formelles - ou denses - (sémantique précise, théorèmes et preuves) [Brisson 2004].

7. Par exemple, recherche sur la maladie de Crohn [Amrani 2005], avec approche générale de fouille de textes spécialisés.

8. Par exemple, dans le contexte particulier du Droit afin de faciliter l'accès à des sites juridiques [Lame 2002].

2. l'**objectif opérationnel** : communication entre utilisateurs, interopérabilité entre systèmes, réutilisabilité, résolution de problèmes.
3. le **sujet** : domaine de connaissance, raisonnement, modèle de représentation.

Il est possible d'ajouter un quatrième critère : la **granularité** (niveau de détail utilisé lors de la conceptualisation de l'ontologie). Deux types de granularité peuvent ainsi être distingués : **fine**, qui correspond à un niveau de détails élevé avec un vocabulaire riche assurant une description détaillée des concepts pertinents d'un domaine ; **large**, qui correspond à un niveau de détail faible (le raffinement faisant l'objet en ce cas de conceptualisations sous-jacentes).

À partir de ces quatre critères (degré de formalisme, objectif opérationnel, sujet et granularité), [Gómez-Pérez 1999, Psyché 2003] distinguent neuf grandes catégories d'ontologies non exclusives :

1. les *ontologies de représentations des connaissances* : formalisation des connaissances (par exemple, les Frame Ontology qui intègrent les primitives de représentation des langages à base de frames : classes, instances, facettes, propriétés/slots, relations, restrictions, valeurs autorisées, etc.)
2. les *ontologies communes (ou de haut-niveau)* : vocabulaire lié aux objets, aux propriétés, aux états, aux valeurs, au temps, à l'espace, à la causalité, au comportement, à la fonction ; elles intègrent les fondements philosophiques dans leur conception ainsi que les primitives cognitives communes à plusieurs domaines (objectif de standardisation).
3. les *ontologies génériques* : noyaux d'ontologies réutilisables.
4. les *ontologies de domaine* : vocabulaire des concepts d'un domaine, relations entre ces derniers, les activités ainsi que les théories et les principes de base de ce domaine ; la plupart des ontologies existantes sont des ontologies du domaine.
5. les *ontologies de tâches* : un ensemble de termes (noms, verbes, adjectifs génériques) au moyen desquels on peut décrire au niveau générique comment résoudre un type de problème (gestion des tâches de diagnostic, de planification, de conception, de configuration, de tutorat, etc.)
6. les *ontologies de domaines-tâches* : ontologie de tâche réutilisable dans un seul domaine.
7. les *ontologies d'application* : structuration d'un domaine particulier ; les concepts dans les ontologies d'application correspondent souvent aux rôles joués par les entités du domaine tout en exécutant une certaine activité. Sa construction peut se faire par spécialisation d'une ontologie générique.
8. les *ontologies légères* : composées d'une hiérarchie de concepts et d'une hiérarchie de relations.

9. les *ontologies denses* : composées d'une hiérarchie de concepts, d'une hiérarchie de relations et d'axiomes.

Le schéma donné par [Psyché 2003] (*cf.* figure 2.6) donne un aperçu assez complet de typologies d'ontologies selon les quatre dimensions de classification. Par rapport aux critères énoncés par [Uschold 1996, Fürst 2002], nous retrouvons dans ce schéma le degré de formalisation, le sujet (objet de la conceptualisation), et la granularité.

Les ontologies - tous types confondus - possèdent des éléments de base que sont les **concepts**, les **relations**, les **fonctions**, les **propriétés**, les **axiomes** (assertions, acceptées comme vraies dans le domaine) et les **instances** (extensions)[9]. Commençons par définir les concepts, avant d'aborder les relations puis les propriétés.

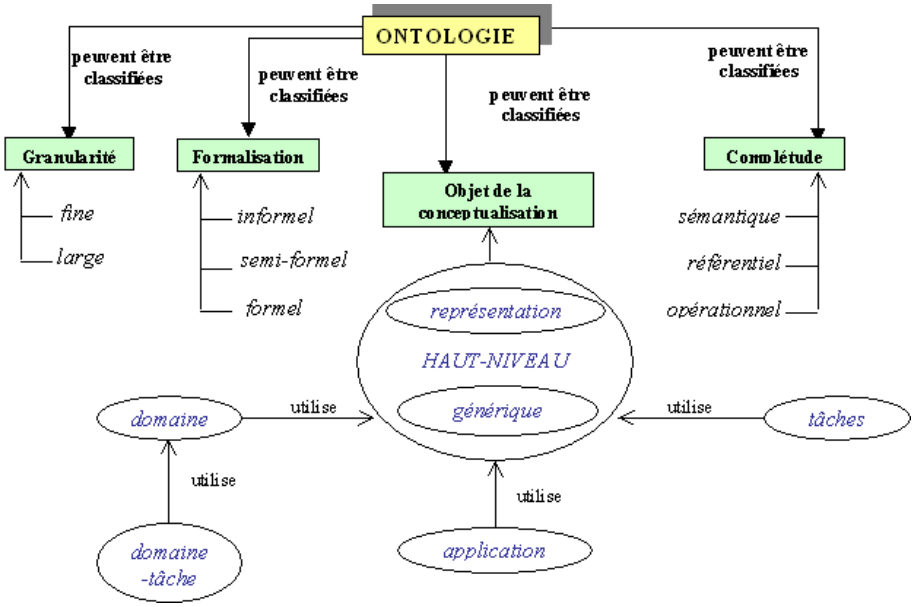


FIGURE 2.6 – Typologies d'ontologies selon quatre dimensions de classification.

2.6.3 Les concepts

Définitions [Concept]

- lat. *concipere*, recevoir ; représentation mentale générale et abstraite d'un objet [Robert 1984].
- contenu psychologique intermédiaire entre le référent (objet, idée, sentiment, action ou personne) et le mot lexical, graphique ou sonore, correspondant [CAF 2005].

9. Cet appartenance ne fait pas aujourd'hui l'objet d'un consensus.

– résultat de l'opération par laquelle l'esprit isole de certaines réalités données dans l'expérience un ensemble dominant et stable de caractères communs qu'on désigne ordinairement, en les généralisant, par le même mot. [Morfaux 1995]

Les éléments potentiellement constitutifs d'un concept sont : le *terme*, l'*intension* et l'*extension*.

Définition [Terme]

– concept représenté par son expression verbale et désignant le sujet ou le prédicat de la proposition [Morfaux 1995].

Définition [Intension / compréhension]

– lat. *comprehensio*, action de saisir ensemble. Ensemble des caractères essentiels communs appartenant à un terme et à un concept, qui s'exprime par la définition. Par exemple : un oiseau est un vertébré, à sang chaud, ovipare, etc. (...) Le concept d'être a une compréhension minimale et une extension maximale, et c'est l'inverse pour l'individu [Morfaux 1995].

Définition [Extension]

– ensemble des êtres, objets ou faits auxquels s'applique un concept et le terme qui les désigne ; l'extension s'exprime par la classification. Par exemple, *vivant* a une plus grande extension que *animal* qui a lui-même une plus grande extension que *vertébré*, etc. [Morfaux 1995]

Prenons pour illustration le concept *Chat* :

– **terme** : chat, matou, minou, greffier, grippeminaud, raminagrobis, mistigri, patte-pelu ;
– **intension** : mammifère carnivore de la famille des félidés, vivant de manière autonome et indépendant, s'exprimant par miaulement ;
– **extension** : pupuce (mon chat), simbad (le chat de ma copine), etc.

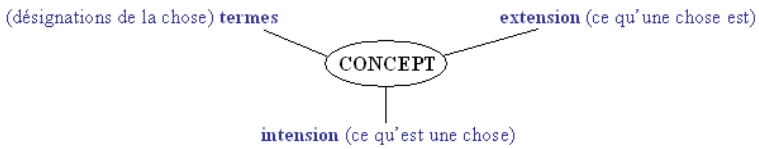


FIGURE 2.7 – Concept.

Pour [Guarino 2001], deux concepts peuvent être :

– **équivalents** : s'ils ont la même extension ;
– **incompatibles** : si leurs extensions sont disjointes (par exemple : chat et

chien) ;
- **dépendants** : un concept c_1 est dépendant d'un concept c_2 si pour toute instance de c_1 il existe une instance de c_2 qui ne soit ni partie ni constituant de l'instance de c_1 (par exemple : *enfant* dépendant de *parent*).

Par rapport à la notion d'intension et d'extension, GUARINO définit un type particulier de concept - le **concept générique** : concept dont l'extension est vide. Par exemple : la vérité.

Enfin, signalons que les concepts peuvent être classés selon trois dimensions :

1. niveau d'abstraction : concret ou abstrait ;
2. atomicité : élémentaire ou composé ;
3. niveau de réalité : réel ou fictif.

2.6.4 Le triangle sémiotique, une modélisation des concepts

On peut associer au graphe orienté de la figure 2.8 un triangle *sémiotique*[10] (exemple sur la figure 2.9), équivalent - selon [Roche 1999] - à la phrase : *Savoir ce que signifie un mot se ramène à connaître l'idée dont il est le signe*. Ce système de connaissance peut se traduire comme la perception d'un système de signes. Ainsi, trois niveaux indissociables interviennent dans un phénomène perceptible :

1. le référent / *signe* - la manifestation ;
2. le *signifié* - le sens ;
3. et le *signifiant* - la désignation.

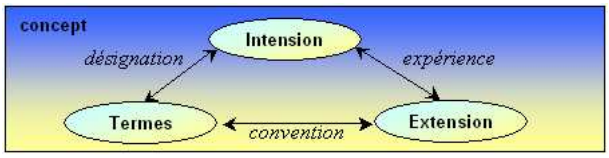


FIGURE 2.8 – Graphe modélisant un concept.

Trois dimensions sont perçues dans un phénomène perceptible :

1. *syntaxique*, l'organisation structurée des signes ;
2. *sémantique*, la signification des signes ;
3. et *pragmatique*, la signification des signes en connaissant le contexte.

A l'origine de cette modélisation, on peut citer en premier lieu ARISTOTE, pour qui l'objet est une substance première représentée dans tout son caractère concret,

10. Théorie générale des signes, de leur signification sous toutes les formes [Peirce 1978].

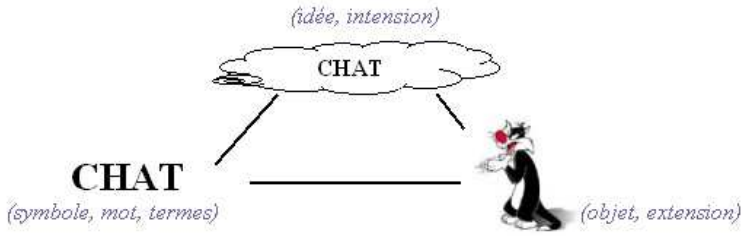


FIGURE 2.9 – Triangle sémiotique.

tandis que l'idée est une affectation de l'esprit (approche nominaliste). Dans *Perl hermeneias*, ARISTOTE fixe les bases de cette triade : *La parole est un ensemble d'éléments symbolisant les états de l'âme, et l'écriture un ensemble d'éléments symbolisant la parole. Et, de même que les hommes n'ont pas tous le même système d'écriture, ils ne parlent pas tous de la même façon. Toutefois, ce que la parole signifie immédiatement, ce sont des états de l'âme qui, eux, sont identiques pour tous les hommes ; et ce que ces états de l'âme représentent, ce sont des choses, non moins identiques pour tout le monde.*

Selon [Rastier 2001], cette triade, schéma de pensée des théories de la signification qui structure l'espace problématique d'une tradition, a su traverser le temps des Anciens jusqu'aux Classiques en passant par les Médiévaux, avec une étonnante stabilité et ce à travers *mille débats et remaniements*. Si un langage (verbal ou non) est le moyen d'exprimer des connaissances personnelles à l'émetteur, alors on peut considérer qu'il associe à une classe de termes des propriétés *(a, b, c, d)* qui délimitent chaque concept. Ainsi, par exemple, [Eco 1984] nous explique que lorsque nous entendons *Tu peux prendre le truc pour la télé sur le meuble de la cuisine ?*, nous devons comprendre que notre interlocuteur - qui nous montre du doigt la télécommande - affirme que dans son monde de référence, sur le meuble de la cuisine, il y a une chose qui possède les propriétés *(a, b, c, d)* définies par le langage comme caractérisant le concept de télécommande associé à toutes les occurrences du terme - y compris truc accompagné de la désignation de l'objet. Autrement dit nous mettons régulièrement en jeu ce triangle pour transformer en informations des données en provenance de notre environnement. [Ogden 1989] représente ce triangle sémiotique d'une manière un peu différente ; à savoir un triangle ABC dont :

- le sommet B est la référence (*ce que la chose est*) ;
- le point A, le signe (*ce qu'est la chose*) ;
- et le point C, le mot lexical (*ce qui désigne la chose*).

Un individu est confronté en un premier temps aux signes, lesquels peuvent être représentés (phase d'émission) sous des formes multiples. Leur perception (phase de réception) fait appel à nos sens - il n'y a à ce stade pas encore d'interprétation (nous sommes essentiellement au niveau de la syntaxe). [Peirce 1978] distingue trois types

de signes :

- l'*icône*, signifiant et signifié sont similaires (par exemple les émoticônes ou certains panneaux de circulation) ;
- l'*indice*, le caractère du signifiant est lié par contiguïté au signifié (par exemple la fumée pour indiquer le feu) ;
- et le *symbole*, qui demande d'être initié à la connaissance d'une règle fonctionnelle pour accéder au signifié (par exemple la croix verte pour indiquer une pharmacie) - c'est un signe établi par convention sociale et culturelle.

[Sebeok 1994] complète cette liste par trois types supplémentaires :

- le *signal*, signe naturel déclenchant une réaction chez l'individu récepteur ;
- le *symptôme*, où il existe un lien naturel et automatique entre un signifiant visible et un signifié (un état) ;
- et le *nom*, plus exactement le mot désignant l'objet (par exemple la suite de caractères o.i.s.e.a.u qui se prononce *waso* et qui désigne un volatile).

Le processus de passage de la forme au concept est qualifié non sans débats de *signification*. Le concept (ou sens, ou encore contenu sémantique) est abstrait et non accessible. Pour Saussure, la relation de signification est une relation établie entre la forme et le concept qui s'appellent l'un l'autre [Saussure 1962]. Si un individu n'a pas le concept en tête, alors sa perception du symbole soit restera sans suite, soit rentrera dans un processus d'apprentissage (et donc de stockage en mémoire) au moyen de recherche de similitudes avec des données connues par exemple. Ce processus de signification est hautement dépendant de l'éducation de l'individu émetteur comme récepteur. L'intension, ici développée, est l'image mentale associée à l'image visuelle ou sonore du mot (dans un cadre textuel par exemple).

La relation de *référence* fait passer de la suite de symbole *c.h.a.t* au concept de félin domestique mangeur de souris dans un premier temps, puis à la réalisation concrète de ce concept dans notre univers sous la forme de la boule de poils nommé Félix. Ce deuxième processus est nettement dépendant du contexte dans lequel le symbole sera émis puis reçu. L'objet peut être réel, imaginaire ou inimaginable.

[Peirce 1998] ajoute à cette triade un quatrième notion : le *fondement*. " *Un signe est quelque chose qui tient lieu pour quelqu'un de quelque chose sous quelque rapport ou à quelque titre. Il s'adresse à quelqu'un, c'est à dire crée dans l'esprit de cette personne un signe équivalent ou peut-être un signe plus développé. Ce signe qu'il crée, je l'appelle interprétant du premier signe. Ce signe tient lieu de quelque chose : de son objet. Il tient lieu de cet objet, non sous tous rapports, mais par référence à une sorte d'idée que j'ai appelée quelque fois le fondement du représentant. Il faut comprendre ici le terme idée ici dans une sorte de sens platonicien, courant dans le langage de tous les jours ; je veux dire dans le sens où nous disons qu'un homme saisit l'idée d'un autre homme : où nous disons quand un homme se souvient de ce qu'il pensait quelque temps auparavant, qu'il se souvient de la même idée, et où*

nous disons quand un homme continue à penser à quelque chose - ne serait-ce qu'un dixième de seconde, dans la mesure où la pensée continue à être cohérente pendant ce laps de temps, c'est-à-dire à avoir un contenu semblable - qu'il a la même idée, et que cette idée n'est pas à chaque instant de ce laps de temps une nouvelle idée. " [Morris 1946] ajoute également une quatrième dimension à cette triade : l'*interprète* - celui pour qui le signe a fonction de signe.

De ce fait, pour aller du mot à l'objet ou réciproquement, il semble qu'il faille passer par la référence, *i.e.* ce qui est contenu dans notre savoir. Le mot n'a de *sens* que par le passage par la référence, et *de facto* l'objet n'est analysable et sans doute perceptible dans son identité que par le même chemin. Le mot est donc attaché mentalement à une intension et cette intension renvoie à une extension [CAF 2002].

Il ne faut cependant pas confondre *dénotation* et *désignation*. Le premier établit le lien entre un lexème et l'ensemble complet des entités du monde que l'on peut désigner par ce lexème. Le deuxième est un acte de référence où le locuteur réfère à un certain individu (le référent) au moyen d'une expression référentielle.

Nous pouvons trouver dans la littérature bon nombre de références à des auteurs ayant travaillé sur ce triangle. [Eco 1993] a établi une synthèse des différents termes employés pour désigner les sommets de ce petit graphe (*cf.* figure 2.10).

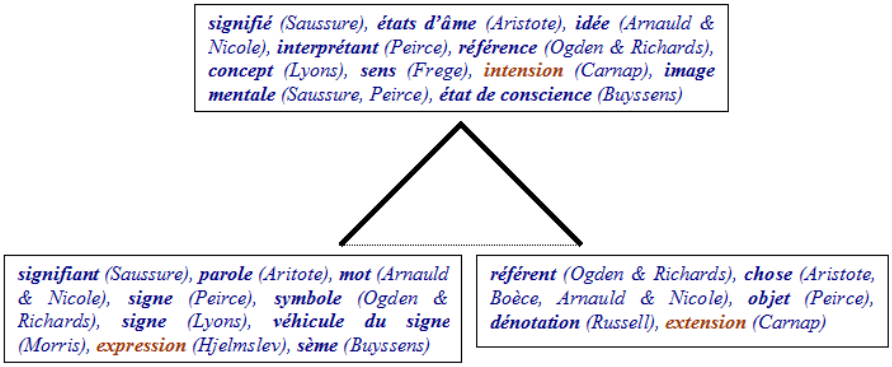


FIGURE 2.10 – Synthèse des triangles sémiotiques.

Un simple triangle modélise l'atome de la connaissance : le concept.

2.6.5 Les relations

Prenons le cas du concept *Canidé*, son intension est : *mammifère carnassier aux molaires nombreuses, aux griffes non rétractiles.* Cette intension est définie par l'utilisation d'autres concepts comme ceux de *Mammifère*, *Molaire*, et *Griffes non rétractiles*. Un canidé sans dents ou végétarien ne correspond plus vraiment à l'image que nous avons de cet animal. L'élaboration d'un concept semble ainsi reposer sur

d'autres concepts, soit en terme d'identité soit en terme de différence. Il est alors considéré comme un élément situé au sein d'un réseau structuré, réseau représentatif d'un domaine de connaissance où les concepts sont liés par des **relations**.

Ces relations traduisent des associations pertinentes existant entre les concepts présents dans le segment analysé, à savoir :

- *sous classe de* : relation de type généralisation/spécialisation ;
- *partie de* : relation de type agrégation/composition ;
- *associée à* ;
- etc.

La relation première, héritée de la vision aristotélicienne, est de type "*est un*". Il s'agit d'une relation de subsomption, qui définit un lien de spécialisation/généralisation. A ces relations s'ajoutent :

- des **fonctions** ;
- des **propriétés algébriques** telles la symétrie, la réflexivité, la transitivité ;
- une **cardinalité**.

Deux relations peuvent être :

- **incompatibles**, si elles ne peuvent pas être attestées sur le même concept (par exemple : *est parent de* et *est fils de*, on ne peut être à la fois fils et père du même concept) ;
- **inverses** : si elles sont l'inverse l'une de l'autre (par exemple : est plus long que, est plus court que) ;
- **exclusives** : si elles sont incompatibles ET si la négation de l'une entraine l'affirmation de l'autre.

Par rapport à la notion de relation, [Guarino 2001] définit un deuxième type de concept : le concept unité.

Définition [Concept unité]

- - si pour chaque instance de ce concept, les différentes parties qui le composent sont liées par une relation qui ne lie pas d'autres instances de concepts entre eux. Par exemple : pour notre cuillère, le manche et la partie creuse sont liées par une relation « emmanchées » ; cette relation ne lie que ces deux sous-parties.

2.6.6 Les propriétés

Comme indiqué ci-dessus, un concept peut être déterminé à l'aide d'autres concepts. Mais il peut également être caractérisé par des **propriétés** valuées. Ainsi, par exemple, les chats ont une couleur de pelage spécifique : gris (les Chartreux) ou noir (les Bombay) par exemple. Si nous établissons une ontologie sur les chats, nous aurons des concepts *Chartreux* ou *Bombay* pour lesquels est précisé le type

de couleur (gris ou noir). La valeur de cette propriété varie suivant le concept auquel il est fait référence. Citons un type de propriété particulier : l'**identité** ; il s'agit d'une propriété qui permet de différencier deux instances d'un même concept (par exemple, deux chats tatoués ne possèdent pas le même numéro d'identifcation.)

Par rapport à la notion de propriété, [Guarino 2001] définit deux autres types de concept : le concept rigide, et le concept anti-rigide.

Définition [Concept rigide]

- Concept possédant au moins un propriété essentielle [11] ; si une instance perd cette propriété, elle perd également son identité ; par exemple, soit h une instance du concept ETRE ; sa propriété *personne physique* est essentielle.

Définition [Concept anti-rigide]

- concept dont toutes les propriétés sont non essentielles ; par exemple, soit h une instance du concept ÊTRE, sa propriété « étudiant » est anti-rigide car une personne n'est pas forcément étudiante ou ne l'est pas toute sa vie.

2.6.7 Des ontologies légères aux ontologies denses

L'axiome, dans sa définition philosophique, est un principe de base d'évidence et non démontrable : un axiome est une expression qui est toujours vraie. Son utilité est multiple : définition de restriction quant à la valeur d'attributs, définition d'arguments d'une relation, validation et déduction d'information... Les axiomes, selon [Staab 2000], représentent les connaissances n'ayant pas un caractère strictement terminologique. Plus précisément, les axiomes permettent de fixer la sémantique des concepts et des relations. Prenons, par exemple, deux relations exclusives *Amis[Humain,Humain]* et *Ennemis[Humain,Humain]*. Telles quelles, ces deux relations ne disent pas grand chose. Ajoutons les règles sous-jacentes à ces relations, à savoir : (1) *Les amis de mes amis sont mes amis*, (2) *Les ennemis de mes ennemis sont mes amis* et (3) *Les ennemis de mes amis sont mes ennemis*. Ces axiomes permettent de fixer l'interprétation de ces relations et en précise le sens.

2.6.8 Définitions formelles des ontologies

[Maedche 2001] définit formellement une ontologie *concrète* comme étant une paire (O, Lex) où O est une ontologie abstraite et Lex un lexique pour O.

Définition [Ontologie] Soit un langage logique L ayant une sémantique formelle dans laquelle des règles d'inférence peuvent être exprimées. Une ontologie abstraite est une structure $O = (C, \leq^c, R, \sigma, \leq^R, IR)$ consistant en :

11. Une propriété est dite essentielle pour un objet s'il l'a possède durant tout le temps où il existe.

- deux ensembles disjoints C et R dont les éléments sont respectivement appelés *Concepts* et *Relations* ;
- un ordre partiel $\leq^C$ sur C, appelé *hiérarchie de concepts* ou taxinomie ;
- une fonction $\sigma : C \times C$ appelée *signature* ;
- un ordre partiel $\leq^R$ sur R, appelé *hiérarchie de relations*, où $r_1 \leq^R r_2$ implique $sigma(r_1) \leq^{C \times C} \sigma(r_2)$ avec $r_1, r_2 \in R$.
- un ensemble IR de règles d'inférences exprimées dans le langage logique L ;
- la fonction $dom : R \to C$ avec $dom(r) = \Pi_1(\sigma(r))$ retourne le domaine de r ;
- la fonction $range : R \to C$ avec $range(r) = \Pi_2(\sigma(r))$ retourne l'échelle de valeurs de r.

Définition [Lexique]
Un lexique pour une ontologie abstraite $O = (C, \leq^C, R, \sigma, \leq^R, IR)$ est une structure $Lex := (S_C, S_R, Ref_C, Ref_R)$ qui consiste en :

- deux ensembles S_C et S_R dont les éléments sont appelés *signes* respectivement pour des concepts et des relations ;
- deux relations $Ref_C \subseteq S_C \times C$ et $Ref_R \subseteq S_R \times R$, appelées *affectation de référence lexicale* respectivement pour des concepts et des relations ;
- à partir de Ref_C nous définissons $\forall s \in S_C, Ref_C(s) = c \in C | (s, c) \in Ref_C$ et $Ref_C^{-1}(s) = s \in C | (s, c) \in Ref_C$
- à partir de Ref_R nous définissons $\forall s \in S_R, Ref_R(s) = r \in R | (s, r) \in Ref_R$ et $Ref_R^{-1}(s) = s \in R | (s, r) \in Ref_R$

2.7 Construction d'une ontologie

2.7.1 Des principes...

On pourrait retenir neuf principes majeurs que l'on retrouve dans de nombreuses publications, la plupart ayant été énoncés par [Gruber 1993b] :

1. *clarté et objectivité* : la définition intensionnelle de chaque concept doit être fournie en langage naturel et objective, les ambiguïtés réduites au maximum ;
2. *complétude* : le définitions intensionnelles de chaque concept doivent être exprimées par toutes les conditions nécessaires et suffisantes à ce concept ;
3. *cohérence* : une ontologie doit être cohérente, les axiomes doivent être consistants et la cohérence des définitions en langage naturel doit être vérifiée ;
4. *extensibilité maximale* : une ontologie doit pouvoir être étendue sans que cela remette en cause le travail effectué auparavant ;
5. *engagements ontologiques minimaux* : une ontologie doit faire un minimum d'hypothèses sur le monde ; elle doit contenir un vocabulaire partagé mais ne doit pas être une base comportant des connaissances supplémentaires sur le monde à modéliser.

6. *diversification des hiérarchies* : plus une ontologie sera détaillée dans ses concepts, plus il sera facile d'y faire rentrer de nouveaux concepts par des liens d'héritages multiples entre autre ;
7. *distance sémantique minimale* entre les concepts enfants d'un même parent : regrouper au maximum les concepts similaires, et créer des sous-classes si nécessaire ;
8. *normalisation* des noms ;
9. *biais d'encodage minimal* : une ontologie doit être conceptualisée indépendamment de tout langage d'implémentation ; il ne faut pas perdre de vue que son objectif est de permettre le partage des connaissances entre des agents (humains ou non) utilisant des langages de représentation différents.

[Noy 2000], dans un document également diffusé par le bureau de normalisation documentaire de la Bibliothèque Nationale de France (BNF), ajoute également les critères suivants :

- il n'y a pas une unique façon correcte de modéliser un domaine, il y a toujours des alternatives viables. La meilleure solution dépend presque toujours de l'application que vous voulez mettre en place et des évolutions que vous anticipez.
- le développement d'une ontologie est nécessairement un processus itératif.
- les concepts dans une ontologie doivent être très proches des objets (physiques ou logiques) et des relations dans votre domaine d'intérêt. Fort probablement ils sont des noms (objets) ou verbes (relations) dans des phrases qui décrivent votre domaine.

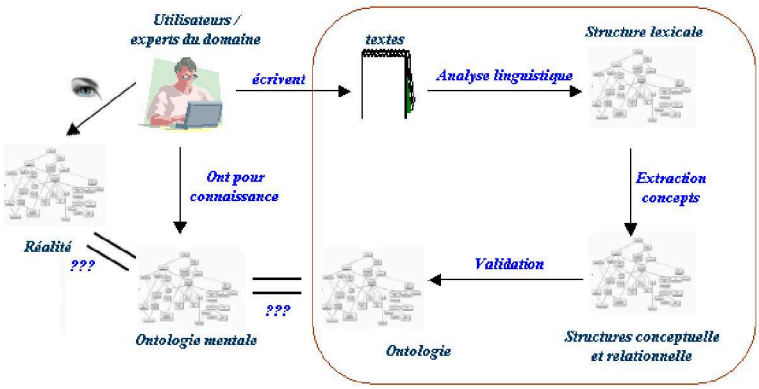


FIGURE 2.11 – Processus de construction d'ontologies à partir de textes.

[Roche 2007] souligne le problème illustré par la figure 2.11 : "*Si l'on considère que les documents scientifiques et techniques d'un domaine en véhiculent les connaissances, l'extraction d'ontologies à partir de textes semble être une idée des plus prometteuses. L'idée principale est que les termes dénotent des concepts, et que les*

relations linguistiques entre termes traduisent une relation entre concepts. (...) Un tel processus soulève néanmoins un certain nombre de questions. La principale est de savoir dans quelle mesure une ontologie construite à partir de textes et une ontologie définie directement par les experts sont comparables. Autrement dit : quelles sont les conséquences de l'utilisation d'un langage donné, qu'il soit naturel ou formel ? que perdons-nous (et introduisons-nous) lors de l'écriture de textes et en quoi les langages formels conditionnent la conceptualisation ?" L'objectif va donc être de réduire au maximum l'écart entre la connaissance des experts et les ontologies générées.

2.7.2 ... et des étapes

Selon [Charlet 2005], *les méthodologies de construction d'ontologies ne sont pas légion. Nous entendons par-là, la donnée argumentée de procédures de travail, d'étapes, qui décrivent le pourquoi et le comment de la conceptualisation puis de l'artefact construit.* Au-delà de la diversité des nombreuses publications sur cette problématique, un grand nombre d'étapes restent communes aux différentes méthodologies, et un certain consensus semble émerger sur les phases majeures du cycle de vie d'une ontologie (*cf.* figure 2.12).

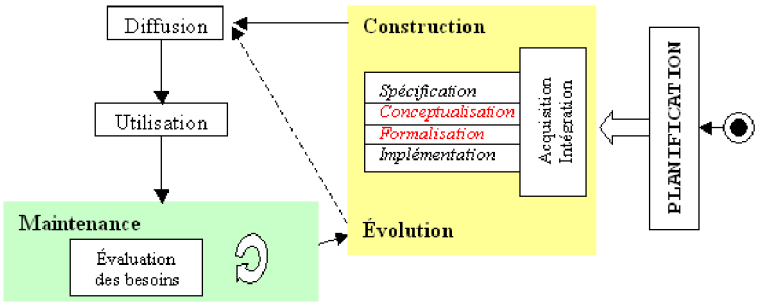


FIGURE 2.12 – Cycle de vie d'une ontologie [Brisson 2004].

2.7.3 Le cycle de vie d'une ontologie

[López 2000] présente les ontologies informatiques comme des composants logiciels, ce qui implique *de facto* que leur développement doit être en accord avec les standards proposés pour une telle activité. Pour ce faire il montre comment suivre - voire adapter si nécessaire - le processus de développement logiciel défini dans la norme *IEEE 1074-1995*[12], à savoir :

- utiliser un cycle de vie logiciel prédéfini par la norme ;
- établir un management de projet ;
- respecter un processus orienté développement logiciel avec :
- *pré-développement, i.e.* étude de faisabilité, cahier des charges ;

12. http ://standards.ieee.org/

- *développement*, *i.e.* collecte des données, conceptualisation, formalisation et enfin implémentation ;
- *post-développement*, *i.e.* évaluation, intégration, documentation ;
- former toutes les personnes concernées par ce produit, des utilisateurs à celles chargées de la maintenance.

L'objectif et l'intérêt majeur d'une telle démarche sont :

- d'une part, de se calquer sur une technique éprouvée et fiable ;
- d'autre part, de favoriser le repérage et la réparation d'erreur, en évitant leur propagation tout au long du processus de développement grâce à une politique d'évaluation tout au long du processus de développement.

La construction d'une ontologie comporte principalement six étapes :

1. *évaluation des besoins* ;
2. *collecte* ;
3. *pré-analyse* ;
4. *modélisation* ;
5. *formalisation* ;
6. *validation.*

2.7.4 Étape N°1 - l'évaluation de besoins

L'objectif de cette étape est de définir les buts et surtout les limitations de l'ontologie que nous souhaitons développer. La construction d'une ontologie se décline selon trois vues avec pour chacune les questions importantes que nous devons nous poser (*cf.* tableau 2.1).

TABLE 2.1 – Questions

Aspect de la construction d'une ontologie	**Questions à poser**
L'objectif opérationnel	*Dans quel but utiliser cette ontologie ? A quels types de questions cette ontologie devra répondre ?*
Le domaine de connaissances	*Quel domaine ? Quelles compétences ?*
Les utilisateurs	*Qui va utiliser l'ontologie ? Qui va maintenir l'ontologie ?*

Il s'agit ici de pouvoir répondre aux différentes questions posées, à savoir :

- *problématique* ;

- *objectif opérationnel* ;
- *domaine de connaissance* ;
- *utilisateurs.*

2.7.5 Étape N°2 - la collecte des données

La collecte de données est la première étape de conceptualisation dans la construction d'une ontologie. Selon [López 1996], il existe plusieurs techniques pour la réaliser :

- les entretiens avec des experts du (ou des) domaine(s) ;
- *et/ou* la composition d'un corpus de textes [13].

Les entretiens avec les experts permettent d'obtenir deux types d'information :

- une ébauche de spécifications ;
- un inventaire détaillé des connaissances spécifiques à leur domaine (concepts, relations, modèles existant).

Il faut cependant être vigilant quant à la qualité des informations. C'est pourquoi une phase de normalisation est nécessaire (cf. étapes N°3a et 3b). Il faut par conséquent essayer d'une part de rapprocher les différentes représentations mentales des experts, et d'autre part de les faire converger vers une vision la plus objective possible (*i.e* indépendante des émotions, de la culture, de l'éducation, etc.). Le consensus est parfois très difficile à trouver, la culture qu'elle soit d'entreprise ou sociale possédant un poids très important sur l'élaboration des connaissances. A cela s'ajoute la diffficulté d'expliciter certaines connaissances. [Bourigault 2000b] insiste ainsi sur le fait que "*les experts ne sont pas en mesure de verbaliser explicitement et complètement un ensemble de termes qui serait la traduction verbale d'un système explicite de concepts dont ils auraient une conscience suffisante pour formuler un verdict définitif.*"
La deuxième méthode, selon [Condamines 2005] indépendante ou complémentaire de la précédente, réside dans la *constitution d'un corpus de textes*. Cette opération se révèle à la fois délicate et d'une importance capitale, car c'est à partir de ces documents que la connaissance est extraite. Se pose alors l'une des questions posées par C. Roche : les textes sont le fruit d'agents humains et de leur langage, quelles sont dès lors les informations perdues ou introduites lors de la phase d'écriture de tels textes ? Seul un travail minutieux avec les experts peut apporter tout ou partie de la réponse. [Bourigault 2003] dit ainsi que : " *au-delà des problèmes techniques ou politiques de disponibilités des textes, cette collecte doit se faire avec l'aide de spécialistes et en fonction de l'application cible visée. Il convient en effet*

13. Nous entendons par corpus, un ensemble de documents utilisés pour une étude, spécialement pour une étude linguistique. Pour [Roche 2007], il s'agit d'un ensemble de textes de même *type* ou *genre* : contenu, conditions de production, pratiques langagières, communautés de pratique d'origine.

de s'assurer auprès de spécialistes que les textes choisis ont un statut suffisamment consensuel pour éviter toute remise en cause ultérieure de la part d'utilisateurs ou de leur part. " En terme de taille, l'ensemble doit être suffisamment important pour couvrir de manière satisfaisante le domaine, mais raisonnablement petit pour pouvoir être appréhendé par un analyste. [Bourigault 2003] estime la taille moyenne d'un corpus entre 50.000 et 200.000 mots.

2.7.6 Étape N°3a - l'étude linguistique

L'analyse d'un corpus de textes - corpus de référence - se fait dans un premier temps de manière informelle afin de pouvoir en extraire les principaux concepts et ébaucher une première représentation des connaissances. Elle se fait, dans un second temps, de manière plus formelle pour pouvoir en retirer les différents concepts sous forme de triplet (terme, intension, extension) ainsi que les relations associées. Cette phase est réalisée au travers de l'extraction puis de l'analyse de termes. Citons, entre autre, la méthodologie TERMINAE, de [Aussenac-Gilles 2000], qui propose une approche afin d'extraire les concepts, leurs propriétés, les relations et leur regroupement.

[Bourigault 2003] présente l'étude linguistique comme consistant à *identifier des termes et des relations lexicales, en utilisant des outils de traitement de la langue naturelle (SYNTEX comme extracteur de termes, UPPERY comme outil d'analyse distributionnelle, CAMALEON pour l'aide au repérage de relations par des patrons linguistiques*[14]*, YAKWA comme concordancier).*

Il peut être intéressant de définir avec précision la nature d'un terme. Du point de vue *Théorie Générale de la Terminologie*[15], le terme serait le *représentant linguistique d'un concept dans un domaine de connaissance.* [Bourigault 2000b] estime qu'il s'agit *du résultat d'un processus d'analyse terminologique. Un mot ou une unité complexe n'acquiert le statut de terme que par décision.* Pour [Roche 2007], *tout terme (désignation) d'une terminologie est un mot d'usage de la langue de spécialité (LSP), la réciproque n'est pas nécessairement vraie. La construction d'ontologies à partir de textes devrait tenir compte de cette distinction.* Parmi les travaux sur l'extraction de candidats termes[16] (ainsi que de relations linguistiques) à partir d'un corpus, nous pouvons citer [Aussenac-Gilles 2005, Buitelaar 2005, Daille 2004], lesquels reposent sur des méthodes statistiques[17] (analyse distributionnelle Harris [Harris 1968]) et/ou linguistique[18].

14. Ou patrons lexicaux-syntaxiques du type *substantif adjectif* ou *substantif préposition substantif*...
15. Théorie fondée par Eugène Wüster à la fin des années 1930, impliquant la linguistique, la logique, l'ontologie, l'informatique et les sciences des objets.
16. Selon [Bourigault 2000b], mots ou groupes de mots susceptibles d'êtres retenus comme termes par un analyste, et de fournir des étiquettes de concepts.
17. Fréquence d'apparition de ces mots dans les textes.
18. Analyse du rôle grammatical des mots dans les textes.

2.7.6.1 Approches syntaxiques

Pour [Hernandez 2006], il s'agit d'extraire des termes (*i.e.* un ou plusieurs mots) au moyen de relations grammaticales entre les mots dans les phrases des documents. Il est ainsi possible d'utiliser - après lemmatisation des documents - des expressions régulières ou patrons lexicaux-syntaxiques ; l'objectif étant le repérage de syntagmes nominaux (*SN*) et verbaux (*SV*). [Moigno 2002] définit ces syntagmes de la manière suivante : *un syntagme verbal (resp. nominal, adjectival) est un groupe de mots dont la tête syntaxique est un verbe (resp. nom et adjectif).* Par exemple, *valoriser les piles usagées* est un syntagme verbal dont la tête syntaxique est le verbe *valoriser* et l'expansion le syntagme nominal *piles usagées*. De même ce dernier a pour tête le nom *piles* et pour expansion l'adjectif *usagées*. Au travers de ces relations linguistiques, nous pouvons par la suite extraire des relations sémantiques. Plusieurs analyseurs syntaxiques existent, citons parmi eux SYNTEX [Bourigault 2000a]. Il s'agit d'un analyseur de corpus spécialisés issu du projet *LEXTER* de [Bourrigault 1994], dont l'un des intérêts est qu'il s'appuie sur un apprentissage endogène du corpus, ce qui lui permet d'être plus performant qu'un analyseur reposant uniquement sur des règles définies manuellement. Syntex prend en entrée des textes étiquetés morpho-syntaxiquement [19] puis effectue une analyse syntaxique du corpus [20]. Les relations syntaxiques relevées sont :

- sujet ;
- complément objet direct ;
- complément prépositionnel [21] (xxx que je xxx, dont nous xxx, etc.) ;
- antécédence relative ;
- modification adjectivale (épithète, attribut) ;
- subordination.

Il s'agit enfin de sélectionner les syntagmes les plus représentatifs du corpus relatif au domaine. Pour ce faire deux modes peuvent être possibles : soit une sélection en raison de leur nature grammaticale, soir une sélection en rapport avec leur poids par rapport au corpus. Pour ce faire, il est possible d'utiliser la *mesure d'information mutuelle* [Velardi 2002] qui sélectionne les syntagmes dont les termes sont les plus liés :

$$IM(x,y) = \frac{nb(x,y)}{nb(x)*nb(y)}$$

où x et y sont deux termes composants un syntagme, $nb(a)$ est le nombre d'apparitions du terme a dans le corpus, et $nb(a,b)$ le nombre d'apparition du terme a avec

19. *I.e.* une étiquette relatif à la nature de chaque mot, des outils comme *TreeTagger*, développé à l'Université de Stuttgart, peuvent être utilisés à cet effet mais nous considérerons ce processus comme étant semi-automatique.
20. Identification de constituants syntaxiques ou de relation de dépendance syntaxique.
21. De nom, de verbe, d'adjectif.

le terme b dans ce même corpus. Ainsi, si l'un des deux termes est plus fréquent que l'autre, la mesure aura une valeur faible.

2.7.6.2 Approches statistiques

Cette démarche est bien différente de la précédente. La première étape consiste à supprimer tous les mots vides de sens (articles, prépositions, conjonctions...) en considérant soit un dictionnaire, soit leur longueur (inférieure à deux caractères par exemple). La seconde étape vise à supprimer les différentes variantes lexicales des mots restant en ne prenant en compte que leur racine. Par exemple, du paragraphe *Les chefs d'états réunis cette semaine pour le Forum international pour la conservation du tigre viennent d'approuver la stratégie de financement et de suivi du plan de restauration des Tigres* ne sont retenus que les termes *chefs d'états réunis semaine Forum international conservation du tigre approuver stratégie de financement suivi plan de restauration des Tigres.*

Pour [Hernandez 2006], il est possible de ne retenir que les mots (dits *termes individuels*) ou des expressions (*séquence de mots juxtaposés*). Tout comme dans la méthode syntaxique, il faut ensuite opérer une sélection à partir de leurs occurrences dans les documents composant le corpus. Pour ce faire plusieurs mesures peuvent être utilisées, citons par exemple *tf-idf* dans [Robertson 1976] ou l'*entropie* dans [Brini 2005]. Plusieurs expérimentations ont été menées, lesquelles ont abouti à des conclusions contradictoires - les domaines couverts ayant apparemment une forte influence sur les résultats obtenus. Selon Hernandez, aucune étude n'a conclu sur le choix entre une extraction syntaxique ou statistique des expressions. Ce choix reste donc difficile. Cependant, intuitivement, on peut penser que pour la construction d'ontologies, l'extraction syntaxique est la plus adaptée. Nous retenons de plus que les mesures de sélection des termes dépendent fortement du domaine traité et du corpus de référence confectionné. L'intervention d'experts pour la validation des termes est nécessaire.

L'engagement ontologique minimal impose que les données collectées (et conservées) portent uniquement sur le domaine concerné. Pour [Fürst 2002], il est important, avant de passer aux prochaines étapes, d'opérer un tri entre données spécifiques au domaine et celles qui ne font que participer à l'expression des connaissances du domaine. Cette partie du développement fournit un ensemble de termes, soit le premier élément du triplet (**terme**, intension, extension) formant un concept. Nous obtenons un vocabulaire organisé, un lexique de termes de la terminologie et de mots d'usage de la langue de spécialité qui devra être validé par les experts. Mais comme le souligne [Bourigault 2000b], il n'existe pas *une* terminologie qui représenterait le savoir sur le domaine, mais autant de terminologies que d'applications dans lesquelles ces terminologies sont utilisées !

Enfin, il y a toutes ces connaissances qui ne sont pas exprimées explicitement, ni dans les textes, ni par les experts, car elles sont évidentes pour les auteurs des premiers comme pour les seconds. Or, elles devraient être présentes au sein de l'ontologie produite. [Lame 2002] suggère de passer d'une version tacite à une version explicite de ces connaissances en les capturant sur un support textuel à partir d'interviews. L'élaboration d'une ontologie revient dès lors à (1) des interviews, et (2) à de l'analyse de corpus de textes, avec les inconvénients que nous venons d'évoquer si nous partons du principe que tout processus d'écriture apporte un biais par rapport aux connaissances initiales.

Toutes ces opérations s'effectuent dans le cadre d'une construction d'ontologies *ex nihilo*. Mais, on peut considérer une ontologie existante, soit en la complétant soit en l'adaptant à des besoins spécifiques.

2.7.7 Étape N°3b - l'étude sémantique

Cette étape, deuxième de la phase de conceptualisation, porte le nom de **normalisation sémantique** [Bachimont 2000]. S'effectuant en collaboration étroite avec les experts du domaine, elle répond à plusieurs objectifs :

- associer aux termes une intension ;
- et fournir une terminologie fiable, rigoureuse et commune à tous les acteurs pour les traitements suivants (modélisation).

Par exemple, pour désigner le dirigeant d'une société nous pouvons avoir de multiples dénominations comme : son nom, son surnom, sa fonction (l'acronyme PDG ou Président Directeur Général), les usages propres à chaque groupe, etc. La normalisation sémantique va impliquer le choix d'une forme canonique pour identifier le concept désigné.
[Uschold 1996] propose de découper ce processus en trois parties successives (*cf.* figure 2.13).

2.7.7.1 Identification des clés de concepts et des relations avec le domaine

D'après B. Bachimont, le corpus contient l'expression des notions qu'il faut modéliser [Bachimont 2000]. Il en découle la nécessité de définir les concepts par les libellés linguistiques rencontrés dans le corpus. Mais choisir un terme pour un concept, c'est lui associer une interprétabilité. Idéalement, il faudrait s'assurer que - pour un domaine précis - l'interprétation du terme choisi sera toujours constante quelque soit le spécialiste utilisant ou consultant l'ontologie.
Le but étant de définir des primitives, qui sont par définition indépendantes du contexte d'utilisation de l'ontologie, il est nécessaire d'appliquer des contraintes d'utilisation afin de décontextualiser certains termes. Le problème est de partir de la sémantique de la langue naturelle pour arriver à la définition non contextuelle d'un

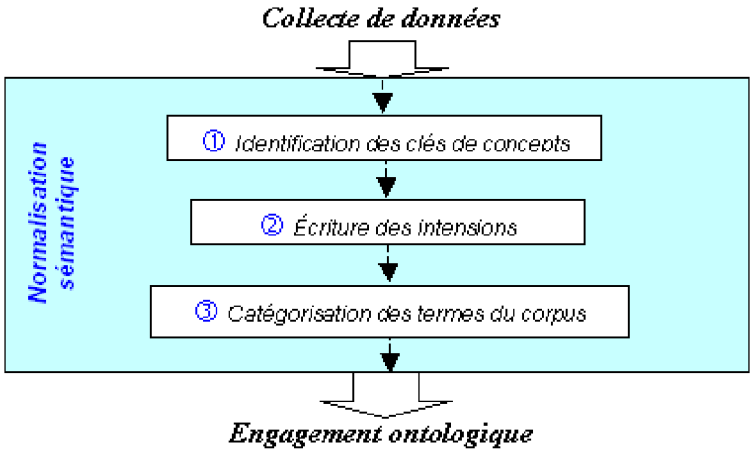


FIGURE 2.13 – Normalisation sémantique.

libellé de concept. Pour ce faire, [Bachimont 2000] suggère l'utilisation du paradigme différentiel.
Le paradigme différentiel sert à exprimer une chose par ce qu'elle peut être et ce qu'elle n'est pas. Il s'agit de définir chaque concept en indiquant en quoi il est semblable à ses concepts parents et en quoi il est différent de ses concepts frères.

2.7.7.2 Définition d'une intension précise et non ambiguë pour chaque concept

Cette définition peut se décomposer en plusieurs actions, à savoir :
- écrire son intension aussi clairement que possible, en langage naturel (attention aux définitions circulaires) ;
- s'assurer de la complétude de l'intension ;
- s'assurer de la consistance avec les termes existants ;
- indiquer les relations avec les autres termes.

2.7.7.3 Catégorisation des termes du corpus qui désignent chaque concept

Ce processus de catégorisation revient à :
- classer les termes approximativement par catégories (créer au besoin des subdivisions) ;
- repérer tous les cas limites (homonymes, synonymes, ceux qui font appel à plusieurs concepts, etc.) et les traiter ;
- identifier les liens sémantiques entre les catégories.

Cette étape fournit un ensemble de termes et d'intensions, soit les deux premiers éléments du triplet (**terme, intension**, extension) formant un concept.

2.7.8 Étape N°4 - création des concepts et d'une taxinomie

Cette étape - marquant la fin de la conceptualisation et le début de l'**ontologisation**. Les définitions de cet engagement varient suivant les auteurs. Pour N. Guarino, il s'agit d'une relation entre un langage logique et un ensemble de structures sémantiques, le sens du concept étant donné par son extension dans l'univers d'interprétation du langage [Guarino 1998]. Pour B. Bachimont, l'engagement ontologique est l'association des extensions aux couples (termes, intensions) [Bachimont 2000]. Pour T. Grüber, c'est une garantie de cohérence entre l'ontologie et son domaine, mais pas une garantie de complétude [Gruber 1993a].
L'apparition des extensions termine la phase de conceptualisation par l'obtention du triplet (**terme, intension, extension**) qui forme les concepts. L'ajout des relations entre concepts, puis d'axiomes, peut suivre et conduire à une ontologie dense. L'ontologisation mène à la construction d'une hiérarchisation de concepts avec relations et propriétés de concepts.
Cette hiérarchisation peut être représentée sous la forme d'un graphe. Pour bâtir un graphe à partir des concepts élaborés jusque-là, trois approches sont possibles :

- *bottom-up* : procédé de développement de bas en haut par généralisation en partant des termes les plus spécifiques, avec un fort niveau de détail. Par exemple, nous partons des concepts *Chartreux* et *Siamois*, que nous généralisons en *Chat* qui lui même sera généralisé en *Felis*, etc.
- *top-down* : procédé de développement de haut en bas par spécialisation en partant des termes les plus génériques, avec un haut niveau d'abstraction. Par exemple, nous partons des concepts *Végétariens* et *Carnivores*, puis spécialisons ce dernier en *Ursidés*, *Canidés* et *Félidés*, lequel est subdivisé en *Chats*, *Panthères*, etc.
- *middle-out* : procédé combiné de développement par généralisation et spécialisations de concepts intermédiaires, avec une forte modularité thématique.

La première version est préconisée par M. Uschold, alors que [Bachimont 2000] propose de construire le graphe suivant les principes différentiels énoncés précédemment, où la signification du nœud se détermine en fonction de ses plus proches voisins (les parents et les unités soeurs). Pour cela, il définit quatre principes :

1. *principe de communauté avec le père (similarité)* : tout concept partage l'intension de son concept père ; définition par le genre proche.
2. *principe de différence avec le père* : tout concept a une intension différente de son concept père (sinon il n'y aurait pas besoin de le définir) ; définition par le genre spécifique.
3. *principe de communauté avec les frères* (sémantique unique) : tout concept enfant possède un (ou des) propriété(s) commune(s) aux concepts frères issus du même concept père, mais s'exprimant différemment ; définition par les différences mutuellement exclusives. Par exemple, la propriété *sexe* issue du concept *humain* est égale à masculin pour le concept enfant *homme* et féminin pour le concept enfant *femme*.

4. *principe de différence avec les frères* : tous les concepts frères issus d'un même père doivent être distincts (sinon il n'y aurait pas besoin de les définir).

2.7.9 Étape N°5 - formalisation

Il s'agit de la dernière étape de traduction des connaissances, en provenance de données exprimées de manière totalement informelle en langage naturel, en notions compréhensibles par des agents par l'intermédiaire de langages plus ou moins formels. Lors de ce processus nous devons coder les différentes classes, les attributs, les types, les contraintes, etc. Il reste ensuite à peupler l'ontologie au moyen d'instances. Pour [Brisson 2004], suivant que les agents seront humains ou non, et suivant le système opérationnel dans lequel l'ontologie sera insérée, nous nous devons d'adapter le degré de formalisme de notre ontologie :

- *degré très informel*, expression de l'ontologie en langage naturel ;
- *degré semi-formel*, (1) soit par expression de l'ontologie dans une forme restreinte et structurée de langage naturel, (2) soit expression de l'ontologie dans un langage artificiel défini formellement ;
- *degré formel*, expression de l'ontologie en termes utilisant une sémantique formelle, théorèmes et preuves.

Cette dernière forme implique de traduire l'ontologie obtenue à l'étape précédente en une ontologie computationnelle, *i.e.* compréhensible par une machine. Il est important en ce cas de garder une version informelle en complément.

2.7.10 Étape N°6 - validation

Il s'agit de la dernière étape dans la construction d'une ontologie. Lors de la première phase, nous avons défini un certain nombre d'objectifs opérationnels. Avec des questions comme : *dans quel but utiliserez vous cette ontologie ? à quels types de questions l'ontologie devra-t-elle fournir des réponses ?* Cette liste n'est bien entendu pas exhaustive, mais elle nous aide à établir la liste des points à tester, à évaluer ; sans oublier les questions de compétences définies par [Fox 1998], qui constituent l'élément clé pour caractériser de façon rigoureuse les connaissances que doit inclure une ontologie. Et la validation doit se faire à toutes les étapes en amont de manière à éviter non seulement la propagation des erreurs, mais aussi de reprendre un travail rendu caduque.

2.8 Quelques outils de construction d'ontologies

Selon un état de l'art du web sémantique réalisé en 2007 par [Cardoso 2007], l'éditeur d'ontologies le plus utilisé était Protégé avec près de 68,2% des utilisateurs. Les outils propriétaires - encore en faible nombre à cette époque - ne représentaient que 2 à 3% chacun. Il est à penser aujourd'hui, avec l'utilisation croissante

des ontologies et l'avènement du web des données, que cette proportion a quelque peu évolué du fait de l'apparition de nouveaux éditeurs et la nécessité de travailler sur des ensembles de données volumineux, même si Protégé reste majoritairement employé et cité. Nous présentons, dans cette section, trois des cinq éditeurs d'ontologies étudiés dans [Norta 2010] pour le projet ContentFactory[22]. Ce rapport a évalué ces outils suivant une vingtaine de critères (flexibilité, interopérabilité, performance, etc.). Les trois outils que nous avons retenus sont : PROTÉGÉ[23], NEON[24] et TOPBRAID COMPOSER[25]. Ils ont la particularité d'être disponibles en version gratuite (PROTÉGÉ est issu du monde universitaire, NEON et TOPBRAID sont des productions industrielles). Ces trois éditeurs possèdent en outre les fonctionnalités suivantes :

- édition et visualisation d'ontologies ;
- import et export en différents formats / langages ;
- extensions par plug-ins ;
- conversion, intégration et annotation ;
- intégration d'axiomes et règles ;
- merging d'ontologies.

2.8.1 PROTÉGÉ

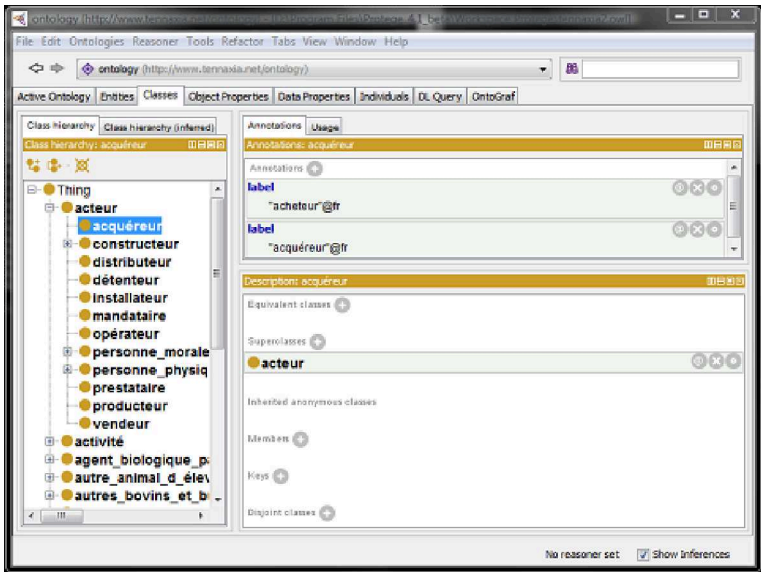


FIGURE 2.14 – Copie d'écran de l'éditeur Protégé.

22. http ://www.verkko-ope.net/cf/
23. http ://protege.stanford.edu
24. http ://neon-toolkit.org/
25. http ://www.topquadrant.com

Protégé est un éditeur d'ontologies (*cf.* figure 2.14) développé et distribué en open source par l'Université de Stanford. Il ne s'agit pas uniquement d'un outil dédié à OWL, c'est également un éditeur modulaire capable de manipuler bon nombre de formats par l'intermédiaire de plugins (y compris pour la visualisation). Protégé, très utilisé notamment par les biologistes, est issu d'une communauté scientifique composée de ces derniers et d'informaticiens utilisant le langage des frames[26] pour faire de la logique. Cet outil est assez puissant et permet de réaliser une ontologie (par défaut en OWL DL, avec possibilité d'extension en OWL Full) relativement aisément à l'aide de nombreuses fenêtres dédiées (classes, propriétés, instances). [Norta 2010] a relevé les caractéristiques citées dans le tableau 2.2.

Table 2.2 – Caractéristiques de Protégé.

OS	Sur tout système disposant d'un VM java.
Langages supportés	OWL, RDF
Multilinguisme	Oui
Vérification d'ontologie	Non
Visualisation	OntoViz, OWLViz, Jambalaya, TGViz
Développement collaboratif d'ontologies	Oui (Colaborative Protégé et WebProtégé)
Interopérabilité (web services...)	Pas d'échanges avec d'autres applications
Outil modifiable	Très flexible et paramétrable
Performance	Moins d'une minute pour toute opération
Sécurité	Peut gérer une politique d'utilisateurs
Flexibilité	Large communauté continue de développer
Facilité d'utilisation	Utilisable essentiellement par des experts

2.8.2 Néon

Néon est un environnement gratuit et Open Source pour le développement d'ontologies (*cf.* figure 2.15), conçu et distribué par l'entreprise *Ontoprise*. Il est fondé sur OntoStudio, l'une des briques de leur suite logicielle de management des connaissances (et de recherche d'information) à destination des entreprises. Il dispose d'un outil graphique dédié à la visualisation, et d'un serveur pour une utilisation dans un environnement multi-utilisateurs. La valeur ajoutée de cet outil réside (1) dans sa très forte modularité dû à l'environnement Eclipse dans lequel il évolue, (2) dans le nombre important de plugins à disposition, et (3) dans ses outils collaboratifs de création d'ontologie (wiki, méthodologie, planning). Il est ainsi l'un des rares à ce jour à implémenter à la fois une partie management et une partie données. [Norta 2010] a relevé les caractéristiques citées dans le tableau 2.3.

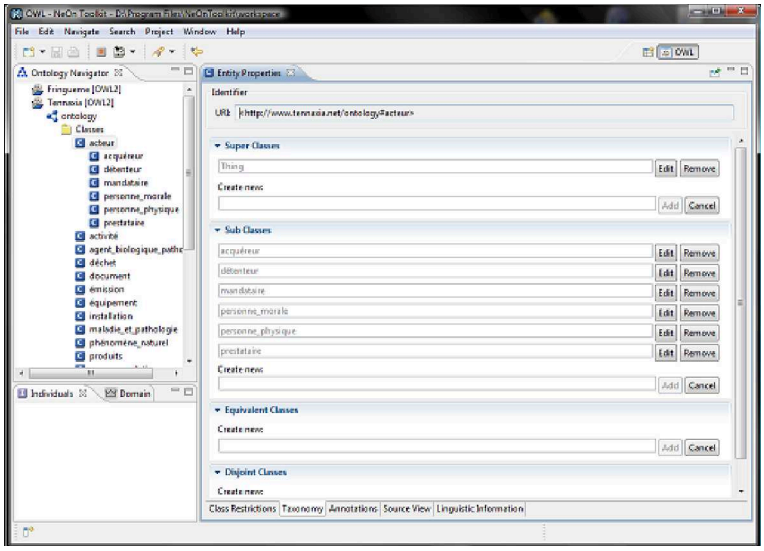


FIGURE 2.15 – Copie d'écran de l'éditeur Néon.

TABLE 2.3 – Caractéristiques de Neon.

OS	Windows, Mac OS, Linux
Langages supportés	OWL, RDF, F-logic, SPARQL
Multilinguisme	Oui, plugin lexical très complet
Vérification d'ontologie	Oui, test de consistance et de cohérence
Visualisation	Nombreux modes graphiques ou tabulaires
Extraction d'information automatique	Bases de données et ressources XML
Développement collaboratif d'ontologies	Oui, plugins + serveur + wiki
Interopérabilité (web services...)	Pas d'échange avec d'autres applications
Outil modifiable	Oui, environnement Eclipse
Performance	OK pour ontologies volumineuses
Sécurité	Management des accès
Flexibilité	Tout passe par le gestionnaire de plugins
Facilité d'utilisation	Intuitif et nombreuses aides

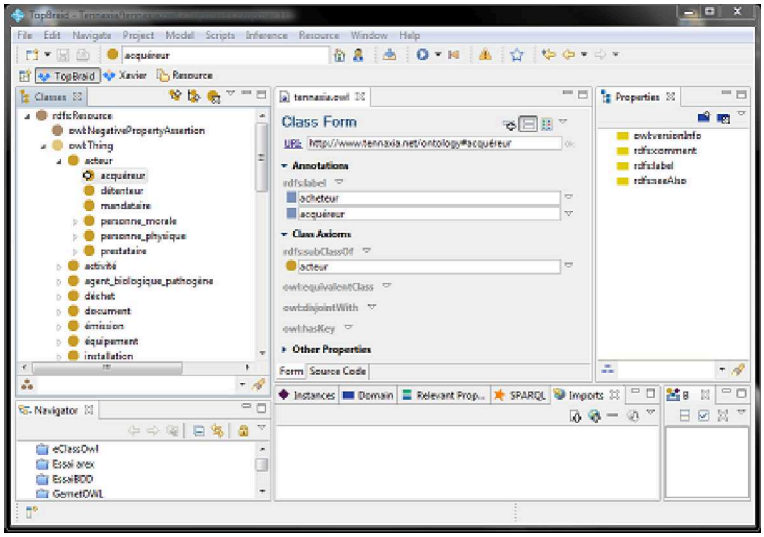


FIGURE 2.16 – Copie d'écran de l'éditeur TopBraid Composer.

2.8.3 TopBraid Composer

TOPBRAID COMPOSER est un environnement pour le développement d'ontologies (*cf.* figure 2.16), conçu et distribué commercialement par l'entreprise TOPQUADRANT. Disponible gratuitement dans sa version de base (déjà très complète), TopBraid, sous un environnement Eclipse, offre toutes les fonctionnalités pour développer et maintenir une ontologie, mais également effectuer du raisonnement. Fondé historiquement sur le module de création d'ontologies de Protégé, TopBraid gère parfaitement les ontologies volumineuses. La version gratuite offre de très nombreuses fonctionnalités utiles à la création d'une ontologie (y compris ontologies denses) avec tous les avantages d'un produit propriétaire en terme d'optimisation et de stabilité. Relativement rapide et fonctionnel, cet outil bénéficie de plus des plugins Eclipse tels *subversive* qui lui permet par exemple de travailler en subversion. [Norta 2010] a relevé les caractéristiques citées dans le tableau 2.4.

2.9 En conclusion

Le *concept* d'ontologie regroupe de multiples faces. Il n'est pas réducteur et simpliste d'estimer que l'ontologie est avant tout de la connaissance propre à une collection d'individus partageant les mêmes intérêts, de la connaissance propre à une communauté d'usage. *De facto* cette connaissance modélisée est fonction des individus qui la créent, de leur culture, de leur éducation, mais aussi de la perception propre de leur univers. Une grande partie des travaux actuels repose encore sur

26. Utilisation de classes pour concepts, de slots pour propriétés et de facets pour valeurs des propriétés et contraintes.

TABLE 2.4 – Caractéristiques de TopBraid.

OS	Windows, Mac OS, Linux
Langages supportés	OWL, RDF
Multilinguisme	Oui, utilisation de SKOS
Vérification d'ontologie	Oui, test de consistance
Visualisation	Arborescente dans la version free
Extraction d'information automatique	Bases de données, XML, Excel
Développement collaboratif d'ontologies	Oui pour la version commerciale
Interopérabilité (web services...)	Web services en version commerciale
Outil modifiable	Oui, environnement Eclipse
Performance	OK pour ontologies volumineuses
Sécurité	Oui, par SVN.
Flexibilité	Oui, en changeant de version (Suite, Live)
Facilité d'utilisation	De nombreuses aides

le modèle aristotélicien des conceptualisations humaines. Cependant, des problématiques récentes incitent à faire évoluer ce modèle ou tout au moins à l'enrichir.

Chapitre 3

Catégorisation, catégories et prototype

3.1 Introduction

Définitions [Catégorisation] :
- Conduite adaptative fondamentale par laquelle nous découpons le réel physique et social. Sa fonction cognitive est la création de catégories (d'objets, d'individus, etc.) nécessaires à la transition du continu au discret. [Dubois 2001]
- Constitution de classes d'équivalences, en étant capable d'extraire des invariants tout en négligeant des caractéristiques non pertinentes. [Rossi 2006]

Le concept de catégorisation[1] se situe aujourd'hui à la croisée des chemins de la philosophie, de la logique, de la psychologie cognitive, de la linguistique et de l'ingénierie des connaissances. [Barsalou 1991] distingue *Conceptualisation* et *Catégorisation* : si la première conçoit les concepts comme des entités isolées, la seconde a pour objectif de les intégrer dans un système cohérent de croyances au moyen de leur traits et des relations qui les relient entre eux, ou des relations qui relient les instances d'un même concept entre elles.

Le processus de catégorisation est une action naturelle chez l'Homme au travers de la fonction de classement et la conscience d'appartenance à un groupe. Ce processus mental est présent dans les perceptions, réflexions, paroles, ... dès lors qu'une chose est prise comme étant *une espèce de chose*, dès lors qu'une organisation - une structuration - d'un environnement est mis en place. Cette structuration est une représentation interne, une modélisation, d'une réalité externe et perçue par nos sens (notion de catégorie perceptive). Les représentations conceptuelles sont des intermédiaires entre le monde perçu et le monde représenté mettant en œuvre tant les perceptions que la cognition [Chemlal 2006].

1. Nous utiliserons dans ce chapitre uniquement le terme *catégorie* en lieu et place de *concept*, ce terme étant le plus courant dans les travaux que nous présentons ici et possédant une définition moins rigide que celle de concept.

La catégorisation est l'objet de nombreux travaux depuis le début des années 1960. [Howard 1963] fut l'un des premiers à l'intégrer dans le domaine du marketing et à prendre en compte les états affectifs au sein de son analyse. Il part ainsi de l'hypothèse que les consommateurs ne peuvent traiter sur un même pied d'égalité toutes les marques qu'ils rencontrent sur le marché lorsqu'ils sont confrontés à une décision d'achat. Lorsqu'un consommateur envisage un achat, la quantité d'options qui lui viennent à l'esprit serait en effet moindre que le nombre qui est objectivement disponible. Et Howard de définir ainsi avant bien d'autres, par ce qu'il qualifie de *marque saillante*, la notion même de *prototype*. La notion de *catégorie sémantique naturelle* apparaîtra dans les travaux de psychologie cognitive d'E. Rosch au début des années 70, avec une théorie qui évolue au fil du temps, pour aboutir au final à l'émergence des concepts de prototype et de typicalité.

3.2 Principes généraux

Il existe dans la littérature maintes définitions de la notion de catégorie. Selon [Chemlal 2006], les catégories sont des unités élémentaires du traitement cognitif, véritables intermédiaires entre le monde perçu et le monde représenté ayant un pied dans la perception et un autre dans la cognition. Elles sont définies, d'après [Piaget 1972], sur la base d'une relation d'appartenance permettant de dire si oui ou non un élément appartient à une catégorie. Selon [Dubois 1991], il existe plusieurs types de catégories, des plus précises au plus abstraites (les plus englobantes et les plus génériques), les différents niveaux de catégorie étant hiérarchisés par une structure interne. [Barsalou 1985] définit deux types de catégories : (1) les catégories dites *naturelles*, fondées sur la ressemblance physique de leurs instances avec des structures préexistantes en mémoire, et (2) les catégories dites *à buts* dont les structures sont composées et modifiées en fonction des buts poursuivis par l'individu (par exemple, le but *gagner de l'argent* peut nous amener à la création d'une catégorie de *placement rentable*).

La catégorisation est un concept multidimensionnel qui peut être vu suivant deux dimensions :

- *verticale*, *i.e.* différents niveaux hiérarchiques où l'ascendance est vue comme une inclusion des ensembles (par exemple *chartreux* → *chat* → *félins* → *mammifère*) ;
- *horizontale*, *i.e.* une segmentation catégorielle.

[Ladwein 1995] assimile ce processus à un stockage d'informations structurées de manière mémorisable et opérante. Pour [Rosch 1981], il s'agit d'une activité cognitive consistant à regrouper des objets ou des événements non identiques dans des ensembles, une catégorie cognitive étant un groupe d'objets considérés comme équivalents par un individu. [Chemlal 2006] assimile la catégorisation à un processus fondamental de prise d'information sur le monde. Selon [Cohen 1987], un individu adopterait trois modèles de catégorisation :

- un *modèle classique* où chaque catégorie est dotée d'un ensemble de propriétés nécessaires et suffisantes (tout élément possédant ces propriétés appartient à la catégorie, et tous les membres sont équivalents - un nouvel élément est affecté à la catégorie avec laquelle il a le plus grand nombre de propriétés en commun) ;
- un *modèle prototypique* où la catégorie est organisée autour d'un élément central (fictif) - le prototype - résumé de la catégorie qui représente la tendance centrale (un nouvel élément est affecté à la catégorie par comparaison entre ce nouvel élément et le prototype de la catégorie) ;
- un *modèle de l'exemplaire* où la catégorie est organisée autour d'un de ses membres (réel) - l'exemplaire - définissant le mieux, aux yeux de l'individu, la catégorie (un nouvel élément est affecté à la catégorie par sa similitude à l'exemplaire).

Catégories et catégorisation, en tant qu'état et processus, ont donc un double rôle : faciliter la classification de nouveaux éléments en simplifiant l'environnement de l'individu, et améliorer l'efficience du traitement de l'information lors de prises de décisions, de jugements d'évaluation, de choix et de production de nouvelles connaissances [Brucks 1985, Sujan 1985, Nedungadi 1985].

3.2.1 Catégorisation naturelle et perceptive

Un environnement donné n'est perçu que par l'intermédiaire d'un flot continu et dynamique de stimuli. Un certain nombre de représentations susceptibles d'être traitées/catégorisées sont ensuite élaborées, stockées en mémoire, puis transmises au travers d'actes de discours. La question fondamentale qui se cache derrière le processus de catégorisation est de savoir sur quels critères il est décidé de l'appartenance ou de la non-appartenance d'un objet à une catégorie.

Selon [Bruner 1957], les perceptions s'inscrivent dans un *processus de catégorisation dans lequel en suivant une logique d'inférence, les individus utilisent des signaux reçus pour construire une identité catégorielle.* Ainsi, par les perceptions, *les individus apprennent les relations entre les propriétés des objets et des événements rencontrés ; ils apprennent à prédire et à vérifier quoi va avec quoi.* "Je perçois, donc je catégorise." J. S. Bruner modélise le processus de catégorisation en une séquence comportant quatre phase successives :

1. **catégorisation primitive** - simple perception des choses, sans aucune signification associée ;
2. **recherche des indices** - recherche de données servant d'indices à la catégorisation des choses perçues ;
3. **épreuve de confirmation** - recherche des éléments supplémentaires pour valider la catégorisation de ces choses ;
4. **conclusion de la confirmation** - de moins en moins ouvert aux perceptions

pouvant infirmer le choix.

La catégorisation perceptive serait ainsi un processus de simplification de la complexité de l'environnement [Berger 2000], une manière de stocker les connaissances en mémoire par des individus confrontés à des stimuli. Ce processus est illustré par l'expérience réalisée par [Cauzinille-Marmèche 1998] (*cf.* figure 3.1). Soit un ensemble d'objets construits sur la base de trois dimensions binaires : la forme (triangle, carré), la taille (grand, petit), la couleur (noir, blanc). Les objets (1), (2) et (3) sont présentés en précisant que tous trois appartiennent à une même catégorie ; la question étant laquelle (et pourquoi). L'objet (1) offre trois possibilités : triangle, grand et blanc. L'objet (2) réduit le nombre d'hypothèses aux deux premières. Enfin l'objet (3) précise la catégorie à celle de *triangle*. La catégorisation s'est fait par une recherche d'attributs perçus communs à tous les objets présentés.

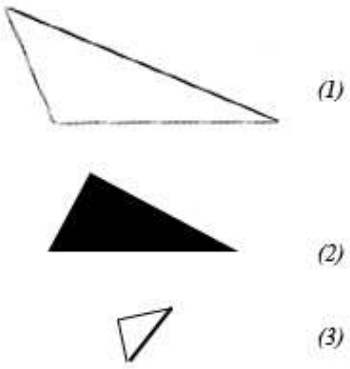


FIGURE 3.1 – Expérience rapportée par E. Cauzinille-Marmèche, D. Dubois et J. Mathieu [Cauzinille-Marmèche 1998]

Dès lors, dans un processus de catégorisation et avec cette approche, comment s'effectue le choix d'une catégorie ? Selon [Bruner 1990], le choix d'une catégorie dépend de deux facteurs : l'appariement et l'accessibilité. Ce choix est fonction du degré de recouvrement entre les indices perçus et les propriétés définitoires du concept (la définition se trouve dans le système cognitif), mais pas uniquement. Il peut en effet dépendre de plusieurs facteurs :

- l'apprentissage d'instances communes entre catégories ;
- les besoins, les motivations, les émotions de l'individu (par exemple, les personnes dépressives ont un accès plus facile aux catégories à connotation négative) ;
- la récence d'utilisation d'une catégorie ;
- le jugement perceptif (par exemple, la présentation des caractère I3 peut-être perçue comme le nombre 13, la lettre B ou des caractères I et 3) ;
- le contexte émotionnel de perception de la catégorie ([Roth 1983]).

3.2.2 Catégorisation suivant l'approche par Conditions Nécessaires et Suffisantes (CNS)

Cette conception aristotélicienne de la catégorisation, dans la perspective des travaux de [Collins 1969], repose sur l'idée que chaque catégorie est définie par un nombre fini de propriétés caractéristiques. Tout élément de l'univers possédant *au moins* l'ensemble de ces propriétés appartient à la dite catégorie. Cette approche, dite *objectiviste*, se fonde sur l'appartenance booléenne d'une entité à une catégorie en utilisant les propriétés comme CNS. Selon [Kleiber 2004], ce modèle s'appuie sur les quatre propositions suivantes :

1. les catégories possèdent des frontières délimitées ;
2. tous les membres d'une même catégorie ont un statut catégoriel identique ;
3. l'intension détermine l'ensemble des traits et conditionne l'extension ;
4. l'appartenance d'un élément de l'univers à une catégorie donnée est de type booléen.

Par exemple, un oiseau est défini comme un animal vertébré ovipare, couvert de plumes, muni d'un bec sans dents, de deux pattes et de deux ailes. Sur un mode définitionnel, une catégorie *oiseau* est déterminé par les traits [*ponds des oeufs*], [*plumes*], [*bec sans dents*], [*deux pattes*] et [*deux ailes*]. Tout animal possédant au moins l'ensemble de ces traits est classé dans la catégorie des *oiseaux*. *De facto*, dans ce modèle, dire qu'un canari est plus un oiseau qu'une autruche ne paraît pas envisageable. Les considérations du style "*est plus que*", "*est moins que*" ou "*plus ou moins que*" au sein d'une même catégorie n'ont pas lieu d'être dans cette catégorisation. L'ensemble des attributs pour une catégorie doit être *suffisant* pour déterminer sa condition d'appartenance à une catégorie. Et, selon [Kleiber 2004], chacune de ces conditions, dont l'importance équivalente et indépendante, doit être *nécessaire*, en sachant qu'aucune n'est suffisante [2]. Chaque catégorie possède donc :

- un ensemble de propriétés correspondant aux CNS (sens du terme lexical qui désigne cette catégorie - du ressort de la sémantique) ;
- et un ensemble de propriétés contingentes [3] (connaissances encyclopédiques sur la catégorie - du ressort de la pragmatique).

La définition zoologique d'une vache indique qu'il s'agit d'un mammifère domestiqué ruminant appartenant à l'espèce Bos taurus, de la famille des bovidés, à poils courts, élevé pour la production de lait ou de viande. Un élément, pour être catégorisé comme *vache*, doit posséder un nombre fini de propriétés biologiques propres à cette espèce. Mais il en existe qui, bien que distribuées à l'ensemble des vaches, ne constitue pas pour autant une condition nécessaire et suffisante. Les vaches de race scandinave ne possèdent pas de cornes. La propriété *possède des cornes* fait partie de nos connaissances encyclopédiques mais ne ressort pas au sens lexical du terme vache.

2. C'est l'ensemble de toutes les conditions qui est suffisant.
3. *i.e.* pouvant être possédées ou non par les instances de la catégorie.

[Kleiber 2004] dénombre plusieurs limites au modèle des CNS :

- la multiplicité des définitions possibles [4] [Kleiber 1987] ;
- les CNS ne peuvent s'appliquer à tous les domaines de connaissance ;
- difficultés avec les cas marginaux (cas des pingouins et des autruches pour la catégorie *oiseau*, ce sont bien des oiseaux mais ils ne possèdent pas toutes les CNS de cette catégorie) [Geerearts 1985, Geeraerts 1986, Lyon 1969].

Le modèle des CNS n'aborde pas la notion d'héritage - simple ou multiple - et ne considère pas non plus la catégorisation sous un angle plus général au sein d'une arborescence ou d'un treillis. Comment, dès lors, conjuguer CNS et hiérarchie ? Il serait possible d'envisager les CNS comme étant des propriétés héritées de la catégorie mère, et les propriétés contingentes comme étant des attributs propres à chaque héritier de cette catégorie. Le principe de la différentialité [5] n'est apparemment pas incompatible avec cette approche ; il faudrait créer des classes intermédiaires de manière à pouvoir traiter les cas marginaux. De plus, une vision ontologique élimine certains problèmes comme celui lié à la polysémie des noms de catégories employés (résolu par le principe d'unicité des termes saillants de catégories). Cette vision ontologique favoriserait de surcroît l'existence d'une hiérarchie interne au sein d'un même niveau catégoriel, éliminant de ce fait la contrainte liée aux CNS pures. Reste alors le problème de la plasticité de la connaissance, plasticité qui est l'une des caractéristiques de la connaissance chez l'être humain. Il est entendu par plasticité, la capacité des représentations à changer de forme ou de fonction selon les altérations du contexte. Par exemple, si dans le cas de la grande majorité des *Oiseaux*, le mode de locomotion est le vol, dans le cas du pingouin (qui est un oiseau) il s'agit de la nage. Un modèle fondé sur les CNS ne semble pouvoir concilier ces deux aspects. À moins de dissocier cette propriété de l'oiseau, et de créer une catégorie *Oiseau volant*, une catégorie *Oiseau nageur* et une catégorie *Oiseau volant et nageur*.

3.3 Théories du prototype

[Barsalou 1983, Rosch 1975c] formulent certaines hypothèses quant aux catégories :

- certaines catégories sont de meilleurs représentants de leur catégorie mère que d'autres ;
- l'appartenance de certaines catégories à une catégorie mère est incertaine ;
- les non-membres d'une catégorie varient dans leur similarité au prototype de cette catégorie ;
- toutes les catégories d'une catégorie mère ne sont pas toutes équivalentes.

4. Ce problème peut être corrigé avec la vision consensuelle d'une ontologie de domaine pour un endogroupe donné.

5. A savoir que chaque catégorie hérite de l'ensemble des attributs de la catégorie mère, auxquels il rajoute ses propres propriétés, différentes de celles de l'ensemble de ses frères.

3.3.1 De l'importance des propriétés

D'un point de vue expérimental, [Rips 1973] compare les temps de catégorisation d'une série de signifiés appartenant à une hiérarchie catégorielle du domaine de la zoologie. Il résulte de ses expériences que les *Poules*, *Rouge-gorges* et *Perroquets* sont jugés plus facilement comme appartenant à la catégorie des *Oiseaux* qu'à la catégorie des *Animaux*. A l'inverse, les *Ours*, *Chameaux* et *Souris* sont plus facilement assimilés aux *Animaux* qu'aux *Mammifères*. Au lieu de supposer que toutes les catégories sont équivalentes, et que seule leur nombre d'occurrences dans notre univers intervient, il est possible d'envisager que le poids de certaines de leurs propriétés est variable, et que, par conséquent, certains d'entre eux conduisent à une décision plus rapidement que d'autres. Par exemple, la propriété *allaiter(x,petits)* aurait un faible poids pour la catégorie *Souris*, classé plus facilement comme *Animal* que comme *Mammifère*.

Une deuxième expérience a été réalisée et consiste à caractériser des objets suivant trois propriétés : (A) la forme, (B) la couleur et (C) la taille. L'objectif de cette expérience est de demander aux cobayes de déduire une règle de catégorisation valable pour tous les objets et concernant une seule propriété. Dans la plupart des cas, les sujets réussissent plus rapidement à résoudre le problème pour certaines propriétés que pour d'autres. Il s'agit de propriétés *saillantes*, de propriétés donnant un *relief* plus important que les autres pour la catégorie. Le choix de cet attribut dépend de l'âge de l'individu (chez les enfants, la couleur passe avant la forme), de la personnalité, du contexte dans lequel les objets sont présentés, etc. La propriété qui donne le relief à une catégorie est celle qui est choisie le plus souvent pour catégoriser ses instances, et qui va la spécifier (et donc la différencier) le plus par rapport aux autres.

Il est dès lors possible de définir un relief structural, caractère différentiel de la catégorie. Cette notion caractérise le fait qu'une propriété est plus ou moins facilement prise en considération dans un trait sémantique quelconque de la catégorie qui la possède. Chaque catégorie possède ainsi deux structures de propriétés : une structure permanente (les propriétés possèdent la même valeur d'importance) et une structure circonstanciée (des propriétés pondérées).

Un autre aspect est à prendre en compte dans la définition du relief structural : la dimension affective de la signification. [Osgood 1953] a mis en avant dans ses travaux l'importance de la dimension *agréable / désagréable*. Pour beaucoup d'individus, l'affectivité possède un relief élevé au sein de leur catégorisation. Ainsi, plus cette dimension est élevée, plus le relief structural l'est également et plus l'indiscrimination est forte.

3.3.2 Typicalité et prototype, approche dite standard

Selon [Kleiber 1987], l'approche CNS semble souffrir de plusieurs insuffisances. Tout d'abord, il est assez difficile de définir une catégorie en termes de conditions nécessaires et suffisantes. Mais, si tel était le cas, alors tous les membres d'une catégorie auraient un statut équivalent. Or, intuitivement, il est possible de dire que certains membres d'une catégorie sont de meilleurs exemplaires de cette catégorie que d'autres (pour les européens, le moineau est un meilleur exemplaire d'oiseau qu'un émeu ou un pingouin). La théorie classique est synonyme de déterminisme quant au rattachement (ou non) d'un objet à une catégorie, alors qu'il existe des cas où l'appartenance est plus questionnable. Enfin, cette théorie poserait le problème de la polysémie des mots, avec une vision trop restrictive du sens.

Afin de pallier aux problèmes liés à une approche CNS, il s'est développé un deuxième courant, qualifié d'*expérientialiste* [Lakoff 1986] [6]. La catégorisation des éléments de l'univers s'établit ici sur la base de formation de prototypes de référence puis de recherche de similitudes globales avec ces derniers. Si la première approche est plutôt analytique car fondée sur l'analyse directe des propriétés, cette deuxième approche peut être qualifiée de comparative car fondée sur l'expérience et la fréquence d'apparition d'éléments dans un environnement. Cette catégorisation des éléments composant un environnement donné est fortement dépendante de sa perception par un individu.

La théorie standard du prototype repose sur les quatre propositions suivantes :

1. les catégories possèdent des frontières floues ;
2. les membres d'une même catégorie ne partagent pas tous les mêmes propriétés [7] ;
3. toute catégorie possède une structure interne (un ensemble de propriétés) prototypique ;
4. l'appartenance d'un élément de l'univers à une catégorie donnée est fonction du degré de similarité entre l'élément en question et le prototype de la catégorie [8].

L'axe central consiste donc à bâtir un *prototype* pour chaque catégorie.

6. Selon G. Lakoff, dans la perspective d'une sémantique cognitive, l'ensemble des métaphores qui structurent nos expériences quotidiennes reposent sur des schèmes sensori-moteurs. C'est parce que nous éprouvons dans notre chair et notre corps propre des variations de tonicité musculaire, des mouvements viscéraux et des changements de posture associés par exemple aux changements d'humeur, que nous pouvons bâtir, comprendre et déployer de telles métaphores en toute cohérence.

7. Ils sont liés par une similitude, et doivent posséder au moins une propriété commune avec le prototype.

8. La décision d'appartenance est ici issue d'un processus beaucoup plus global, et non plus d'un processus purement analytique fondé sur chacune des propriétés.

3.3.2.1 Prototype

On entend par prototype le "meilleur" exemplaire (le plus représentatif) pour chacune des catégories selon un consensus social de l'endogroupe concerné. Selon E. Rosch, ce meilleur exemplaire est une sous-catégorie et non une instance particulière [Rosch 1973]. Si tel était le cas, cela signifierait que tous les individus composant cet endogroupe ont la même représentation mentale d'une catégorie donnée. Pour [Kleiber 2004], les notions de *prototype* et de *stéréotype* sont à différencier : le premier est une vision extensionnelle (objet mental selon [Reed 1972], schéma, image cognitive, etc.) et le second une vision intensionnelle. Par exemple, un *moineau* pourra être jugé comme étant le prototype de la catégorie *oiseau*, alors que le stéréotype de cette même catégorie sera une définition abstraite regroupant toutes les propriétés considérées comme typiques de la catégorie (mais non comme CNS) - et ce pour l'ensemble des locuteurs de l'endogroupe. En ce cas, comment choisir ces traits typiques pour une catégorie ? Selon E. Rosch, deux critères peuvent être utilisés : la fréquence d'usage (familiarité) et la *cue validity* (validité des traits). [Kleiber 2004] décrit cette dernière comme *le degré de prédictibilité pour une catégorie d'une propriété ou d'un attribut d'un objet (cue) (...) correspondant à la fréquence de l'attribut associé à la catégorie en question divisée par la fréquence totale de cet attribut pour toutes les catégories pertinentes. Un attribut présente donc une cue validity élevée pour une catégorie si un grand nombre de membres de la catégorie le possèdent et si, en revanche, peu de membres de catégories opposées le vérifient.* Il en résulte que les sous-catégories les plus prototypiques sont celles qui (1) partagent le plus de propriétés avec les autres sous-catégories, et qui (2) ont le moins de propriétés en commun avec les sous-catégories opposées. Il en résulte également que le prototype ne possède pas forcément d'extension dans l'univers, dans le sens où il peut être une construction mentale *idéale* rassemblant tous les traits typiques pour une catégorie donnée ; le prototype serait un modèle cognitif idéalisé.

Selon [Prat 2006], les jeunes enfants peuvent extraire des prototypes, cela signifie que la construction de prototypes ne nécessite rien d'autre que l'exposition à un ensemble d'exemplaires. Le futur adulte en gardera sûrement la trace et peut être un comportement ultérieur. *A contrario*, selon [Lemaire 1999], il faut attendre un certain âge, âge permettant la mise en œuvre de certaines stratégies et l'acquisition d'une quantité suffisante de connaissances. Il suggère que la construction d'un prototype nécessite la mise en œuvre de stratégies particulières et, surtout, un minimum de connaissances générales ou spécifiques. Cette seconde hypothèse amène à l'idée d'une construction de prototypes en fonction d'un contexte tant cognitif qu'émotionnel. Intuitivement, les connaissances sont construites en se fondant sur une classification, sur une catégorisation de représentations mentales autour de stéréotypes, de prototypes. Ceux-ci peuvent prendre une forme abstraite comme concrète, ils possèdent l'ensemble des attributs que doivent ou devraient avoir les éléments d'une catégorie. D'un point de vue commun, un *prototype* se définirait comme étant le meilleur exemplaire d'une catégorie. [Rosch 1973] le définit comme

étant un *stimulus, qui prend une position saillante dans la formation d'une catégorie parce qu'il est le premier stimulus associé à cette catégorie.* Mais cette définition évolue au fil des ans, tout comme la théorie du même nom, pour devenir *membre le plus central d'une catégorie, fonctionnant comme un point de référence cognitif.* Le processus de catégorisation s'établit alors en deux phases :

- recherche du meilleur exemplaire de chaque catégorie ;
- *classement* graduel autour du prototype par recherche de similitude (partage d'une ou plusieurs caractéristiques) avec ce dernier.

Plus on descend dans les niveaux de catégorisation, plus le nombre de propriétés augmente. Ainsi, les niveaux dits *supérieurs* sont caractérisés par peu de propriétés (essentiellement la fonction), les niveaux *intermédiaires* sont descriptifs, enfin les niveaux *subordonnés* sont surtout des niveaux d'expertise. [Rosch 1975a, Rosch 1978] associent ainsi intimement la notion de sémantique à celle de catégorisation. Selon [WoodField 2004], E. Rosch et ses collègues *ont démontré expérimentalement que la corrélation entre les traits distinctifs est élevée dans le cadre des catégories fondamentales. Ces catégories ont en effet une grande cohésion interne et sont mutuellement exclusives. Beaucoup de raisons permettent de penser que ce niveau des catégories fondamentales a une réalité psychologique. Rosch indique que les catégories fondamentales sont les catégories les plus abstraites dont les membres ont une forme semblable, dont on peut former une image typique, et qui peuvent être traitées par les mêmes routines psychomotrices.*

3.3.2.2 Typicalité

Nous trouvons dans la littérature aussi bien les termes *typicalité* et *prototypicalité*. Selon [Cordier 1993], ces deux termes sont souvent employés comme synonymes, en sachant que le prototype serait l'exemple catégoriel qui représenterait à lui seul le meilleur résumé de la catégorie à laquelle il appartient au sein d'un classement effectué par degré de typicalité. Il est ainsi entendu par *typicalité*, l'aspect caractéristique et représentatif d'une catégorie vis-à-vis d'une catégorie. La typicalité d'une catégorie pour une sur-catégorie donnée peut être évaluée par un individu de plusieurs manières :

- par degré de satisfaction d'un idéal associé à une catégorie, selon [Barsalou 1983] ;
- par ressemblance à une famille de catégories, selon [Rosch 1975c] ;
- par familiarité avec des instances de la catégorie et fréquence d'exposition ;
- par expériences / expertises liées à la catégorie, impliquant au final une réaction de préférence.

[Rips 1973] ont trouvé, lors d'expériences de mesure de temps de catégorisation du type *un x est un y* - où x est un sous-catégorie de la catégorie y, que toutes les sous-catégories ne sont pas équivalentes dans leur rattachement à une catégorie. De ce fait, certaines catégories étant de meilleurs représentants de leur catégorie que d'autres, [Rosch 1975c] estiment qu'il est possible d'organiser les connaissances suivant des

gradients de typicalité - tout élément variant en typicalité dans toute catégorie. Selon [Nedungadi 1985], le contexte (importance des propriétés, évaluation, préférence et usage) rentre également en compte dans l'évaluation de la typicalité d'une catégorie. Les phénomènes de typicalité se manifestent de plusieurs façons :

- par le temps de catégorisation (plus il est court, plus la sous-catégorie est typique) ;
- par le nombre d'erreurs de catégorisation (moins elles sont nombreuses, plus la sous-catégorie est typique) ;
- par l'ordre d'apprentissage (les noms des sous-catégories les plus typiques sont retenus en premier) ;
- par l'ordre de production (quand il est demandé à des individus de citer des sous-catégories d'une catégorie, les exemplaires les plus typiques sont cités en premier) ;
- par points de référence cognitive (les sous-catégories les plus typiques sont plus facilement choisies comme points de référence cognitive).

S'il est demandé à un individu (citoyen européen, qui plus est français) de citer un type de chien (*i.e.* en fait un sous-catégories de la catégorie *Chien*), il est très probable qu'il nous donne la race de **son** chien et qu'il nous cite le labrador [9].

3.3.2.3 Approche standard

Selon [Cordier 1993], le processus de décision d'appartenance d'un élément à une catégorie était initialement prise selon des critères de propriétés nécessaires et suffisantes. Or, toujours selon cet auteur, les choix d'un individu sont fonctions au moins de deux éléments : le degré de typicalité [10] d'une part, le niveau d'abstraction d'autre part. Ces points se mettent en place progressivement au fur et à mesure de l'apprentissage de l'individu, dès l'âge de six mois suivant certains spécialistes.

Il existe au moins trois approches pour établir un tel classement et définir la valeur des différents gradients.

La première se fonde sur le *jugement moyen* des individus appartenant à un endogroupe donné. Selon [Cordier 1993], il s'agit alors d'intégrer dans le processus d'engagement ontologique la participation d'experts fiables, représentatifs de l'endogroupe, pour la pondération des liens « *is-a* » de la hiérarchie catégorielle (en essayant de tenir compte autant que possible des biais éventuels), et ce à l'aide d'une échelle d'évaluation graduée, par exemple, de 1 à 7. La valeur 1 signifiant que l'exemplaire est un très bon élément de la catégorie, 4 une valeur moyenne et 7 que cet élément correspond très peu à l'idée que nous pouvons nous faire de cette catégorie. La moyenne des valeurs obtenues pour chaque lien lui est alors affectée. Selon

9. D'après les résultats du site http ://fr.woopets.com/races-de-chiens
10. Un élément est jugé typique s'il représente un bon exemple de la catégorie, selon des critères propres à la culture, à l'éducation et aux émotions de l'individu.

[Cordier 1993], cette méthode offre l'intérêt de donner des résultats relativement stables dès lors qu'ils sont obtenus à partir d'un nombre conséquent d'individus interrogés. De nombreux travaux ont étudié ou utilisé cette méthode, citons parmi eux [Rosch 1975a, Lecocq 1987, Boster 1988].

La deuxième vise à établir une norme à partir des fréquences de citations dans les supports appartenant au domaine et servant à l'élaboration de l'ontologie (plus un terme est fréquent, plus sa catégorie afférente est représentative.) D'un point de vue pratique, cette méthode se fonde, comme la précédente, sur la consultation d'un nombre conséquent d'individus appartenant à un endogroupe donné et experts du domaine traité. Il est demandé à cette population de produire pour chaque item une liste ordonnée de sous-catégories parmi les plus représentatives. Selon [Arucri 1986, McEvoy 1982], le degré de typicalité de chacun de ces items est alors défini à partir de la fréquence de citation. Cette méthode a l'énorme inconvénient d'être parasitée par la forte présence d'éléments de nature extensionnelle.

La troisième méthode, selon [Cordier 1985], consiste à fournir l'ensemble des items et à demander (toujours aux individus appartenant à un endogroupe donné et experts du domaine traité) d'effectuer un classement, lequel est effectué intuitivement par regroupement de propriétés.

En résumé, pour un individu appartenant à un endogroupe donné, le jugement d'appartenance et le degré de typicalité d'une sous-catégorie à une catégorie, sont dûs à la fréquence des mots dans sa langue et à la quantité de caractéristiques ou d'attributs que les sous-catégories possèdent en commun avec la catégorie d'appartenance.

[Niedenthal 2004, Zammuner 1998] ont mené différentes expériences afin d'établir les principaux facteurs de prototypicalité. Ils arrivent à la conclusion que, outre la dimension lexicale et sémantique, la composante émotionnelle est capitale dans une telle catégorisation. Selon [Blanc 2006], cette composante émotionnelle peut être caractérisée suivant deux dimensions : l'intensité et la valence. Mais cette notion n'est pas nouvelle. Descartes avait déjà évoqué, dans ses *Passions de l'âme*, le fait que l'émotion nous indique quelles sont les informations qui doivent être considérées avec attention.

3.3.3 Approche dite étendue

Cette théorie est inspirée par l'approche standard (*cf.* section 3.3.2.3) et par les travaux de Wittgenstein où il est postulé que (1) les référents d'un mot n'ont pas besoin d'avoir d'éléments en commun pour être compris et employés dans le fonctionnement normal du langage, et (2) il s'agit plus d'une ressemblance de famille qui relie les différents référents d'un mot [Wittgenstein 1973]. L'approche étendue propose une structure de ressemblance de famille de la forme AB, BC, CD, DE, et

non de type AB, AC, AD, AE comme l'approche standard. L'idée développée ici, selon [Rosch 1975c] est qu'il existe peu d'attributs communs à tous les instances d'une même catégorie, mais que chaque instance a au moins un, et probablement plusieurs attributs en commun avec une ou plusieurs autres instances. La typicalité devient ici un degré avec lequel une instance d'une catégorie possède des attributs semblables à ceux des autres instances de la dite catégorie. Ainsi, seulement deux hypothèses de la théorie standard se retrouvent dans la théorie étendue : (1) il y a des effets prototypiques d'usage et (2) les différentes instances d'une catégorie sont regroupées par air de famille.

La théorie étendue du prototype repose sur les quatre propositions suivantes :

- il y a à la fois une pluralité des références et une unité intuitive de la signification. Par exemple, le terme *chien* dénotant la catégorie *Chien* désigne les caniches, les labradors, les terre-neuves... Il est donc polysémique, mais nous le percevons comme monosémique.
- il y a ressemblance de famille, ce qui se traduit par des recouvrements de sens ou d'attributs.
- tous les membres ne sont pas équivalents, *i.e.* ils possèdent des degrés de représentativité différents.
- la catégorie a des frontières mal explicitées.

Comme nous pouvons le voir, cette "théorie" s'applique plus aux termes qu'aux catégories et s'apparente d'avantage à l'expression d'une polysémie de type métonymique ou métaphorique. En passant de l'approche standard à l'approche étendue, nous passons d'une approche monosémique (tous les individus qu'un terme rassemble sont semblables) à une approche polysémique (un terme désigne des individus différents). La signification serait donc une association (terme, domaine/contexte).

G. Kleiber pense néanmoins que cette notion ne peut suffire à bâtir une catégorie, qu'il s'agit surtout d'une catégorisation linguistique polysémique (phénomène linguistique partiel) et non une catégorisation naturelle de sens [Kleiber 2004]. Ainsi, l'appartenance à une catégorie n'obéit plus aux mêmes critères suivant que le terme correspondant est monosémique ou polysémique. La notion même de prototype comme meilleur exemplaire - dont l'image mentale constituerait la signification du terme correspondant - est abandonnée. Le jugement des individus n'aurait plus d'intérêt puisqu'il n'y a plus de meilleur exemplaire sur quoi faire porter ce jugement.

3.4 Conclusion

À l'heure du Web Sémantique et du Web des données avec la mise à disposition de milliards de triplets RDF modélisant de la connaissance sur différents sujets, la question de la création d'une catégorisation universelle et ses conséquences est loin d'être illégitime. Cependant, en 2005, bien avant cet avènement, C. Shirky, dans

un article intitulé *Ontology is Overrated : Categories, Links, and Tags*[11], montrait à quel point une catégorisation universelle - et son approche - pouvait poser de nombreux problèmes.

Afin d'illustrer son propos, Shirky se fonde sur la classification DMOZ[12], l'Open Directory Project est le plus grand et le plus complet des répertoires du Web édités par des êtres humains - selon leurs auteurs. Shirky pose ici le problème soulevé par le choix d'utiliser des experts et non l'ensemble des individus d'une communauté pour concevoir une telle catégorisation. Il résume cette pensée sous la forme de propos attribués aux dits experts : *nous comprenons mieux que vous comment le monde est organisé, parce que nous sommes des professionnels formés.* En confiant cette tâche à une minorité d'individus désignés et par forcément représentatifs, la modélisation peut alors être guidée par des intérêts, par exemple commerciaux, destinés à déterminer la vision que l'utilisateur devrait adopter pour utiliser le système.

Le deuxième problème soulevé par Shirky est que toute catégorisation du plus grand nombre doit satisfaire tout un chacun. Pour lui, là encore, le choix de l'utilité de la catégorisation de concepts particuliers par les catalogueurs d'une part, et le placement de ceux-ci dans la modélisation d'autre part, n'est pas sans conséquence et souffre d'absence de neutralité. *Les visions des catalogueurs surpassent nécessairement les besoins de l'utilisateur et la vision du monde de l'utilisateur. Si vous voulez quelque chose qui n'a pas été catégorisé dans le sens où vous y pensez, pas de chance pour vous.* Ainsi un utilisateur lambda établit des correspondances - et donc des transformations - entre ce qu'il cherche et la façon dont l'univers est modélisé au sein de cette catégorisation du plus grand nombre.

11. http ://www.shirky.com/writings/ontology_overrated.html
12. http ://www.dmoz.org/World/Fran%C3%A7ais/

Chapitre 4

Mesures sémantiques

4.1 Introduction

Avec l'avènement du Web des données lié à l'essor considérable des capacités de stockage et de traitement des données, la question de la comparaison d'entités a connu un nouveau regain d'intérêt. Les mesures sémantiques sur lesquelles s'appuient bien souvent les comparaisons doivent s'adapter tant aux nouveaux usages, comme la recherche d'information par exemple, mais également pouvoir supporter un passage à l'échelle sur des volumes de données de plusieurs centaines de milliers de concepts. La plupart des mesures sémantiques de la littérature peuvent être regroupées en quatre catégories :

1. les mesures de type structurel ;
2. les mesures de type intensionnel (fondées sur les propriétés des concepts) ;
3. les mesures de type extensionnel (fondées sur les instances des concepts) ;
4. les mesures de type expressionnel (fondées sur les termes dénotant les concepts).

Pour les définir dans la suite de ce chapitre, nous utilisons les notations suivantes :

- $prof(c_i)$, la profondeur du concept c_i dans la hiérarchie de concepts (telle qu'elle a été définie dans la section 2.6.8) ;
- c_{com}, le plus petit concept père commun aux concepts c_1 et c_2 ;
- max, la profondeur maximale de la hiérarchie ;
- $dist_{edge}(c_1, c_2)$, la longueur du plus court chemin entre deux concepts c_1 et c_2 ;
- $|I_c|$, le nombre d'instances du concept c ;
- $|I|$, le nombre total d'instances de l'ontologie.

4.2 Mesures de type structurel

A. Collins et M. Quillian, un psychologue et un informaticien, ont élaboré dans [Collins 1969] les premiers réseaux sémantiques sur la base de temps de réponse à des questions comme "*Un canari est-il un oiseau ?*" ou encore "*Un canari est-il un animal ?*". Ils ont supposé que plus le temps de réponse à ces questions était long,

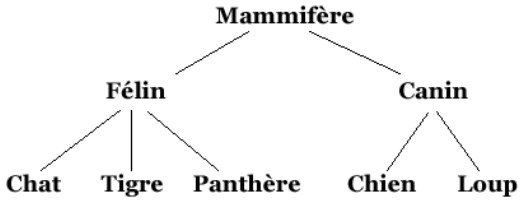


FIGURE 4.1 – Extrait d'une hiérarchie de concepts dédiée à l'univers des mammifères.

plus le processus de recherche de l'information était complexe, plus les concepts étaient distants l'un de l'autre. En terme de combinatoire, ils ont modélisé cette interprétation de leurs tests par un nombre d'arcs plus élevé dans une hiérarchie entre concepts distants qu'entre concepts similaires.

Un exemple connu d'une mesure structurelle est la mesure de Rada qui définit la similarité entre concepts par l'inverse de la longueur du plus court chemin qui les sépare. La hiérarchie de concepts est considérée ici comme un graphe dont les arcs sont les liens is-a et les nœuds les concepts, et au sein duquel des indices combinatoires (profondeur, densité, plus court chemin) sont utilisés pour comparer les nœuds.

4.2.1 Mesure de Rada

[Rada 1989] définit une mesure de similarité comme étant la longueur du plus court chemin entre deux concepts dans la hiérarchie de concepts, notée $dist_{edge}(c_1, c_2)$. La similarité entre $c_1, c_2 \in \mathcal{C}$ est définie par :

$$Sim_{Rad}(c_1, c_2) = \frac{1}{dist_{edge}(c_1,c_2)}$$

Par exemple, sur la hiérarchie [1] de la figure 4.1, nous obtenons les résultats suivants :
$Sim_{Rad}(Panthere, Tigre) = \frac{1}{dist_{edge}(Panthere,Tigre)} = \frac{1}{2} = 0.5$
$Sim_{Rad}(Panthere, Chat) = \frac{1}{dist_{edge}(Panthere,Chat)} = \frac{1}{2} = 0.5$
$Sim_{Rad}(Chat, Chien) = \frac{1}{dist_{edge}(Chat,Chien)} = \frac{1}{4} = 0.25$
$Sim_{Rad}(Chat, Canin) = \frac{1}{dist_{edge}(Chat,Canin)} = \frac{1}{3} = 0.33$

4.2.2 Mesure de Resnik

[Resnik 1995] complète la mesure de Rada avec la profondeur maximale de la hiérarchie (max). La similarité entre les concepts c_1 et c_2 est égale au ratio entre la profondeur maximale de la hiérarchie et le plus court chemin entre ces concepts. La similarité entre $c_1, c_2 \in \mathcal{C}$ est définie par :

1. Sur cet exemple, la hiérarchie est de type arborescente. Dans le cas général, il peut également s'agir d'un treillis.

$$Sim_{Res}(c_1, c_2) = \frac{2*max}{dist_{edge}(c_1,c_2)}$$

Par exemple, sur la hiérarchie de la figure 4.1, avec $max = 3$, nous obtenons les résultats suivants :
$Sim_{Res}(Panthere, Tigre) = \frac{2*3}{dist_{edge}(Panthere,Tigre)} = \frac{6}{2} = 3$
$Sim_{Res}(Panthere, Chat) = \frac{2*3}{dist_{edge}(Panthere,Chat)} = \frac{6}{2} = 3$
$Sim_{Res}(Chat, Chien) = \frac{2*3}{dist_{edge}(Chat,Chien)} = \frac{6}{4} = 1.5$
$Sim_{Res}(Chat, Canin) = \frac{2*3}{dist_{edge}(Chat,Canin)} = \frac{6}{3} = 2$

4.2.3 Mesure de Leacock

[Leacock 1998] normalise la mesure de Resnik au moyen de la fonction *log* de manière à obtenir des résultats dans l'intervalle [0,1], avec la valeur 0 pour des concepts distincts et la valeur 1 pour des concepts totalement similaires. La similarité entre les concepts c_1 et c_2 est égale au ratio entre le plus court chemin entre ces concepts et la profondeur maximale de la hiérarchie. La similarité entre $c_1, c_2 \in \mathcal{C}$ est définie par :

$$Sim_{Lea}(c_1, c_2) = -log\left(\frac{dist_{edge}(c_1,c_2)}{2*max}\right)$$

Par exemple, sur la hiérarchie de la figure 4.1, nous obtenons les résultats suivants :
$Sim_{Lea}(Panthere, Tigre) = -log\frac{dist_{edge}(Panthere,Tigre)}{2*3} = -log\frac{2}{6} = 0.48$
$Sim_{Lea}(Panthere, Chat) = -log\frac{dist_{edge}(Panthere,Chat)}{2*3} = -log\frac{2}{6} = 0.48$
$Sim_{Lea}(Chat, Chien) = -log\frac{dist_{edge}(Chat,Panthere)}{2*3} = -log\frac{4}{6} = 0.18$
$Sim_{Lea}(Chat, Canin) = -log\frac{dist_{edge}(Chat,Canin)}{2*3} = -log\frac{3}{6} = 0.30$

4.2.4 Mesure de Wu et Palmer

[Wu 1994] propose une autre mesure de similarité prenant en compte à la fois (1) la profondeur des concepts comparés dans la hiérarchie et (2) la structure de cette hiérarchie par la proximité relative de leur père commun. La similarité entre $c_1, c_2 \in \mathcal{C}$ est définie par :

$$Sim_{Wu}(c_1, c_2) = \frac{2*prof(c_{com})}{prof(c_1)+prof(c_2)}$$

Par exemple, sur la hiérarchie de la figure 4.1, nous obtenons les résultats suivants :
$Sim_{Wu}(Panthere, Tigre) = \frac{2*prof(Felin)}{prof(Panthere)+prof(Tigre)} = \frac{2*2}{3+3} = 0.66$
$Sim_{Wu}(Panthere, Chat) = \frac{2*prof(Felin)}{prof(Panthere)+prof(Chat)} = \frac{2*2}{3+3} = 0.66$
$Sim_{Wu}(Chat, Chien) = \frac{2*prof(Mobilier)}{prof(Chat)+prof(Chien)} = \frac{2*1}{3+3} = 0.33$
$Sim_{Wu}(Chat, Canin) = \frac{2*prof(Mobilier)}{prof(Chat)+prof(Canin)} = \frac{2*1}{3+2} = 0.4$

4.3 Mesures de type intensionnel

Le processus de catégorisation repose en grande partie sur un regroupement d'objets par similarité [Thibaut 1997]. Ce processus répond en partie à la théorie

Gestaltiste [Koffka 1935]. Cette dernière distingue plusieurs lois, dont deux qui nous intéressent plus particulièrement :

- la loi de **proximité**, qui permet de regrouper des éléments qui apparaissent souvent ensemble, qui sont proches dans une même zone perceptive. C'est le cas des lettres qui forment un mot, des mots qui forment un syntagme. Il s'agit d'un regroupement présentant une certaine cohérence.
- la loi de **similarité**, qui permet de regrouper les éléments qui nous paraissent semblables. Il peut s'agir de similitudes descriptives (au sens perceptibles) ou fonctionnelles.

4.3.1 Similarité intensionnelle

D'un point de vue ensembliste, deux entités sont similaires si le cardinal de l'intersection des ensembles de leurs caractéristiques est plus grand que celui des sous-ensembles restant.

4.3.2 Mesure de Tversky

Cette approche (*cf.* figure figure 4.2) a été notamment développée par le psychologue A. Tversky qui a proposé dans [Tversky 1977] la définition suivante d'une similarité entre deux concepts :

$$sim_{tversky}(A,B) = \alpha.comm(A,B) - \beta.diff(A,B) - \gamma.diff(B,A)$$

où α, β, γ sont des constantes. Si $\beta = \gamma = 0$ et $\alpha = 1$, la similarité entre A et B correspond à la quantité de propriétés en commun. Si $\alpha = 0$, $\beta > 0$ et $\gamma > 0$, les entités A et B sont évaluées suivant ce qui les différencie - nous avons alors une mesure non pas de similarité mais de dissimilarité.

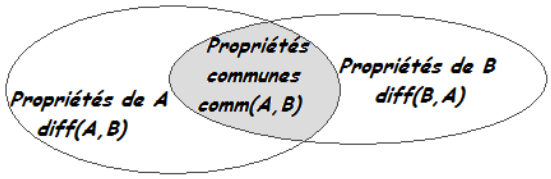


Figure 4.2 – Similarité entre concepts selon Tversky.

Cette mesure est dépendante du cardinal des propriétés de chaque concept. Une seconde version de la formule précédente permet de palier ce problème :

$$sim_{tversky}(A,B) = \frac{comm(A,B)}{\beta.diff(A,B)+\gamma.diff(B,A)+comm(A,B)}$$

La similarité de Tversky a été reprise par [Poitrenaud 1998] pour définir une distance sémantique entre deux catégories par le nombre de propriétés partagées par ces deux catégories au sein d'une hiérarchie conceptuelle. En adoptant le principe de différentialité (*i.e.* héritage des propriétés du concept père auquel sont ajoutées

ses propres caractéristiques - différentes des concepts frères), il faut remonter dans la hiérarchie pour obtenir le plus petit concept père partageant les propriétés communes aux deux concepts comparés. De ce fait, selon [Léger 2005], plus la distance entre deux concepts est grande, plus la différence conceptuelle est importante (les propriétés communes étant également moins nombreuses). [Rips 1973] définit la distance sémantique entre deux concepts comme étant égale au nombre de nœuds du plus court chemin reliant ces concepts (impliquant, selon [Leger 2004], pour avoir un sens le fait que les deux concepts considérés soient sur le même chemin dans la hiérarchie de concepts).

4.4 Mesures de type extensionnel

Un autre moyen d'évaluer la similarité entre deux concepts consiste à comparer de manière ensembliste les instances de chaque concept.

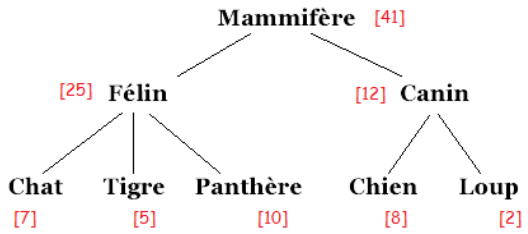


FIGURE 4.3 – Représentation partielle de la classification des mammifères avec cardinalité des instances.

Dans la suite de cette section, par souci de simplification de notation, nous utiliserons : I_P pour $I_{Panthere}$, I_T pour I_{Tigre}, I_C pour I_{Chat}, I_D pour I_{Chien}, I_M pour $I_{Mammifere}$, I_S pour I_{Felin} et I_R pour I_{Canin}. Nous supposons pour les besoins de l'exemple qu'il existe une instance commune aux chats et aux panthères.

4.4.1 Coefficients de Jaccard et Dice

La plupart des mesures de ce type sont inspirées par la similarité de Jaccard, définie dans [Jaccard 1901] et par la fonction :

$$Sim_{Jaccard}(c_1, c_2) = \frac{|I_{c_1} \cap I_{c_2}|}{|I_{c_1}| + |I_{c_2}| - (|I_{c_1} \cap I_{c_2}|)}$$

Par exemple, sur la hiérarchie de la figure 4.3, nous obtenons les résultats suivants :

$Sim_{Jaccard}(Panthere, Tigre) = \frac{|I_P \cap I_T|}{|I_P| + |I_T| - (|I_P \cap I_T|)} = \frac{0}{5+10-0} = 0$

$Sim_{Jaccard}(Panthere, Chat) = \frac{|I_P \cap I_C|}{|I_P| + |I_C| - (|I_P \cap I_C|)} = \frac{1}{10+7-0} = 0.06$

$Sim_{Jaccard}(Chat, Chien) = \frac{|I_C \cap I_D|}{|I_C| + |I_D| - (|I_C \cap I_D|)} = \frac{0}{7+8-0} = 0$

$Sim_{Jaccard}(Chat, Canin) = \frac{|I_C \cap I_R|}{|I_C| + |I_R| - (|I_C \cap I_R|)} = \frac{0}{7+12-0} = 0$

[Dice 1945] possède la même ordonnance que la formule de Jaccard. La mesure de Dice est égal au ratio entre le nombre d'instances en commun et la somme du nombre d'instances des concepts comparés :

$$Sim_{Dice}(c_1, c_2) = \frac{2*|I_{c_1} \cap I_{c_2}|}{|I_{c_1}|+|I_{c_2}|}$$

En reprenant la hiérarchie de la figure 4.3 avec cette mesure, nous obtenons les résultats suivants :
$Sim_{Dice}(Panthere, Tigre) = \frac{2*|I_P \cap I_T|}{|I_P|+|I_T|} = \frac{2*0}{5+10} = 0$
$Sim_{Dice}(Panthere, Chat) = \frac{2*|I_P \cap I_C|}{|I_P|+|I_C|} = \frac{2*1}{10+7} = 0.12$
$Sim_{Dice}(Chat, Chien) = \frac{2*|I_C \cap I_D|}{|I_C|+|I_D|} = \frac{2*0}{7+8} = 0$
$Sim_{Dice}(Chat, Canin) = \frac{2*|I_C \cap I_R|}{|I_C|+|I_R|} = \frac{2*0}{7+12} = 0$

4.4.2 Mesure de d'Amato et al.

[d'Amato 2008] estime que deux concepts peuvent être similaires sans avoir d'instances en commun ; ce qui est le cas pour une similarité de type intensionnelle ou expressionnelle uniquement. Il propose une nouvelle mesure qui évalue, non pas l'intersection entre les ensembles d'instances de chaque concept, mais la variation de la cardinalité du plus petit subsumant commun.

$$Sim_{Ama}(c_1, c_2) = \frac{min(|I_{c_1}|,|I_{c_2}|)}{|I_{c_{com}}|}(1 - \frac{|I_{c_{com}}|}{|I|})(1 - \frac{min(|I_{c_1}|,|I_{c_2}|)}{|I_{c_{com}}|})$$

En reprenant la hiérarchie de la figure 4.3 avec cette mesure (en supposant que cette hiérarchie est un extrait d'une hiérarchie plus importante dont le nombre total d'instances $|I| = 56$), nous obtenons les résultats suivants :
$Sim_{Ama}(Panthere, Tigre) = \frac{min(|I_P|,|I_T|)}{|I_S|}(1 - \frac{|I_S|}{|I|})(1 - \frac{min(|I_P|,|I_T|)}{|I_S|}) = \frac{min(10,5)}{25}(1 - \frac{25}{56})(1 - \frac{min(10,5)}{25}) = 0.09$
$Sim_{Ama}(Panthere, Chat) = \frac{min(|I_P|,|I_C|)}{|I_S|}(1 - \frac{|I_S|}{|I|})(1 - \frac{min(|I_P|,|I_C|)}{|I_S|}) = \frac{min(10,7)}{25}(1 - \frac{25}{56})(1 - \frac{min(10,7)}{25}) = 0.11$
$Sim_{Ama}(Chat, Chien) = \frac{min(|I_C|,|I_D|)}{|I_M|}(1 - \frac{|I_M|}{|I|})(1 - \frac{min(|I_C|,|I_D|)}{|I_M|}) = \frac{min(7,8)}{41}(1 - \frac{41}{56})(1 - \frac{min(7,8)}{41}) = 0.04$
$Sim_{Ama}(Chat, Canin) = \frac{min(|I_C|,|I_R|)}{|I_M|}(1 - \frac{|I_M|}{|I|})(1 - \frac{min(|I_C|,|I_R|)}{|I_M|}) = \frac{min(7,12)}{41}(1 - \frac{41}{56})(1 - \frac{min(7,12)}{41}) = 0.04$

4.5 Mesures de type expressionnel

Les concepts peuvent être également comparés sur un plan expressionnel, avec les termes qui les dénotent, à l'aide notamment du contenu en information défini par Resnik [Resnik 1993].

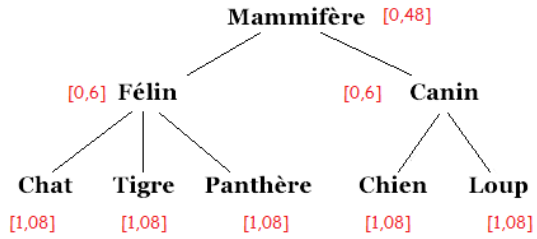


FIGURE 4.4 – Représentation partielle de la classification des mammifères avec contenu en information.

4.5.1 Mesure de Resnik

[Resnik 1995] calcule le contenu en information d'un concept $c \in (C)$ en se fondant sur la probabilité $p(c)$ d'avoir ce concept dans un corpus donné. Ce contenu en information est défini par :

$$\Psi(c) = -log(p(c)) \text{ avec } p(c) = \frac{\sum_{t \in monde(c)} count(t)}{N}$$

où :
- N représente le nombre total d'occurrences des termes de tous les concepts dans le corpus ;
- $monde(c)$ représente l'ensemble des termes possibles pour le concept c mais également pour l'ensemble de ses descendants dans la hiérarchie. Par exemple, avoir les mots " *chat* ", " *chien* " et " *clébard* " va renforcer la présence du concept *Mammifère*, il faut donc les prendre en compte.

Plus le concept est générique, plus son contenu en information est faible - *i.e.* il apporte peu d'information - et inversement plus il est spécifique plus son contenu en information est fort. Mais cette mesure est très sensible à l'ambiguïté, ce qui peut avoir pour effet d'augmenter de façon erronée l'importance du concept étudié. Ceci pré-suppose donc que chaque terme est attribué de façon unique à un concept.

[Sanderson 1999] tente de corriger ce problème en modifiant la calcul de la fréquence d'apparition d'un terme ($\sum count(n)$) par :

$$freq(n) = \sum \frac{count(t)}{nb_{classe}(t)}$$

où $nb_{classe}(t)$ est égal au nombre de concepts dont le terme t est label.

Pour établir la similarité entre deux concepts, [Resnik 1995] propose de calculer la valeur du contenu informatif qu'ils partagent, à savoir la plus grande de celle de l'un de leurs concepts ascendants (*i.e.* le concept le plus spécifique les subsumant tous les deux). La similarité entre $c_1, c_2 \in \mathcal{C}$ est définie par :

$$Sim_{Res2} = max_{c \in S(c_1,c_2)} \Psi(c)$$

Par exemple, sur la hiérarchie de la figure 4.4, nous obtenons les comportements suivants :
$Sim_{Res2}(Panthere, Tigre) = \Psi(Felin) = 0.6$
$Sim_{Res2}(Panthere, Chat) = \Psi(Felin) = 0.6$
$Sim_{Res2}(Chat, Chien) = \Psi(Mammifere) = 0.38$
$Sim_{Res2}(Chat, Canin) = \Psi(Mammifere) = 0.38$

4.5.2 Mesure de Lin

[Lin 1998] propose d'évaluer la similarité entre deux concepts, en tenant compte à la fois de leur contenu informatif commun (comme la mesure de Resnik) mais également de leurs caractéristiques propres. La similarité entre $c_1, c_2 \in \mathcal{C}$ est définie par :

$$Sim_{Lin} = \frac{2*\Psi(c_{com})}{\Psi(c_1)+\Psi(c_2)}$$

Par exemple, sur la hiérarchie de la figure 4.4, nous obtenons les résultats suivants :
$Sim_{Lin}(Panthere, Tigre) = \frac{2*\Psi(Felin)}{\Psi(Panthere)+\Psi(Tigre)} = \frac{2*0.6}{1.08+1.08} = 0.56$
$Sim_{Lin}(Panthere, Chat) = \frac{2*\Psi(Felin)}{\Psi(Panthere)+\Psi(Chat)} = \frac{2*0.6}{1.08+1.08} = 0.56$
$Sim_{Lin}(Chat, Chien) = \frac{2*\Psi(Mobilier)}{\Psi(Chat)+\Psi(Chien)} = \frac{2*0.38}{1.08+1.08} = 0.35$
$Sim_{Lin}(Chat, Canin) = \frac{2*\Psi(Mobilier)}{\Psi(Chat)+\Psi(Canin)} = \frac{2*0.38}{1.08+0.6} = 0.45$

4.5.3 Mesure de Jiang & Conrath

Fondées sur le contenu en information, d'autres mesures ont été proposées telle que celle de Jiang & Conrath [Jiang 1997] :

$$distsem_{Jiang}(c_1, c_2) = \sum[\Psi(c) - \Psi(pere(c))] * TC(c, pere(c))$$

où :
- $c \in chemin(c_1, c_2)$ dans la hiérarchie de concepts, privé de c_{com} ;
- $TC(c_i, c_j)$ une pondération de l'arc reliant le concept c_i au concept c_j tel $c_j = pere(c_i)$.

Appliquée à une hiérarchie de concepts avec une pondération égale à 1 pour les arcs de type hiérarchique, la distance de Jiang & Conrath est définie par :

$$distsem_{Jiang}(c_1, c_2) = (\Psi(c_1) + \Psi(c_2)) - 2\Psi(c_{com})$$

Par exemple, sur la hiérarchie de la figure 4.4, nous obtenons les résultats suivants :
$distsem_{Jiang}(Panthere, Tigre) = (\Psi(Panthere) + \Psi(Tigre)) - 2 * \Psi(Felin) = (1.08 + 1.08) - (2 * 0.6) = 0.96$
$distsem_{Jiang}(Panthere, Chat) = (\Psi(Panthere)+\Psi(Chat))-2*\Psi(Felin) = (1.08+1.08) - (2 * 0.6) = 0.96$

$distsem_{Jiang}(Chat, Chien) = (\Psi(Chat) + \Psi(Chien)) - 2 * \Psi(Mammifere) = (1.08 + 1.08) - (2 * 0.38) = 1.4$
$distsem_{Jiang}(Chat, Canin) = (\Psi(Chat) + \Psi(Canin)) - 2 * \Psi(Mammifere) = (1.08 + 0.6) - (2 * 0.38) = 0.92$

4.6 Conclusion

	Panthère, Tigre	*Panthère, Chat*	*Chat, Chien*	*Chat, Canin*
Rada	0,5	0,5	0,25	0,33
Resnik	3	3	1,5	2
Leacock	0,48	0,48	0,18	0,3
Wu Palmer	0,66	0,66	0,33	0,4
Jaccard	0	0.06	0	0
Dice	0	0.12	0	0
Amato	0,09	0,11	0,04	0,04
Resnik2	0,6	0,6	0,38	0,38
Lin	0,56	0,56	0,36	0,45
Jiang	0,96	0,96	1,4	0,92

TABLE 4.1 – Récapitulatif des différentes valeurs de similarités.

Les mesures de similarité sémantiques décrites dans ce chapitre représentent avant tout un outil permettant un ordonnancement des concepts. Le tableau 4.1 donne les valeurs obtenues avec chacune des mesures étudiées dans ce chapitre. Chacune reflète un point de vue qui influence parfois fortement l'interprétation du résultat.

Les mesures fondées sur les propriétés nous semblent à même de répondre aux critères définis par la loi de similarité de la théorie Gestaltiste. Deux concepts ne peuvent être similaires que s'ils partagent un grand nombre de leur propriété fonctionnelles et descriptives. De ce point de vue, il est possible de classer par ordre croissant de similarité les paires de concepts suivants : *(Chat,Chien), (Chat,Canin), (Panthère,Tigre) et (Panthère,Chat).*

Les mesures de type structurel n'offrent pas la même précision que les mesures reposant sur les propriétés, lesquelles correspondent beaucoup plus à un mode de fonctionnement fondé sur l'intuition. Contrairement aux résultats donnés par les mesures fondées sur les propriétés, la panthère est autant similaire à un chat qu'à un tigre. Néanmoins, l'ordonnancement est semblable aux premières mesures. Un second point à noter sur ce type de mesure est la modélisation adoptée. En effet, tous les arcs " *is-a* " de ces hiérarchies de concepts sont supposés posséder la même "valeur sémantique". Or, ils sont fortement dépendants de la granularité de la conceptualisation. Si celle-ci n'est pas uniforme dans toute la hiérarchie, alors les liens n'ont pas

la même intensité. Par exemple, pour certains individus, la relation " *is-a* " entre les concepts *Chat* et *Félin* est plus faible que la relation " *is-a* " entre les concepts *Poireau* et *Végétal*. Pourtant, il est possible de tenir compte de ces différences dans la modélisation.

Hormis avec la mesure de d'Amato et al., si des concepts ne possèdent aucune instance commune, alors ils ne peuvent être similaires. Notons que les coefficients de Dice et de Jaccard peuvent être utilisés également pour calculer des similarités de manière expressionnelle en comparant les ensembles de termes propres à chaque concept. Concernant la mesure de d'Amato et al., il est possible de faire deux remarques. Tout d'abord, sur l'exemple qui nous a servi à illustrer ce chapitre, si nous avons un nombre d'instances inférieur à sept pour chaque concept descendant de *Mammifère* (hormis Félin), alors les similarités entre le concept *Chat* et tout concept descendant de *Canin* ont la même valeur. Ensuite, si nous nous référons à la figure 4.5, dans les deux cas, les concepts c_1 et c_2 ont la même valeur de similarité pour la formule de d'Amato. Or, dans le second cas, l'ensemble des instances de c_2 est inclus dans l'ensemble des instances de c_1. D'un point de vue intensionnel, la similarité entre c_1 et c_2 devrait être beaucoup plus élevée dans le second cas que dans le premier.

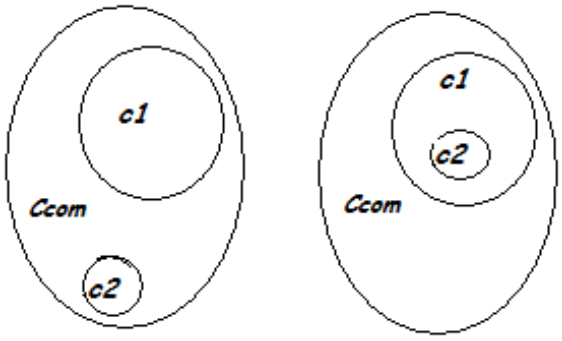


FIGURE 4.5 – Similarité par la mesure de d'Amato and al..

Les mesures fonctions du contenu informationnel d'un concept suivent une approche que l'on peut qualifier de mixte (arc, termes), même si elles sont essentiellement expressionnelles. Elles possèdent comme inconvénient d'être fortement dépendantes de l'interprétation du domaine par leurs auteurs, et non du domaine lui seul. Elles sont également fortement dépendantes de la nature du corpus et surtout de la qualité de l'algorithme de détection des termes. Citons comme frein à la qualité de cette approche :

- l'homonymie (un même terme peut dénoter plusieurs concepts) ;
- les anaphores (le terme est écrit une fois mais cité indirectement plusieurs fois, par exemple *j'ai rencontré* SA SŒUR, *je* L'*aime bien*) ;

– le contexte (l'entourage des mots en modifie le sens, par exemple *le* PETIT*, il n'est vraiment* PAS GRAND).

Enfin, selon [Hernandez 2006], " *l'inconvénient de la mesure de Resnik est que deux paires de concepts qui ont le même subsumant le plus spécifique ont la même similarité. Ceci est par exemple le cas entre (Persan et Labrador) et (Chat et Mammifère).* ". Dans notre exemple, *Félin* et *Mammifère* ont la même valeur de similarité que *Chien* et *Chat*.

Partie II

Contributions scientifiques

L'objet applicatif de cette thèse est (1) la réalisation d'une ontologie du domaine Hygiène-Sécurité-Environnement (HSE), et (2) son exploitation dans le cadre d'une recherche d'information au sein d'une base de textes réglementaires. L'ontologie construite dans le cadre de cette thèse s'inscrit dans une approche pragmatique de la connaissance. Il s'agit de formaliser une perception particulière du domaine HSE, perception de ce domaine par les consultants de l'entreprise Tennaxia au travers d'un ensemble de textes réglementaires, textes jugés représentatifs du domaine par ces consultants. Notre approche de l'ingénierie ontologique nous a conduit à proposer une nouvelle typologie des ontologies qui distingue (1)les ontologies de domaine, (2) les ontologies vernaculaires de domaine et (3) les ontologies pragmatisées vernaculaires de domaine. Cette pragmatisation se concrétise par une différence de typicalité pour (1) des termes dénotant un concept, (2) des instances rattachées à un concept, et (3) un ensemble de sous-concepts pour un concept donné. Ces différences de représentativité sont respectivement exprimées au moyen de gradients de prototypicalité lexicale, extensionnelle, conceptuelle descendante et ascendante. L'exploitation de cette ontologie pragmatisée, dans le cadre d'une recherche d'information au sein d'une base de textes réglementaires, se fait également au moyen de mesures sémantiques. Aujourd'hui utilisées au sein d'un ensemble grandissant d'applications, leur dénomination, qu'il s'agisse de distances ou de simples mesures, entraine une certaine confusion entre similarité et proximité. Nos travaux nous ont amenés à préciser la distinction entre ces deux notions. Nous avons par conséquent défini deux mesures distinctes pour évaluer chacune de ces notions dans le cadre sémiotique d'une conceptualisation.

—

Chapitre 5

Approche sémiotique des ontologies de domaine

5.1 Introduction

Les ontologies que nous voulons personnaliser intègrent les trois dimensions introduites par Peirce dans sa sémiotique : le *signifié*, c'est-à-dire le concept en intension, le *signifiant*, c'est-à-dire les termes désignant le concept, et le *référent*, c'est-à-dire le concept en extension. Nous proposons d'exploiter ces trois dimensions dans nos travaux (gradient de prototypicalité conceptuelle (chapitre 6) et mesures sémantiques (chapitre 7), ce qui nous a conduit à qualifier notre approche de "sémiotique".

5.2 Coordonnées cognitives d'un individu

L'utilisation de ces trois dimensions permet de moduler, dans le calcul, l'importance des aspects intensionnel, extensionnel et expressionnel dans la conceptualisation des utilisateurs. Ces différences d'importance sont conditionnées par le domaine traité, l'univers cognitif des utilisateurs et le contexte d'utilisation. Ainsi, dans le domaine des mathématiques, les concepts sont plutôt manipulés en intension. Dans le domaine des espèces animales, un zoologue a tendance à les conceptualiser en intension (par des propriétés biologiques), alors que la plupart d'entre nous utilisons davantage des conceptualisations extensionnelles (basées sur les animaux rencontrés au cours de leur vie).

Les importances relatives des différentes composantes intensionnelle, extensionnelle et expressionnelle sont modélisées par des coefficients α, β et γ qui sont des coefficients positifs ou nuls. Dans un souci de normalisation, nous imposons que les mesures et leurs composantes varient de 0 à 1, et que $\alpha + \beta + \gamma = 1$. Les valeurs de ces trois coefficients peuvent être fixées arbitrairement ou calibrées par expérimentations. Mais nous proposons une méthode pour les évaluer automatiquement, méthode basée sur le principe suivant. Les rapports entre α, β et γ expriment d'une

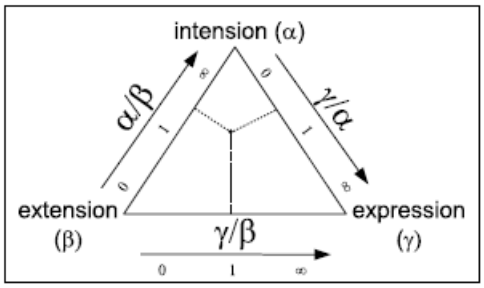


FIGURE 5.1 – Les coefficients de pondération des composantes du SPG comme coordonnées dans le triangle sémiotique. γ/α proche de 0 indique que l'utilisateur a une approche beaucoup plus intensionnelle qu'expressionnelle du domaine, le même rapport proche de 1 indique le contraire, de même pour les autre rapports.

certaine façon les coordonnées cognitives de l'utilisateur dans le triangle sémiotique (cf. figure 5.1).

Nous avons choisi de calculer les valeurs de γ/α et γ/β, les valeurs α, β et γ étant déduites de ces rapports et de l'équation $\alpha+\beta+\gamma = 1$. Dans le cas où aucun corpus n'est disponible, seul le rapport α/β peut être calculé. Il est évalué par la moyenne, sur l'ensemble des concepts, des rapports entre l'importance des propriétés portées par un concept et le nombre des instances de ce concept. En effet, un concept portant des propriétés importantes, mais ayant peu d'instances (par exemple un dragon) est conceptualisé davantage de façon intensionnelle (on se souvient des propriétés descriptives ou fonctionnelles), alors qu'un concept portant des propriétés banales mais ayant de nombreuses instances (par exemple une voiture) est conceptualisé de façon extensionnelle (on se souvient de la voiture qu'on rencontre le plus souvent).

Le rapport γ/α représente le rapport entre ce qui est conceptualisé par l'utilisateur et ce qui est exprimé dans le corpus. Il s'agit du rapport entre ce qui est expressionnel, c'est-à-dire les termes du corpus désignant des concepts non présents dans l'ontologie, et ce qui est intensionnel, c'est-à-dire les concepts de l'ontologie non exprimés dans le corpus. Cependant, nous considérons que l'ontologie couvre tout le corpus, c'est-à-dire que tous les termes du corpus renvoient bien à des concepts, relations ou instances de l'ontologie. Aussi, le rapport γ/α évolue entre 0 et 1 et est approximé par le taux de couverture des concepts de l'ontologie par le corpus. Ce taux est égal au nombre de concepts dont au moins un des termes apparaît dans le corpus, divisé par le nombre total de concepts. De même, γ/β est approximé par le taux de couverture des instances de l'ontologie par le corpus, qui est égal au nombre d'instances dont au moins un des termes apparaît dans le corpus, divisé par le nombre total d'instances.

5.3 Typologie des ontologies de domaine

Toute donnée perçue par un individu appartenant à une communauté devient une connaissance relative à un endogroupe [1], chaque endogroupe possédant un référentiel spécifique, plus ou moins différent de celui d'un autre endogroupe sur un même domaine.

Nous considérons ainsi plusieurs niveaux dans la typologie des ontologies, laquelle tente d'appréhender les notions, complexes à définir et à modéliser, de "subjectivité" et "d'objectivité" des connaissances. Elle est donc complémentaire des typologies d'ontologies déjà proposées en Ingénierie des Ontologies (IO), comme celle introduite dans [Gomez-Perez 2003] :

- *ontologies de représentation*, définissant les primitives d'un paradigme de représentation des connaissances (par exemple la Frame Ontology pour le paradigme des frames [Gruber 1993b]) ;
- *ontologies fondationnelles*, définissant des notions abstraites et/ou de sens commun telles que le temps, l'espace, les quantités, par exemple SUO, Standard Upper Ontology proposée par un groupe de travail IEEE [2], ou encore DOLCE ;
- *ontologies linguistiques*, définissant des ensembles d'unités lexicales liées par des relations linguistiques telles que la synonymie ou l'hyperonymie (par exemple Wordnet [3]) ;
- *ontologies de domaine*, définissant les connaissances d'un domaine borné (par exemple UMLS, Unified Medical Language System, pour le domaine médical [4]) ;
- *ontologies de PSM - Problem Solving Method*, définissant des méthodes de résolution de problèmes génériques (par exemple les tâches et méthodes de Chandrasekaran [Chandrasekaran 1998]).

Notre typologie (*cf.* figure 5.2) distingue trois types d'ontologies : (1) les *ontologies de domaine* (OD), les plus globales, (2) *les ontologies vernaculaires de domaine* (OVD), des ontologies de domaine propres à un endogroupe, et (3) *les ontologies pragmatisées vernaculaire de domaine* (OPVD), des ontologies vernaculaires de domaine dans un contexte donné.

1. [Tajfel 1979, Tajfel 1986] établit une définition du concept d'endogroupe comme étant une collection d'individus qui se perçoivent comme membres d'une même catégorie, qui attachent une certaine valeur émotionnelle à cette définition d'eux-mêmes et qui ont atteint un certain degré de consensus concernant l'évaluation de leur groupe et de leur appartenance à celui-ci.
2. http ://suo.ieee.org
3. http ://www.wordnet.princeton.edu/
4. http ://www.nlm.nih.gov/research/umls/

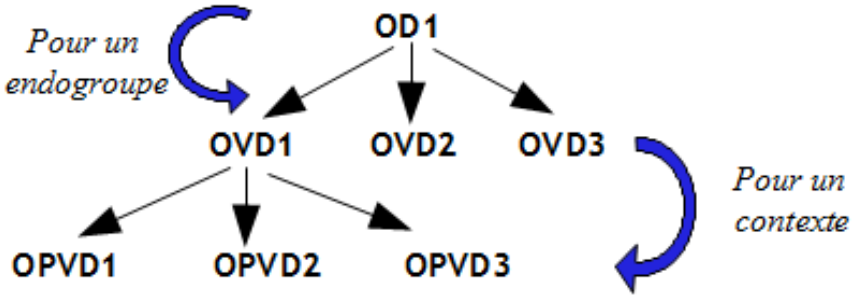


FIGURE 5.2 – Typologie des ontologies.

5.4 Ontologies de domaine (OD)

Une ontologie de domaine est souvent considérée comme une spécification formelle et explicite d'une conceptualisation partagée [Gruber 1993a]. Elle correspond à une modélisation des connaissances spécifiques à un domaine particulier, manipulable et intelligible par des agents tant logiciels qu'humains (*cf.* section 2.6).

5.5 Ontologies vernaculaires de domaine (OVD)

Dans la pratique, définir une ontologie pour un domaine particulier consiste à établir une synthèse consensuelle de connaissances d'individus appartenant à un endogroupe – endogroupe formant une communauté d'usage ou d'intérêt propre à ce domaine[5]. Cette modélisation repose souvent sur : un treillis de concepts, un treillis de relations, un ensemble d'axiomes et un ensemble d'instances. Elle est construite à partir des connaissances relatives exprimées par leurs détenteurs, c'est-à-dire les membres de l'endogroupe. Ces connaissances sont extraites à partir de documents tels des textes, des images, des sons, des documents produits par ces membres (ces documents sont le reflet d'une perception au travers de leur sens, mais également de leur raisonnement et de leur culture).

Nous qualifions ces ontologies de domaine d'**ontologies vernaculaires de domaine**. Le qualificatif *vernaculaire* provient du latin *vernaculus*, qui - à l'origine - fait référence aux esclaves nés dans la maison. Historiquement, ce terme a été utilisé pour désigner tout ce qui était fabriqué, dressé, élevé, tissé, cultivé au sein d'un pays (au sens de terroir, d'une unité géographique et écologique déterminée). Aujourd'hui, le mot *vernaculaire* est surtout usité au sens de *indigène*. Ainsi, une langue est dite vernaculaire lorsqu'elle est parlée uniquement à l'intérieur d'une communauté, d'un endogroupe. De même, une architecture est dite vernaculaire lorsqu'elle se réfère à un type de construction propre à une époque spécifique ou un terroir précis (et qui

5. Il est ainsi possible d'avoir plusieurs ontologies pour un même domaine, à partir du moment où elles sont définies pour des endogroupes différents. Par exemple, pour un domaine donné, il peut y avoir plusieurs points de vue, chacun correspondant à une "école de pensée".

utilise les matériaux propres à ce terroir). Une ontologie de domaine est qualifiée de vernaculaire lorsqu'elle est fabriquée à partir des matériaux cognitifs en provenance d'individus appartenant à un même endogroupe [6].

Le terme *vernaculaire* est très lié à la notion d'écologie, dans le sens de l'interaction entre la biocénose (l'ensemble des êtres vivants dans un environnement donné) et son biotope (le milieu dans lequel ils vivent). E. Rosch [Rosch 1975a, Rosch 1975b] qualifie d'*écologique* le processus d'élaboration de connaissances. Elle estime que ce processus est certes fonction de l'endogroupe, mais avec une influence certaine du contexte / milieu dans lequel il évolue. Ce contexte, pour un domaine D et un endogroupe G, peut être défini par une liste de paramètres, où tout changement de valeur d'au moins l'un d'entre eux entraîne *de facto* une transformation de l'ontologie. Cela peut être - par exemple - des facteurs entraînant des modifications d'état comme le temps (avec des unités temporelles variables, suivant les domaines et les endogroupes considérés).

5.5.1 Ontologies personnalisées vernaculaires de domaine (OPVD)

Les OVD peuvent être vues d'une manière pragmatique [7], c'est-à-dire suivant différents points de vue (contexte d'utilisation des connaissances stockées), sans pour autant remettre en cause la sémantique (formelle) inhérente à l'ontologie considérée. On peut retrouver dans cette approche une inspiration de la conception du langage selon Saussure qui différenciait la "face sémantique" dont le siège serait l'esprit, et la face "sensorielle" pour la communication (perception et émission) influencée par le contexte.

Nous qualifions les OVD placées dans un contexte d'utilisation particulier d'**ontologies pragmatisées vernaculaires du domaine**.

Le passage du stade d'OVD à OPVD se fait par le biais d'un processus de pragmatisation d'une OVD. Cette transformation d'un modèle en un autre consiste dans un premier temps à définir un contexte d'utilisation des connaissances stockées. De manière plus générale, il est possible de parler de pragmatisation d'une ontologie lorsqu'il s'agit de l'adapter à un contexte d'utilisation, selon les utilisateurs (personnalisation) mais aussi selon l'application envisagée (opérationnalisation). Ce contexte, propre à un individu ou un sous-ensemble d'individus de l'endogroupe, est décrit par un ensemble de valeurs affectées à des paramètres liés à la culture, aux modes d'apprentissage, ou encore au contexte émotionnel.

6. [Troadec 2007] établit que ces conceptions dans l'esprit humain sont très liées à sa culture même si elles visent l'universel, et ce de manière asymptotique.

7. Au sens de la pragmatique linguistique, qui considère le contexte comme indispensable à l'interprétation des textes.

5.6 Conclusion

Dans ce chapitre, nous avons introduit notre approche pragmatique des ontologies de domaine à travers une typologie distinguant trois niveaux. Chaque niveau prend en compte un nouvel élément permettant une personnalisation plus avancée des ontologies : le domaine, l'endogroupe, et le contexte d'utilisation. Nous avons intégré une approche "sémiotique" de la conceptualisation d'un point de vue utilisateur, à travers un triplet de coordonnes cognitives, *i.e.* une pondération des dimensions intensionnelle, extensionnelle et expressionnelle. C'est dans un tel cadre que nous proposons le calcul de plusieurs mesures visant à évaluer le degré de typicalité d'un élément par rapport à un autre d'une part, et deux mesures sémantiques d'autre part, le tout pour un domaine, un endogroupe et un contexte donné.

Chapitre 6

Gradients de prototypicalité

6.1 Introduction

On considère ici la définition du prototype introduite dans la section 3.3.2, à savoir le meilleur exemplaire (le plus représentatif) pour chacun des concepts selon un consensus social de l'endogroupe concerné. Les prototypicalités s'expriment par des gradients numériques qui pondèrent les liens *is-a* entre concepts, mais également les propriétés des concepts, les termes qui les désignent, et leurs instances. Trois types de prototypicalité sont ainsi distinguées :

- la **prototypicalité conceptuelle** : deux concepts liés hiérarchiquement peuvent être plus ou moins proches sémantiquement [Kleiber 2004]. Plus précisément, au sein d'une fratrie de concepts, certains seront plus prototypiques de leur père commun que les autres. Par exemple, parmi tous les modèles d'avion, le modèle le plus représentatif, celui auquel on pense le plus volontiers lorsqu'on pense à un avion, est plutôt du type des avions commerciaux modernes que du type des premiers biplans ou d'un avion mu par la force musculaire.
- la **prototypicalité lexicale** : pour un concept donné pouvant être dénoté par plusieurs termes, certains termes sont utilisés plus volontiers que d'autres. Par exemple, de nos jours, on utilise plus souvent le terme *avion* que le terme *aéroplane* pour dénoter le concept de véhicule de navigation aérienne plus lourd que l'air.
- la **prototypicalité extensionnelle** : pour un concept donné possédant plusieurs instances, certaines d'entre elles sont plus représentatives que d'autres. Par exemple, pour toute une génération, *Milou* et *Idéfix* sont plus représentatifs, typiques, du concept de *Chien de fiction* que *Patmol*.

Nous utilisons l'expression *gradient de prototypicalité* car :

- ce qui est calculé n'est pas une distance (pas de symétrie et pas d'inégalité triangulaire) mais une mesure qui reflète la typicalité d'un concept par rapport à un autre, ou par rapport à un terme ou une instance ;
- nous définissons un degré de typicalité, degré qui peut varier sur une échelle arbitraire (l'élément qui a le plus haut gradient est considéré comme le plus prototypique de sa catégorie) et qui n'a de sens que relativement aux autres

degrés de typicalité mesurés pour le même concept ;
– ces mesures modélisent une différence intuitive de degré de vérité dans un processus de catégorisation [Kleiber 2004].

Nous considérons une ontologie O, pour un domaine D et un endogroupe G, comme un t-uple :

$$O_{(D,G)} = \{\mathcal{C}, \mathcal{P}, \mathcal{I}, \leq^C, \leq^P, dom, codom, \sigma, L\} \text{ où}$$

– $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$ sont respectivement les ensembles de concepts, de relations et d'instances ;
– $\leq^C$: $\mathcal{C} \times \mathcal{C}$ et $\leq^P$: $\mathcal{P} \times \mathcal{P}$ sont des ordres partiels sur les hiérarchies de concepts et de relations [1] ;
– $dom : \mathcal{P} \to \mathcal{C}$ et $codom : \mathcal{P} \to (\mathcal{C} \cup Datatypes)$ associent à chaque relation son domaine et son co-domaine ;
– $\sigma : \mathcal{C} \to \wp(\mathcal{I})$ associe à chaque concept ses instances ;
– $L = \{L_C \cup L_P \cup L_I, term_c, term_p, term_i\}$ est le dialecte utilisé par G pour évoquer le domaine D, où L_C, L_P et L_I sont les ensembles de termes associés à $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$, où $term_c : \mathcal{C} \to \wp(L_C)$, $term_p : \mathcal{P} \to \wp(L_P)$ et $term_i : \mathcal{I} \to \wp(L_I)$ sont les fonctions qui associent chaque concept, relation et instance aux termes utilisés pour les dénoter.

6.2 Gradients de prototypicalité conceptuelle

Situés dans un cadre sémiotique, les gradients de prototypicalité conceptuelle (Semiotic Prototypicality Gradient, SPG) sont un agrégat de trois composantes :

1. une composante **intensionnelle**, fondée sur les propriétés des concepts ;
2. une composante **extensionnelle**, fondée sur les instances ;
3. une composante **expressionnelle**, fondée sur les termes apparaissant au sein du corpus.

Le calcul de chaque composante fait appel à une ressource propre à l'univers cognitif de l'endogroupe, ou de l'utilisateur, considéré. Ainsi, le calcul de la composante intensionnelle nécessite l'existence de propriétés dans l'ontologie. Le calcul de la composante extensionnelle nécessite un ensemble d'instances propres à l'endogroupe, ou à l'utilisateur, considéré. Le calcul de la composante expressionnelle requiert un corpus de documents jugés représentatifs, par l'endogroupe ou l'utilisateur, du domaine de connaissance modélisée dans l'ontologie.

Une autre ressource propre à l'univers cognitif de l'endogroupe, ou de l'utilisateur, peut être utilisée : la pondération des liens entre propriétés et concepts. Ces poids expriment, pour chaque concept et chaque propriété, l'importance d'une propriété dans

1. $c_1 \leq^C c_2$ signifie que le concept c_2 subsume le concept c_1.

la définition du concept [2]. Pour chaque propriété, tous les concepts qui la partagent sont ordonnés dans l'intervalle [0,1]. Par exemple, pour la propriété *a_pour_auteur*, le concept *Article scientifique* peut être placé en premier (proche de 1), le concept *Article de presse* peut être placé après selon un ordre défini (parce qu'avoir un auteur est une propriété un peu moins importante dans la définition d'un article de presse), et le concept *Notice d'utilisation* peut être placé près de 0 (importance de cette propriété quasi nulle dans la définition du concept).

Toutes ces ressources (instances, corpus, propriétés et poids des propriétés) ne sont pas toujours disponibles en pratique, mais la méthode que nous proposons pour calculer ces gradients de prototypicalité demeure valide quelque soit le type de ressources disponibles. Par exemple, s'il n'y a aucune instance, seules les composantes intensionnelles et expressionnelles peuvent être calculées. De même, si l'ontologie est la seule ressource disponible, sans aucune instance, aucune pondération sur les propriétés ni aucun corpus, seule la composante intensionnelle est calculée avec toutes les pondérations sur propriétés égales à 1.

Ce modèle permet également de pondérer chaque composante des *spg* pour prendre en compte l'importance des aspects intensionnel, extensionnel et/ou expressionnel dans la conceptualisation du domaine, au moyen des coordonnées cognitives respectives α, β et γ définies dans le chapitre précédent. Cette modulation est conditionnée par le domaine, mais aussi par la nature de l'endogroupe (ou de l'individu) et le contexte applicatif. Par exemple, dans le domaine des mathématiques, l'aspect intensionnel prévaut. Dans le domaine de la zoologie, un expert construit sa conceptualisation sur les propriétés biologiques (l'aspect intensionnel prévaut), mais la majeure partie des individus utilisent habituellement une conceptualisation extensionnelle, fondée sur les animaux rencontrés dans leur univers.

6.2.1 Objectif

L'objectif du gradient de prototypicalité conceptuelle est de mesurer la typicalité d'un sous-concept par rapport à un concept donné. Ce gradient est défini sur les liens hiérarchiques $\leq^C$ reliant deux concepts c_p et c_f avec c_p sur-concept de c_f. Il permet d'évaluer le fait qu'au sein d'une descendance d'un concept donné, certains sous-concepts sont plus prototypiques de leur père que d'autres.

6.2.2 Définition générale

Formellement, nous définissons le gradient de prototypicalité conceptuelle comme une fonction $spg : \mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ qui, à tout couple de concepts $(c_f, c_p) \in \mathcal{C} \times \mathcal{C}$ tel que $c_f \leq^C c_p$ associe la valeur :

2. Plusieurs méthodes peuvent être envisagées pour établir de telles pondérations. La tâche peut soit être confiée à l'ingénieur des connaissances, soit répartie sur plusieurs experts du domaine et représentatifs de l'endogroupe considéré.

$$spg(c_f, c_p) = \alpha * intension(c_f, c_p) + \beta * extension(c_f, c_p) + \gamma * expression(c_f, c_p)$$

Les fonctions *intension*, *extension* et *expression* sont respectivement détaillées en section 6.2.3, 6.2.5 et 6.2.4.

6.2.3 Composante intensionnelle

La composante intensionnelle de *spg* mesure la typicalité d'un concept c_f par rapport à son sur-concept c_p en comparant les propriétés qui leur sont rattachées. Il est possible de calculer la composante intensionnelle comme un rapport entre le nombre de propriétés ajoutées par le concept fils et le nombre de propriétés totales du fils [Aimé 2008a] : un concept est ainsi d'autant plus représentatif de son père qu'il ajoute moins de propriétés à son intension.

La méthode de calcul utilisée ici s'inspire de [Au Yeung 2006] et s'appuie sur la représentation des concepts par des vecteurs dans l'espace des propriétés de l'ontologie. Le principe consiste à calculer dans cet espace un vecteur prototype du concept père c_p. La prototypicalité du concept fils c_f est une distance entre le vecteur représentant c_f et le vecteur prototype du père c_p. Cependant, [Au Yeung 2006] utilise des vecteurs de valeurs de vérité floues comme coordonnées, alors que nos coordonnées sont des valeurs mesurant l'importance de la propriété dans la définition du concept. Par exemple, la propriété *peut flotter* est plus importante dans la définition du concept *Canard* que dans celle du concept *Avion*. Formellement, à tout concept $c \in \mathcal{C}$, est associé le vecteur caractéristique $\vec{v_c} = (v_{c1}, v_{c2}, ..., v_{cn})$ avec $n = |\mathcal{P}|$ et $v_{ci} \in [0, 1], \forall i \in [1, n]$. v_{ci} est la pondération fixée par l'utilisateur pour le concept c par rapport à la propriété i (v_{ci} vaut 1 si l'utilisateur n'a pas fixé ces pondérations).

Le vecteur prototype d'un concept c_p a été originellement introduit dans [Au Yeung 2006] comme une moyenne des vecteurs des concepts fils de c_p. Cependant, [Au Yeung 2006] ne prend en compte dans la moyenne que les concepts qui héritent directement de c_p, alors que nous étendons le calcul à tous les concepts de la descendance. En effet, des propriétés qui apparaissent uniquement sur des descendants indirects du concept père peuvent pourtant apparaître dans le prototype de c_p, en particulier si l'aspect intensionnel est important. Par exemple, dans le cas du concept *chercheur*, le fait d'avoir une blouse blanche n'est pas une propriété du concept, mais peut très bien apparaître dans le prototype du concept. Le vecteur prototype p_{c_p} est donc un vecteur dans l'espace des propriétés, où l'importance de la propriété i est la moyenne des importances des propriétés des concepts de la descendance de c_p possédant i. Nous définissons le vecteur prototype d'un concept c comme suit :

$$\vec{p_{c_p}} = \frac{1}{\sum_{s \in S} \lambda(s)} \sum_{s \in S} \lambda(s) \, \vec{v_s}$$

Où :

- $\lambda(s)$ est égal à $\frac{depth_{tree}(c_p)-depth(s)+1}{depth_{tree}(c_p)}$ avec :
  - $depth_{tree}(c_p)$, la profondeur de la hiérarchie de concepts ayant pour racine le concept c_p ;
  - $depth(s)$, la profondeur du concept s dans la hiérarchie de concepts ayant pour racine le concept c_p.
- S, l'ensemble des concepts de la descendance du concept c_p.

L'objectif du coefficient $\lambda(s)$ (pour un concept s) est de "relativiser" les propriétés qui sont hiérarchiquement distantes du concept c. La composante intensionnelle est donc :

$$intension(c_f, c_p) = 1 - d(\vec{v_{c_f}}, \vec{p_{c_p}})$$

Où d est la distance euclidienne usuelle normée dans l'espace des propriétés. Plus la valeur de cette fonction est proche de 1, plus le concept c_f est *représentatif / typique* du concept c_p d'un point de vue purement intensionnel.

6.2.4 Composante expressionnelle

La composante expressionnelle de *spg* mesure la représentativité d'un concept c_f par rapport à son concept père c_p en comparant leurs expressions. Une première mesure de l'expression d'un concept est définie par le nombre de termes qui le désignent : plus le nombre de termes désignant un concept est grand, plus ce concept occupe de place dans l'univers cognitif de l'utilisateur. Par exemple, le concept *Cheval*, qui possède de nombreux termes pour le dénoter (cheval, bourrin, canasson, dada, destrier), est plus prototypique du concept *Animal* que le concept *Raton-laveur*, qui n'a qu'un seul terme le dénotant. Cette première estimation est donnée par le rapport entre le nombre de termes dénotant c_f et le nombre maximum de termes dénotant les fils directs de c_p. Nous la définissons comme suit :

$$expression_{OVD}(c_f, c_p) = \frac{|term_c(c_f)|}{\max\limits_{c_i \leq^C c_p, \nexists c_j, c_i \leq^C c_j \leq^C c_p} (|term_c(c_i)|)}$$

Si le concept *Cheval* est celui qui possède le plus de termes (au nombre de cinq) pour le dénoter parmi l'ensemble des concepts-fils du concept, alors $expression_{OVD}(raton, animal) = 1/5 = 0,2$ et $expression_{OVD}(cheval, animal) = 5/5 = 1$.

Mais cette évaluation repose sur les termes fixés dans l'OVD, elle est donc la même pour tous les utilisateurs. Si un individu fournit un corpus révélateur de son univers cognitif, il est alors possible de l'utiliser pour personnaliser le calcul de cette composante expressionnelle selon le principe suivant. Plus les termes dénotant c_f, et ses descendants, sont présents dans ce corpus, plus c_f est exprimé dans l'univers cognitif de l'utilisateur, et donc plus il est prototypique de c_p. La prégnance d'un concept dans le corpus dépend dès lors du nombre d'occurrences des termes dénotant le concept et de ceux dénotant sa descendance, rapporté au nombre total de

termes du corpus.

Les occurrences sont de plus pondérées en fonction de la structure du document où elles apparaissent. Par exemple, une occurrence apparaissant dans un titre ou dans une liste de mots-clés a plus de poids qu'une occurrence située à l'intérieur d'un paragraphe. De même pour la nature du document : un terme apparaissant dans un document capital pour l'endogroupe, ou pour l'individu, possède une pondération plus élevée que s'il est présent dans un journal de petites annonces. Comme illustration, prenons comme documents dans le corpus : un codex, des exemplaires de la revue *Nature* et du quotidien *Métro* de poids respectifs 20, 4 et 2. Prenons ensuite les pondérations liées à la structure : *titre* = 10, *résumé* = 5 et *corps de texte* = 1. Le nombre d'occurrences pondérées du mot *cheval* présents dans un codex, à raison de sept fois dans des titres, trois fois dans des résumés et 76 fois dans les corps de texte, est de $20 * (10 * 7 + 5 * 3 + 1 * 76) = 3220$ occurrences.

Nous voulons également tenir compte dans le calcul de la prégnance du nombre de documents dans lesquels les occurrences apparaissent, car un terme qui apparaît souvent mais dans un nombre très restreint de documents doit avoir une prégnance moins élevée qu'un terme présent peu de fois dans chaque document mais de façon uniforme dans la majorité des documents du corpus. Nous définissons la fonction $pregnance_t : L_C \rightarrow [0,1]$ donnant la prégnance d'un terme comme suit :

$$pregnance_t(t) = \frac{count_{occ}(t)}{N_{occ}} * \frac{count_{doc}(t)}{N_{doc}}$$

où $count_{occ}(t)$ est le nombre d'occurrences pondérées de t dans les documents du corpus, $count_{doc}(t)$ est le nombre de documents du corpus où t apparaît, N_{occ} est la somme de tous les nombres d'occurrences pondérées de tous les termes contenus dans le corpus et N_{doc} est le nombre total de documents du corpus.

Et, la fonction $pregnance_c$ définie sur $\mathcal{C}$ et donnant la prégnance d'un concept est définie comme suit :

$$pregnance_c(c) = \sum_{t \in S_{term}(c)} pregnance_t(t)$$
$$\text{où } S_{term}(c) = \Big(\bigcup_{c_i \leq^C c} term_c(c_i) \Big) \bigcup term_c(c)$$

Nous définissons la composante expressionnelle par :

$$expression(c_f, c_p) = expression_{OVD}(c_f, c_p) \times \frac{pregnance_c(c_f)}{pregnance_c(c_p)}$$

ou bien par $expression(c_f, c_p) = expression_{OVD}(c_f, c_p)$ si aucun corpus n'est fourni par l'utilisateur.

6.2.5 Composante extensionnelle

La composante extensionnelle du *spg* mesure la typicalité d'un concept c_f par rapport à son concept père c_p en évaluant la place relative occupée par les instances

de ce concept dans l'extension de c_p. Ainsi, plus l'extension de c_f a d'importance au sein de l'extension de c_p, plus c_f est prototypique de c_p. Par exemple, un individu qui possède une douzaine de chats et un poisson rouge trouve ce félin plus prototypique du concept *Animal domestique* que son poisson rouge.

L'évaluation de cette composante nécessite que les concepts considérés possèdent des instances, définies par l'utilisateur et représentatives de son univers cognitif. Pour le calcul, toutes les instances des concepts sont prises en compte, celles de l'OVD et celles ajoutées par l'utilisateur. De plus, on utilise une forme logarithmique, pour obtenir un comportement de la composante moins linéaire (les prototypicalités des concepts ayant très peu d'instances ne sont pas trop proches de 0). La composante extensionnelle est ainsi donnée par :

$$extension(c_f, c_p) = \frac{1}{1/\left(1-\log\left(\frac{|\sigma(c_f)|}{|\sigma(c_p)|}\right)\right)}$$

6.3 Gradient de prototypicalité lexicale

6.3.1 Objectif

Le gradient de prototypicalité lexicale (Lexical Prototypicality Gradient, LPG) évalue, pour un concept donné et un terme le dénotant, la représentativité de ce terme dans l'univers cognitif de l'endogroupe pour lequel on veut adapter l'ontologie. Par exemple, au sein d'une ontologie de domaine de la chimie inorganique, le concept d'Acide chlorhydrique peut être dénoté par les termes *Acide chlorhydrique*, *HCl*, $H_3O^+Cl^-$, et *123-456-78*. Tous ces termes sont synonymes et dénotent le même concept d'acide chlorhydrique. Cependant, un chimiste a plus tendance à utiliser les termes *HCl*, et $H_3O^+Cl^-$ que les autres termes. Un consultant en risque chimique se sert davantage des termes *Acide chlorhydrique* et *123-456-78*. Pour chacun de ces individus, chaque terme possède donc une valeur de représentativité différente du même concept dans son univers cognitif.

6.3.2 Principe

Le calcul du LPG repose sur l'utilisation d'un corpus jugé représentatif par l'endogroupe en question. Le principe du calcul est que plus le rapport entre le nombre d'apparitions du terme et le nombre d'apparitions de l'ensemble des termes utilisés pour dénoter le concept est proche de 1, plus ce terme est prototypique, au sens lexical, du concept. Les occurrences des termes sont pondérées selon la place qu'ils occupent dans les documents, et selon les types de documents dans lesquels ils apparaissent. Ainsi, suivant des conventions fixées au sein de l'endogroupe, un terme présent dans un titre de document peut être pondéré de manière plus importante que s'il est présent dans un résumé ou dans le corps d'un texte. Il en est de même, et suivant d'autres conventions, pour les termes se trouvant dans des documents fondateurs ou de grande importance pour l'endogroupe considéré (textes réglementaires

par exemple) par rapport à des termes se trouvant dans des documents d'usage plus courant (articles de journaux par exemple). Ces pondérations doivent être fixées par l'endogroupe ou par l'ingénieur des connaissances en fonction des priorités fixées par l'endogroupe.

6.3.3 Définition

Le gradient de prototypicalité lexicale $lpg : L_C \times \mathcal{C} \rightarrow [0,1]$, est défini pour tout couple (t,c), où t dénote le concept c, par :

$$lpg(t,c) = \frac{1}{1-\log\left(\frac{pregnance_t(t)}{\sum_{m \in term_c(c)} pregnance_t(m)}\right)}$$

La prégnance $pregnance_t : L_C \rightarrow [0,1]$ d'un terme t dans un corpus est mesurée par :

$$pregnance_t(t) = \frac{count_{occ}(t)}{N_{occ}} * \frac{count_{doc}(t)}{N_{doc}}$$

où $count_{occ}(l)$ est le nombre pondéré d'occurrences de t dans les documents du corpus, $count_{doc}(t)$ le nombre de documents où t apparaît, N_{occ} la somme de toutes les occurrences pondérées de l'ensemble des termes du corpus et N_{doc} le nombre de documents du corpus.

6.4 Gradient de prototypicalité extensionnelle

6.4.1 Objectif

Le gradient de prototypicalité extensionnelle (Instance Prototypicality Gradient, IPG) évalue, pour un concept donné et une de ses instances, la représentativité de cette instance dans l'univers cognitif du groupe d'utilisateurs pour lequel on souhaite adapter l'ontologie. Comme pour les termes, les instances de concept peuvent être plus ou moins représentatives d'un concept donné. Par exemple, tous les chats sont instances du concept *Chat*, mais une personne qui élève un chat considère son chat comme étant plus représentatif de ce concept que les autres chats (principe de familiarité [3]).

6.4.2 Principe

Le calcul du gradient de prototypicalité extensionnelle peut être réalisé en se fondant uniquement sur un corpus, comme pour le LPG. La fonction $pregnance_i$, similaire à $pregnance_t$, évalue la prégnance d'une instance i par le ratio entre le nombre d'occurrences pondérées d'un terme dénotant l'instance i et le nombre d'occurrences pondérées de tous les termes utilisés dans le corpus, en tenant compte de la nature et de la structure des documents.

3. En psychologie sociale, selon [Berscheid 1998], ce qui est familier est connu, prédictible, fixé, peut-être semblable à nous-mêmes et lié à notre univers cognitif de notre endogroupe.

6.4.3 Définition

Le gradient de prototypicalité extensionnelle $ipg : \mathcal{I} \times \mathcal{C} \rightarrow [0,1]$, est défini pour tout couple (i, c) où i est une instance du concept c par :

$$ipg(i,c) = \frac{1}{1 - \log\left(\frac{pregnance_i(i)}{\sum_{j \in term_i(\sigma(c))} pregnance_i(j)}\right)}$$

La prégnance $pregnance_i(t) : L_I \rightarrow [0,1]$ d'un terme t dans un corpus est définie comme suit :

$$pregnance_i(t) = \frac{count_{occ}(t)}{N_{occ}} * \frac{count_{doc}(t)}{N_{doc}}$$

où $count_{occ}(t)$ est le nombre pondéré d'occurrences de t dans les documents du corpus, $count_{doc}(t)$ le nombre de documents où t apparaît, N_{occ} la somme de toutes les occurrences pondérées de l'ensemble des termes du corpus et N_{doc} le nombre de documents du corpus.

6.5 Exemples

Nous illustrons le calcul des différents gradients sur une hiérarchie de cinq concepts (*cf.* figure 6.1) :

- *Matière première*, dénoté par les termes *matière première* et *ressource naturelle* ;
- *Bois*, dénoté par les termes *bois* et *tissu végétal*, avec comme propriétés *est isolant* et *est résistant* et *est durable* ;
- *Sapin*, dénoté par les termes *sapin* et *pin*, avec comme propriétés *produit de la résine* ;
- *Chêne*, dénoté par le termes *chêne*, avec comme propriété *peut servir à la fabrication de tonneaux.*

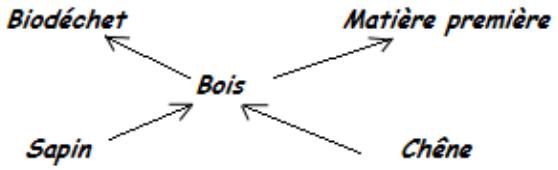


FIGURE 6.1 – Hiérarchie de concepts.

Les gradients sont calculés sur un corpus fictif qui comporte 85 236 documents et 1 267 948 termes.

6.5.1 Composante intensionnelle du spg

Le calcul de la composante intensionnelle revient à évaluer une distance entre un concept et le prototype d'un autre concept. La première étape consiste donc

à calculer les coordonnées de chaque prototype. Pour ce faire, nous nous référons aux coordonnées de chaque concept dans l'espace des propriétés (*cf.* tableau 6.1), coordonnées reflétant l'importance de chaque propriété dans la définition du concept pour un utilisateur donné.

Prop. / cpt	*biodéchet*	*matière première*	*bois*	*sapin*	*chêne*
est biodégradable	0,9	0	0,6	0,6	0,1
source énergie	0	0,8	0,9	0,7	0,4
produits finis	0	0,6	0,8	0,7	0,9
isolant	0	0	0,9	0,9	0,5
résistant	0	0	0,7	0,3	0,8
durable	0	0	0,5	0,2	0,9
produit résine	0	0	0	0,7	0
sert tonneaux	0	0	0	0	0,8

TABLE 6.1 – Liste des propriétés et pondérations.

Chaque concept hérite des propriétés de ses sur-concepts auxquelles sont ajoutées ses propres propriétés. Un poids est associé à chaque propriété pour chaque concept. Par exemple, la pondération de la propriété *est biodégradable* est de 0,9 pour le concept *biodéchet* (c'est une propriété fondamentale pour la définition de ce concept), et de 0,1 pour le concept *Chêne* (plutôt assimilé à quelque chose de durable).

À titre d'illustration, nous détaillons le calcul du vecteur prototype du concept *Biodéchet*, à partir de tous les concepts de sa descendance (*i.e.* des concepts *Bois*, *Sapin* et *Chêne*). Pour la propriété *est biodégradable*, la coordonnée du vecteur prototype est égale à la somme de :

$\frac{3-1+1}{3} * 0,9$ à partir de *Biodéchet*

et $\frac{3-2+1}{3} * 0,6$ à partir de *Bois*

et $\frac{3-3+1}{3} * 0,6 + \frac{3-3+1}{3} * 0,1$ à partir de *Sapin* et *Chêne*

Et le produit de la somme de toutes ces éléments par :
$\frac{1}{\frac{3-1+1}{3}+\frac{3-2+1}{3}+\frac{3-3+1}{3}+\frac{3-3+1}{3}}$.

La coordonnée du vecteur prototype suivant la dimension *est biodégradable* est de 0.66. A la fin du processus, nous obtenons le vecteur prototype suivant (défini dans un espace vectoriel à 8 dimensions) :
$\vec{v}_{Biodechet}$= (0.66, 0.41, 0.45, 0.45, 0.35, 0.3, 0.1, 0.11).
La seconde étape consiste à évaluer la distance entre un concept et ce vecteur prototype dans cet espace des propriétés.
Etudions maintenant la prototypicalité d'un point de vue intensionnel des concepts

Sapin et *Chêne* par rapport au concept *Bois*. Cela nécessite le calcul de la composante intensionnelle du *spg* entre le concept *Bois* et chacun de ses sous-concepts. Cette composante évalue la distance entre chacun de ces sous-concepts et le prototype du concept étudié. Le degré de prototypicalité du concept *Sapin* par rapport au concept *Bois* est égale à $1 - d(\vec{v_{Sapin}}, \vec{p_{Bois}}) = 0,25$. Le degré de prototypicalité du concept *Chêne* par rapport au concept *Bois* est égale à $1 - d(\vec{v_{Chene}}, \vec{p_{Bois}}) = 0,05$. D'un point de vue intensionnel, et dans le cas de notre OPVD, le *Bois* le plus prototypique est le *Sapin* - ses caractéristiques se rapprochent beaucoup plus du prototype du bois que le Chêne.

Ainsi, d'un point de vue intensionnel, le gradient de prototypicalité conceptuel permet de distinguer le bois le plus prototypique.

6.5.2 Composante expressionnelle du spg

Le concept *Bois* est dénoté par deux termes : *bois* et *tissu végétal*. Calculons dans un premier temps la prégnance de l'un de ses termes [4]. Dans le cas du terme *bois*, $pregnance_t(bois) = \frac{count_{occ}(bois)}{N_{occ}} * \frac{count_{doc}(bois)}{N_{doc}} = \frac{98567}{1267948} * \frac{420}{85236} = 0,00038$. La comparaison des prégnances des différents termes (*cf.* tableau 6.2) montre qu'un terme présent beaucoup de fois mais dans peu de documents (cas du terme *bois*) possède une prégnance moins élevée qu'un terme présent beaucoup plus uniformément dans l'ensemble du corpus (cas du terme *chêne*).

terme	$count_{occ}(terme)$	$count_{doc}(terme)$	$pregnance_t(terme)$
biodéchet	1750	920	0,00001
matière première	1783	810	0,00001
bois	98567	420	0,00038
sapin	7643	2190	0,00015
chêne	86200	84132	0,06710
résidus	78900	1220	0,00089
ressource naturelle	3210	536	0,00002
tissus végétal	967	394	0,00000
pin	29746	12700	0,00350

TABLE 6.2 – Liste des termes

Le concept *Bois* est dénoté directement par deux termes (*bois* et *tissus végétal*) et indirectement par l'ensemble des termes dénotant sa descendance (concepts *Sapin* et *Chêne*). La prégnance du concept *Bois* est égale à la somme des prégnances des termes le dénotant directement et indirectement. De manière formelle :
$pregnance_c(Bois) = pregnance_t(bois) + pregnance_t(tissusvegetal) + pregnance_t(pin) + pregnance_t(sapin) + pregnance_t(chene)$
Soit $pregnance_c(Bois) = 0,07$.

4. Le nombre d'occurrences donné dans le tableau 6.2 tient compte de la pondération liée à la structure et à la nature des documents

Etudions maintenant la prototypicalité du point de vue expressionnel des concepts *Sapin* et *Chêne* par rapport au concept *Bois*. Cela nécessite le calcul de la composante expressionnelle du *spg* entre le concept *Bois* et chacun de ses sous-concepts. Cette composante évalue la quantité d'information fournie par chacun des sous-concepts au concept étudié au travers du rapport de leur prégnance respective, en tenant compte du nombre de termes utilisés pour les dénoter ($expression_{OVD}(c_f, c_p)$). Le degré de prototypicalité du concept *Sapin* (deux termes pour le dénoter) par rapport au concept *Bois* est égale à $\frac{2}{2} * \frac{0,003}{0,07} = 0,05$. Le degré de prototypicalité du concept *Chêne* (un seul terme pour le dénoter) par rapport au concept *Bois* est égale à $\frac{1}{2} * \frac{0,067}{0,07} = 0,47$. D'un point de vue expressionnel, et pour cette OPVD, le *Bois* le plus prototypique est le *Chêne* - il est beaucoup plus présent dans notre univers au travers des termes le dénotant dans le corpus que le concept *Sapin*.

Ainsi, d'un point de vue expressionnel, le gradient de prototypicalité conceptuel permet de distinguer le bois le plus prototypique.

6.5.3 Composante extensionnelle du spg

D'après le tableau 6.3, le concept *Bois* possède 612 instances, en partie issues de son concept-fils *Sapin* (au nombre de 51) et de son concept-fils *Chêne* (au nombre de 465). Le degré de prototypicalité du concept *Sapin* par rapport au concept *Bois* est égale à $\frac{1}{1-\log(51/612)} = 0,22$. Le degré de prototypicalité du concept *Chêne* par rapport au concept *Bois* est égale à $\frac{1}{1-\log(465/612)} = 0,72$. D'un point de vue extensionnel, et pour cette OPVD, le *Bois* le plus prototypique est le *Chêne* - il est beaucoup plus présent dans notre environnement au travers de ses instances que le concept *Sapin*.

concept	$\lvert\sigma(concept)\rvert$
Bois	*612*
Sapin	*51*
Chêne	*465*

TABLE 6.3 – Nombre d'instances par concept

Ainsi, d'un point de vue extensionnel, le gradient de prototypicalité conceptuel permet de distinguer le bois le plus prototypique.

6.5.4 Valeur du spg

Etudions maintenant, d'un point de vue global, les prototypicalités des concepts *Sapin* et *Chêne* par rapport au concept *Bois*. Cela nécessite une évaluation dans quatre contextes différents, définis par quatre triplets (pondérations de chaque composante) représentant quatre coordonnées cognitives distinctes :

1. contexte 1 : $(\alpha = 0,33,\ \beta = 0,33,\ \gamma = 0,33)$;
2. contexte 2 : $(\alpha = 0,75,\ \beta = 0,125,\ \gamma = 0,125)$;

3. contexte 3 : $(\alpha = 0,125,\ \beta = 0,75,\ \gamma = 0,125)$;
4. contexte 4 : $(\alpha = 0,125,\ \beta = 0,125,\ \gamma = 0,75)$;

Pour chacun de ces contextes, le tableau 6.4 donne les valeurs du *spg*, et la figure 6.2 nous montre l'écart entre les valeurs de prototypicalités obtenues. Premièrement, quelques soient les coordonnées cognitives, le concept *Chêne* est plus prototypique du concept *Bois* que le concept *Sapin*. Maintenant, si nous regardons l'écart entre les valeurs pour un même contexte, nous constatons une nette différence pour un contexte ayant une composante expressionnelle dominante (contexte #3) et au contraire un rapprochement assez sensible pour un contexte ayant une composante intensionnelle dominante (contexte #2). Le cas d'un contexte orienté extensionnel (contexte #4) se rapprochant assez du cas moyen (contexte #1).

Contexte	*Bois-Sapin*	*Bois-Chêne*
#1	**0,171**	**0,563**
#2	**0,221**	**0,243**
#3	**0,097**	**0,803**
#4	**0,201**	**0,661**

TABLE 6.4 – Prototypicalités des concepts *Sapin* et *Chêne* par rapport au concept *Bois*.

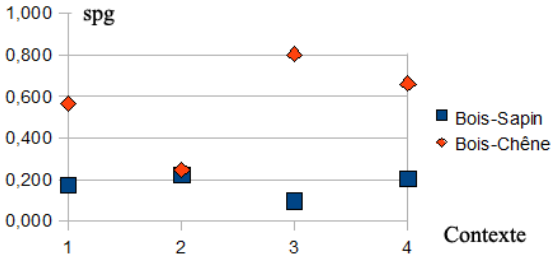


FIGURE 6.2 – Prototypicalités des concepts *Sapin* et *Chêne* par rapport au concept *Bois*.

Ainsi, quelque soient les contextes et d'un point de vue qualitatif, le *spg* permet de distinguer le bois le plus prototypique. Mais d'un point de vue quantitatif, certains contextes vont distinguer des prototypicalités plus ou moins fortes.

6.5.5 Gradient de prototypicalité lexicale

Le concept *Matière première* est dénoté dans le corpus au moyen des termes *Matière première* et *Ressource naturelle* (*cf.* tableau 6.5). La somme des prégnances des termes le dénotant est égale à 0,00003. Le calcul du gradient de prototypicalité lexicale nous donne pour valeur 0,53 pour le terme *Ressource naturelle*, et 0,47 pour le terme *Matière première*. Nous pouvons dès lors conclure que, sur la base de ce

corpus, le terme *Ressource naturelle* est nettement plus prototypique pour dénoter le concept *Matière première* que le terme *Matière première*.

t	$count_{occ}(t)$	$count_{doc}(t)$	$pregnance_t(t)$	$lpg(t,c)$
Matière première	1783	810	0,00001	0,46909
Ressource naturelle	3210	536	0,00002	0,53213

TABLE 6.5 – Termes dénotant le concept Matières premières.

Ainsi, d'un point de vue lexical, le gradient de prototypicalité lexical permet de distinguer le terme dénotant le mieux le concept *Matière première*.

6.5.6 Gradient de prototypicalité extensionnelle

Le concept de *Sapin* possède des instances exprimées dans le corpus au moyen des termes *Pin argenté* et *Pin maritime* (*cf.* tableau 6.6). La somme des prégnances des termes dénotant directement ses instances est égale à 0,00176. Le calcul du gradient de prototypicalité extensionnelle nous donne pour valeur 0,32 pour le terme *Pin argenté*, et 0,78 pour le terme *Pin maritime*. Nous pouvons dès lors conclure que, sur la base de ce corpus, l'instance *Pin maritime* est nettement plus prototypique du concept *Sapin* que l'instance *Pin argenté*.

t	$count_{occ}(t)$	$count_{doc}(t)$	$pregnance_i(t)$	$ipg(t,c)$
Pin argenté	6794	6260	0,00039	0,31625
Pin maritime	22952	6440	0,00137	0,73269

TABLE 6.6 – Instances du concept Sapin.

Ainsi, d'un point de vue extensionnel, notre gradient de prototypicalité extensionnelle nous permet de distinguer l'instance la plus prototypique du concept *Sapin*.

6.6 Paramètre émotionnel

Des travaux en psychologie cognitive (*e.g.* [Mikulinger 1990a, Mikulinger 1990b]) ont montré que l'état émotionnel d'une personne influe sur sa perception des catégories d'objets. Bien au delà de la simple valence de l'état émotionnel (positive comme la joie, ou négative comme la colère), il semble que ce soit le degré d'excitation ("arousal") de cet état émotionnel qui conditionne l'accès à l'information stockée. Plus ce degré est élevé, et plus l'accès à l'information doit non seulement être rapide mais également pertinente pour apporter une réponse efficace. Ainsi face à un danger réel, vital et immédiat, il est impératif pour la survie d'accéder à l'information la plus utile pour déterminer la marche à suivre - c'est de l'ordre du réflexe. Dualement, face à un doute quant à la nature d'un danger, il est possible de prendre le temps d'examiner différentes stratégies et fouiller plus en amont dans la mémoire.

Nous introduisons donc un facteur émotionnel qui permet de moduler les gradients

eux-mêmes en fonction de l'état d'esprit de l'utilisateur. Ce facteur émotionnel est modélisé par un coefficient δ qui peut varier entre 0 et 1 pour un "arousal" faible, et entre 1 et ∞ pour un "arousal" fort. Les valeurs des trois gradients sont élevées à la puissance δ, ce qui a pour effet, en cas d' "arousal" fort de l'utilisateur, de réduire fortement les prototypicalités qui sont déjà faibles, et inversement, d'augmenter les prototypicalités faibles.

L'extension de requête en recherche d'information constitue un exemple d'utilisation de ce facteur émotionnel : il est possible d'étendre plus ou moins des requêtes en ajoutant aux concepts apparaissant dans la requête initiale les concepts qui en sont les plus prototypiques. Cette extension tente d'occorder plus finement la requête avec l'attente de l'utilisateur en terme de pertinence de réponse à un problème. Avec un coefficient δ positionné à une valeur proche de 0, l'utilisateur est dans un état émotionnel qui lui offre une plus grande disponibilité, l'extension de requête est très large avec un accès à des concepts ayant un degré de prototypicalité initialement faible à l'état neutre ($\delta = 1$). A l'inverse, pour un coefficient δ positionné à une valeur élevée, le nombre de concepts ajoutés à la requête se restreint pour ne pouvoir accéder qu'aux concepts les plus prototypiques.

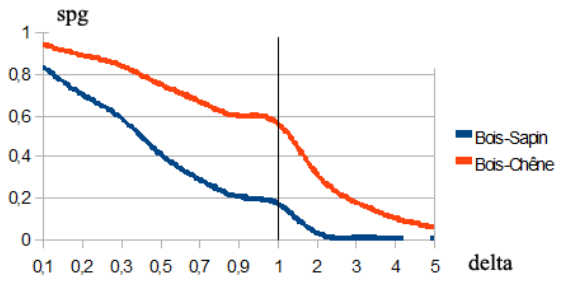


FIGURE 6.3 – Prototypicalités des concepts *Sapin* et *Chêne* par rapport au concept *Bois* en fonction de la valeur de l'arousal.

A titre d'exemple, la figure 6.3 reprend les prototypicalités des concepts *Sapin* et *Chêne* par rapport au concept *Bois* calculées précédemment. Le sapin qui avait une valeur de prototypicalité de 0,17, à l'état neutre, a une valeur de prototypicalité avoisinant les 0,7 pour un "arousal" faible. A l'inverse, pour un "arousal" fort, il atteint très rapidement des valeurs assez proches de zéro et deviendrait difficilement accessible en terme d'extension de requêtes.

6.7 Conclusion

La norme française ISO 704[5] définit les concepts comme des représentations mentales d'objets dans un contexte ou un domaine spécialisé. En tenant compte de

5. Cette norme établit et harmonise les principes fondamentaux et les méthodes pour préparer et compiler des terminologies, qu'il s'agisse d'activités menées dans le cadre de la normalisation

cette définition, et en s'inspirant de travaux en psychologie cognitive sur le processus de catégorisation, nous pouvons dès lors considérer que pour chaque individu (1) tous les termes dénotant un concept n'ont pas la même représentativité, (2) de même pour les instances rattachées à un concept, mais aussi (3) pour tous les sous-concepts et pour tous les sur-concepts d'un concept donné. Nous avons cherché à modéliser cette différence de représentativité par l'intermédiaire des gradients de prototypicalité, respectivement lexicale, intensionnelle et conceptuelle.

ou non. Elle décrit les liens existant entre les objets, les concepts, et leurs représentations par des terminologies. Elle fixe également des principes généraux régissant la formation des désignations et la formulation des définitions.

Chapitre 7

Mesures sémiotiques de similarité et de proximité

7.1 Introduction

Dans de nombreux travaux, les termes *proximité* et *similarité* ne sont pas distingués. Nous mêmes avions proposé une première version de la mesure de similarité SemioSem [Aimé 2009b, Aimé 2009d]. Cependant, dans certains cas, cette distinction est importante.

Pour illustration, prenons une tasse et du café. Une tasse est, par définition, un petit récipient de forme ovoïde, cylindrique ou demi-sphérique, généralement muni d'une anse, et qui sert à boire. Le café est, également par définition, une infusion préparée avec des fèves de caféier torréfiées et moulues. Ces deux concepts n'ont pas du tout les mêmes propriétés (la même intension), l'un est un contenant, l'autre un contenu. Par abus de langage, les dénominations peuvent se mélanger ; on emploie souvent l'expression "prendre un café" dans le sens de "prendre une tasse de café". Les instances de ces deux concepts sont en revanche régulièrement associées. Ainsi, ces deux concepts sont très proches, mais en aucun cas similaires.

Prenons maintenant cette même tasse et un bol. Un bol est, par définition, un récipient hémisphérique dans lequel on sert certaines boissons. Ces deux concepts sont donc par définition deux récipients destinés à la consommation de boisson, ils partagent la même intension. Le terme de bol ou de tasse est également utilisé pour désigner un volume (un bol de cidre, une tasse de café). Ces deux concepts sont certes proches, mais ils sont surtout similaires.

D'un point de vue étymologique, *proximité* vient du latin *proximus* qui signifie *très proche*, c'est-à-dire situé dans un espace qui se trouve à faible distance du point (de vue) considéré. *Similarité* vient du latin *similis* qui signifie *semblable*. En psychologie cognitive, il a été montré que la perception visuelle s'effectue par l'application de principes d'organisation dont ceux qui ont été mis en évidence par la théorie Gestaltiste [Koffka 1935]. Cette psychologie de la perception de la forme est apparue en

Allemagne dans les années 1920. Elle repose sur les travaux de ses trois fondateurs : K. Koffka, W. Köhler et M. Wertheimer. Pour les théoriciens de la Gestalt, les êtres vivants ne perçoivent pas une suite de sensations, mais une configuration globale. Cette vision vient contredire celle des structuralistes, pour qui les différentes perceptions sont décomposées en sensations primaires, sensations primaires que nous analysons élément par élément. Ainsi, le tout est considéré comme différent de la somme des parties. De cette idée découle la première des six lois fondamentales de la Gestalt Theory, dite *loi de la bonne forme* : un ensemble d'éléments tend à être perçu d'abord comme une forme organisée, simple et stable. Cinq autres lois sont déduites de cette première : loi de *bonne continuité*, loi de *destin commun*, loi de *clôture*, loi de *proximité* et loi de *similarité*. Nous nous intéressons ici plus particulièrement à ces deux dernières :

- la loi de **proximité**, qui permet au cerveau de regrouper des éléments qui apparaissent souvent ensemble, qui sont proches dans une même zone perceptive. C'est le cas des lettres qui forment un mot, des mots qui forment un syntagme. Il s'agit d'un regroupement présentant une certaine cohérence, et la plupart du temps inconscient.
- la loi de **similarité**, qui nous permet de regrouper les éléments qui paraissent semblables. Il peut s'agir de similitudes descriptives (perceptibles) ou fonctionnelles.

Ainsi, "similarité" et "proximité" sont deux concept proches mais pas similaires. C'est pourquoi nous avons développé deux mesures sémiotiques distinctes : une mesure de similarité, SemioSem, et une mesure de proximité, ProxSem. La première se fonde uniquement sur les caractéristiques des concepts (propriétés, termes, instances) indépendamment de la structure de l'ontologie et du corpus de textes. La seconde se fonde sur la représentation simultanée des deux concepts (relations, présence de termes dans un même document, présence simultanée des instances).

7.2 SemioSem, une mesure de similarité sémiotique

7.2.1 Principe

SemioSem s'inspire des critères définis par la *"loi" de similarité*, issue de la *Gestalt Theory*, qui établit que des concepts partageant des propriétés (fonctionnelles ou perceptives) sont considérés comme appartenant à un même groupe. De cette manière, la loi de similarité permet de regrouper les éléments qui paraissent semblables.

Le cerveau humain ayant la faculté de regrouper les éléments qui se ressemblent (suivant leur forme, leur taille, leur couleur, etc.), il est possible de rendre opérationnelle cette loi pour - par exemple - permettre une navigation plus intuitive au sein d'une interface en regroupant les pictogrammes de même couleur et de même taille, ou en appliquant un même style de mise en page sur des blocs traitant d'un même sujet. Par exemple, les différents symboles de la figure 7.1 semblent être da-

vantage disposés en colonnes qu'en lignes : nous regroupons intuitivement les ronds d'un côté et les carrés de l'autre.

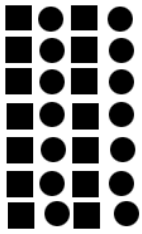


Figure 7.1 – Ensemble de symboles illustrant la loi de similarité.

En s'inspirant de cette approche, SemioSem évalue suivant les trois dimensions d'une conceptualisation (intension, extension et expression) la similarité entre deux concepts. Ainsi deux concepts sont similaires si :

- d'un point de vue intensionnel, la distance entre leurs prototypes est faible ;
- d'un point de vue expressionnel, ils partagent un grand nombre de termes pour les dénoter ;
- d'un point de vue extensionnel, ils possèdent un grand nombre d'instances en commun.

Comme pour le calcul des gradients de prototypicalité [Aimé 2009a, Aimé 2010d], le modèle que nous proposons nécessite une pondération de chaque composante de SemioSem pour prendre en compte l'importance des aspects intensionnel, extensionnel et/ou expressionnel dans la conceptualisation du domaine, au moyen des coordonnées cognitives α, β et γ définies dans le chapitre 5.

Cette mesure présente l'intérêt d'être indépendante de la structure de l'ontologie (arbre ou treillis, profondeur...), mais dépendante de l'univers cognitif de l'utilisateur au travers non seulement des coordonnées cognitives α, β et γ, mais aussi de l'univers de l'utilisateur par ses instances et son dialecte.

7.2.2 Définition

La mesure de similarité SemioSem : $\mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ est définie par :

$$SemioSem(c_1, c_2) = \alpha * intens(c_1, c_2) + \beta * extens(c_1, c_2) + \gamma * express(c_1, c_2)$$

Les fonctions *intens*, *extens* et *express* sont respectivement détaillées dans les sections 7.2.3, 7.2.4 et 7.2.5.

7.2.3 Composante intensionnelle

Les vecteurs prototypes des concepts c_1 et c_2 sont calculés suivant la méthode proposée dans le chapitre précédent pour le calcul de la composante intensionnelle

du *spg*.

D'un point de vue *intensionnel*, plus les prototypes respectifs de c_1 et c_2 sont proches, *i.e.* plus leurs propriétés sont proches, plus ces concepts sont similaires. La composante intensionnelle $intens : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est donc dépendante de la distance entre les vecteurs prototypes des deux concepts. Cette fonction est définie par :

$$intens(c_1, c_2) = 1 - dist(\vec{p_{c_1}}, \vec{p_{c_2}})$$

Nous retrouvons ici les critères définis dans la loi de similarité de la *Gestalt Theory* : plus deux concepts partagent de propriétés descriptives et fonctionnelles, plus ils sont similaires. Nous apportons cependant une petite différence en tenant compte du point de vue de l'utilisateur. Supposons que nous nous intéressions à la similarité entre les concepts *Avion*, *Bateau* et *Canard*. *Avion* et *Bateau* sont plus similaires que *Bateau* et *Canard*. Dans le premier cas, les deux concepts partagent la propriété *transporte des objets ou des personnes*, laquelle est une propriété importante pour ces concepts, alors que dans le second cas les deux concepts partagent la propriété *peut flotter* qui n'est pas importante dans la définition du concept *Canard*.

7.2.4 Composante extensionnelle

La prise en compte du point de vue *extensionnel* s'appuie sur la mesure de similarité de [Dice 1945], qui est définie par le ratio entre le nombre d'instances communes et la moyenne du nombre d'instances des deux concepts. Ainsi, deux concepts sont similaires s'ils possèdent un grand nombre d'instances en commun et très peu d'instances distinctes. Cette mesure possède le même ordre et se situe dans le même intervalle $[0, 1]$ que la mesure de similarité de [Jaccard 1901]. Elle offre de plus une plus grande régularité.

La composante extensionnelle $extens : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie par :

$$extens(c_1, c_2) = \frac{|\sigma(c_1) \cap \sigma(c_2)|}{Moyenne(|\sigma(c_1)|, |\sigma(c_2)|)}$$

Considérons deux concepts : *Animal de guerre* (ou de combat) et *Chien*. Le premier possède comme instances un éléphant, un chien anti-char et un chien d'attaque. Si un utilisateur a un point de vue plutôt belliqueux, son concept *Chien* peut avoir comme instances un chien de chasse, un chien anti-char et un chien d'attaque. La similarité entre le concept *Animal de guerre* et le concept *Chien* est alors de 0.66 (deux concepts communs sur une moyenne de trois concepts). S'il a une vision plus pacifique du monde canin, il a comme instance un chien de sauvetage, un chien d'attelage, un chien guide d'aveugle, un chien de berger et à la rigueur un chien d'attaque. La similarité entre le concept *Animal de guerre* et le concept *Chien* est de 0.25 (un seul concept commun sur une moyenne de quatre concepts). Cette similarité, d'un point de vue expressionnel, dépend ainsi des instances propres à l'utilisateur (ou à l'endogroupe) considéré.

7.2.5 Composante expressionnelle

La prise en compte de la composante expressionnelle est assez similaire à la composante extensionnelle. La fonction *express* est définie par le ratio entre le nombre de termes communs et le nombre total de termes dénotant les deux concepts. Ainsi, deux concepts sont similaires s'ils possèdent un grand nombre de termes communs les dénotant et très peu de termes propres à chacun. Des comparaisons de deux types peuvent être effectuées entre les deux ensembles de termes pour déterminer leur intersection :

- soit une comparaison exacte, où il s'agit de comptabiliser les termes identiques pour les deux concepts ;
- soit une comparaison approximative, où il s'agit de comptabiliser, en plus de ces termes identiques, les termes synonymes [1] entre chaque concept au moyen d'un dictionnaire de synonymes.

La composante expressionnelle $express : \mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ est définie par :

$$express(c_1, c_2) = \frac{|term_c(c_1) \cap_{approx} term_c(c_2)|}{|term_c(c_1) \cup term_c(c_2)|}$$

où :

$$\begin{gathered} term_c(c_1) \cap_{approx} term_c(c_2) = \\ (terme_c(c_1) \cap term_c(c_2)) \\ \cup (term_c(c_1) \cap synonymes(term_c(c_2))) \\ \cup (term_c(c_2) \cap synonymes(term_c(c_1))) \end{gathered}$$

Considérons deux concepts : *Voiture* et *Véhicule utilitaire*. Le premier est dénoté par les termes suivants : voiture, véhicule, caisse, bagnole. Le second est dénoté par les termes suivants : véhicule utilitaire, carrosse, automobile. Si nous prenons une comparaison exacte, nous avons une similarité nulle entre les deux concepts d'un point de vue expressionnel ; aucun des termes présents dans le premier ensemble n'existe dans le second. Maintenant, utilisons une comparaison approximative et étudions le second ensemble (celui concernant le véhicule utilitaire) par rapport au premier en prenant terme par terme :

- *véhicule utilitaire* est quasi synonyme de *véhicule* ;
- *carrosse* ne possède aucun équivalent dans le concept *Voiture* ;
- *automobile* est au moins synonyme de *voiture*.

Etudions le premier ensemble (celui concernant la voiture) par rapport au second en prenant terme par terme :

- *véhicule* est quasi similaire à *véhicule utilitaire* ;
- *caisse* est synonyme de *automobile* ;
- *bagnole* est synonyme de *automobile* ;

1. Nous pouvons considérer, pour chaque concept et pour un utilisateur donné, uniquement les termes dont la prototypicalité lexicale est non nulle.

– *voiture* est synonyme de *automobile*.

Au total, nous obtenons une intersection comprenant six termes sur sept, soit une similarité expressionnelle de 0,86.

7.3 ProxSem, une mesure de proximité sémiotique

ProxSem s'inspire des critères définis par la *"loi" de proximité* issue de la *Gestalt Theory*, qui établit que des concepts proches dans une même scène perceptive sont considérés comme appartenant à un même groupe. La loi de proximité réunit les éléments physiquement proches pour leur affecter une même valeur sémantique.

Le cerveau humain a la faculté de considérer comme un tout les éléments qui nous paraissent proches physiquement (par exemple les lettres de chaque mot que nous lisons). A titre d'exemple, les différents symboles de la figure 7.2 semblent disposés en trois groupes distincts.

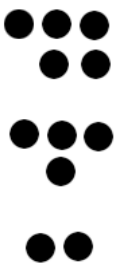


Figure 7.2 – Ensemble de symboles illustrant la loi de proximité.

En s'inspirant de cette approche, ProxSem évalue la proximité entre deux concepts suivant les trois dimensions sémiotiques (intension, extension et expression). Ainsi deux concepts sont d'autant plus proches, si :

– d'un point de vue *intensionnel*, il existe une proportion plus importante de relations (*ObjectProperty)* les reliant ;
– d'un point de vue *expressionnel*, les termes qui les dénotent sont souvent présents ensemble dans les mêmes documents ;
– d'un point de vue *extensionnel*, leurs instances sont souvent présentes ensemble dans l'univers de l'utilisateur.

Comme pour le calcul des gradients de prototypicalité, le modèle que nous proposons nécessite une pondération de chaque composante de ProxSem pour prendre en compte l'importance des aspects intensionnel, extensionnel et/ou expressionnel dans la conceptualisation du domaine, au moyen des coordonnées cognitives α, β et γ définies dans le chapitre 5.

Cette mesure présente l'intérêt d'être indépendante de la structure de l'ontologie, mais dépendante de l'univers cognitif de l'utilisateur au travers non seulement des coordonnées cognitives α, β et γ, mais aussi par ses instances propres et son dialecte (termes dont la prototypicalité lexicale est différente de zéro pour les concepts étudiés).

7.3.1 Définition

La mesure de proximité ProxSem : $\mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ est définie par :

$$ProxSem(c_1, c_2) = \alpha * intens_{prox}(c_1, c_2) + \beta * extens_{prox}(c_1, c_2) + \gamma * express_{prox}(c_1, c_2)$$

Les fonctions $intens_{prox}$, $extens_{prox}$ et $express_{prox}$ sont respectivement détaillées dans les sections 7.3.2, 7.3.3 et 7.3.4.

7.3.2 Composante intensionnelle

Soit deux concepts c_1 et c_2, et p_1, p_{12}, p_2, p_{21} les quatre ensembles suivants :
$p_1 = \{p_k \in \mathcal{P} : c_1 \in dom(p_k)\}$, l'ensemble des relations ayant pour domaine c_1 ;
$p_2 = \{p_k \in \mathcal{P} : c_2 \in dom(p_k)\}$, l'ensemble des relations ayant pour domaine c_2 ;
$p_{12} = \{p_k \in \mathcal{P} : c_1 \in dom(p_k) \wedge c_2 \in codom(p_k)\}$, l'ensemble des relations ayant pour domaine c_1 et co-domaine c_2 ;
$p_{21} = \{p_k \in \mathcal{P} : c_2 \in dom(p_k) \wedge c_1 \in codom(p_k)\}$, l'ensemble des relations ayant pour domaine c_2 et co-domaine c_1.

D'un point de vue intensionnel, plus deux concepts possèdent de relations entre eux, plus ils sont proches. La composante intensionnelle $intens_{prox} : \mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ est définie par :

$$\text{Si } |p_1| + |p_2| = 0,\ intens_{prox}(c_1, c_2) = 0$$
$$\text{Sinon } intens_{prox}(c_1, c_2) = \frac{1}{1 - \log\left(\frac{|p_{12}| + |p_{21}|}{|p_1| + |p_2|}\right)}$$

7.3.3 Composante expressionnelle

D'un point de vue expressionnel, plus deux concepts ont des termes les dénotant présents ensemble dans les mêmes documents, plus ils sont proches. La composante expressionnelle $express_{prox} : \mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ est définie par :

$$express_{prox}(c_1, c_2) = \frac{nbDocPond_t(c_1, c_2)}{nbDoc_t(c_1, c_2)}$$

Où :

- $nbDocPond_t(c_1, c_2)$ est le nombre pondéré de documents où les termes dénotant les concepts c_1 et c_2 sont présents ensemble ;
- $nbDoc_t(c_1, c_2)$ est le nombre pondéré de documents où au moins un des termes dénotant les concepts c_1 et c_2 est présent.

Une version simplifiée consisterait à calculer le ratio entre le nombre de documents où les termes dénotant les concepts c_1 et c_2 sont présents ensemble, et le nombre de documents où au moins un des termes dénotant les concepts c_1 et c_2 est présent.

Soit M_{term}, un matrice définie dans $[0,1]^{t \times d}$, où :
- chaque ligne i correspond à un terme t_i dénotant un concept ;
- chaque colonne j correspond à un document d_j où est présent - au moins - un des termes dénotant le concept c_1 ou le concept c_2 ;
- $t = |term_c(c_1)| + |term_c(c_2)|$
- $d = |d_j : \exists t \in ((term_c(c_1) \cup term_c(c_2)), t \in d_j|$
- chaque élément $M_t(i,j) \in [0,1]$ est égale à la pondération *tf-idf* pour le terme t_i dans le document d_j.

Soit $\zeta_t(d_j)$, une pondération de pertinence du document d_j pour les concepts c_1 et c_2, égale à la somme de toutes les pondérations *tf-idf* pour les termes dénotant ces concepts dans le document d_j. Cette fonction est définie par :

$$\zeta_t(d_j) = \sum_{i=1}^{t} M_{term}(i,j) = \sum_{i=1}^{t} \left(\frac{n_{i,j}}{\sum_k n_{k,j}} * log \frac{|D|}{|\{d_k : t_i \in d_k\}|} \right),$$

Où :
- $n_{i,j}$ est le nombre d'occurrences du terme t_i dans le document d_j ;
- $\sum_k n_{k,j}$ est le nombre d'occurrences de tous les termes dans le document d_j ;
- $|D|$ est le nombre de documents du corpus ;
- $|\{d_k : t_i \in d_k\}|$ est le nombre de documents où le terme t_i apparaît.

La fonction $nbDocPond_t(c_1, c_2)$ est définie par :

$$nbDocPond_t(c_1, c_2) = \sum \zeta_t(d_k)$$

Où $d_k \in \{d_z : \exists t_1 \in term_c(c_1), \exists t_2 \in term_c(c_2), t_1 \in d_z \wedge t_2 \in d_z\}$.

La fonction $nbDoc_t(c_1, c_2)$ est définie formellement par :

$$nbDoc_t(c_1, c_2) = \sum \zeta_t(d_k)$$

Où $d_k \in \{d_z : \exists t_1 \in term_c(c_1), \exists t_2 \in term_c(c_2), t_1 \in d_z \vee t_2 \in d_z\}$.

7.3.4 Composante extensionnelle

D'un point de vue extensionnel, plus deux concepts ont des termes dénotant leurs concepts présents ensemble dans les mêmes documents, plus ils sont proches. La composante extensionnelle $extens_{prox} : \mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ est définie par :

$$extens_{prox}(c_1, c_2) = \frac{nbDocPond_i(c_1, c_2)}{nbDoc_i(c_1, c_2)}$$

Où :

- $nbDocPond_i(c_1, c_2)$ est le nombre pondéré de documents où les termes dénotant les instances des concept c_1 et c_2 sont présents ensemble ;
- $nbDoc_i(c_1, c_2)$ est le nombre pondéré de documents où au moins un des termes dénotant les instances des concepts c_1 et c_2 est présent.

Une version simplifiée consisterait à calculer le ratio entre le nombre de documents où les termes dénotant les instances des concepts c_1 et c_2 sont présents ensemble, et le nombre de documents où au moins un des termes dénotant les instances des concepts c_1 et c_2 est présent.

Soit M_{inst}, un matrice définie dans $[0,1]^{t \times d}$, où :
- chaque ligne i correspond à un terme t_i dénotant une instance d'un concept ;
- chaque colonne j correspond à un document d_j où est présent - au moins - un des termes dénotant une instance du concept c_1 ou du concept c_2 ;
- $t = |term_i(c_1)| + |term_i(c_2)|$
- $d = |d_j : \exists t \in (term_i(c_1) \cup term_i(c_2)), t \in d_j|$
- chaque élément $M_{inst}(i,j) \in [0,1]$ est égale à la pondération *tf-idf* pour le terme t_i dans le document d_j.

Soit $\zeta_i(d_j)$, une pondération de pertinence du document d_j pour les concepts c_1 et c_2, égale à la somme de toutes les pondérations *tf-idf* pour les termes dénotant les instances de ces concepts dans le document d_j. Cette fonction est définie formellement par :

$$\zeta_i(d_j) = \sum_{i=1}^{t} M_{inst}(i,j) = \sum_{i=1}^{t} \left(\frac{n_{i,j}}{\sum_k n_{k,j}} * log \frac{|D|}{|\{d_k : t_i \in d_k\}|} \right),$$

Où :
- $n_{i,j}$ est le nombre d'occurrences du terme t_i dans le document d_j ;
- $\sum_k n_{k,j}$ est le nombre d'occurrences de tous les termes dans le document d_j ;
- $|D|$ est le nombre de documents du corpus ;
- $|\{d_k : t_i \in d_k\}|$ est le nombre de documents où le terme t_i apparaît.

La fonction $nbDocPond_i(c_1, c_2)$ est définie formellement par :

$$nbDocPond_i(c_1, c_2) = \sum \zeta_i(d_k)$$

Où $d_k \in \{d_z : \exists t_1 \in term_i(c_1), \exists t_2 \in term_i(c_2), t_1 \in d_z \wedge t_2 \in d_z\}$.

La fonction $nbDoc_i(c_1, c_2)$ est définie formellement par :

$$nbDoc_i(c_1, c_2) = \sum \zeta_i(d_k)$$

Où $d_k \in \{d_z : \exists t_1 \in term_i(c_1), \exists t_2 \in term_i(c_2), t_1 \in d_z \vee t_2 \in d_z\}$.

7.4 Exemple de calculs des différentes mesures

Afin d'exemplifier nos différentes mesures, voici de manière détaillée un exemple de calculs effectués sur deux concepts (*cf.* figure 7.3) : *Fibre acrylique* et *Fibre antibactérienne*, notés respectivement : c_{acr} et c_{antib}.

Le concept *Fibre acrylique* :
- est dénoté par les termes *fil*, *fibre acrylique*, *acrylique*, *polyacrylonitrile* et *PAN* ;
- possède huit instances : *Acrilan, Courtelle, Crylor, Dralon, Dynel, Léacryl, Orlon* et *Amicor*.

Le concept *Fibre antibactérienne* :
- est dénoté par les termes *fibre*, *fibre antibactérienne* et *fibre antibactérienne de soins et de santé* ;
- possède cinq instances : *SilveRStat, Alibaba, ThermovylZCB, Bamboo, Nanobon* et *Amicor*.

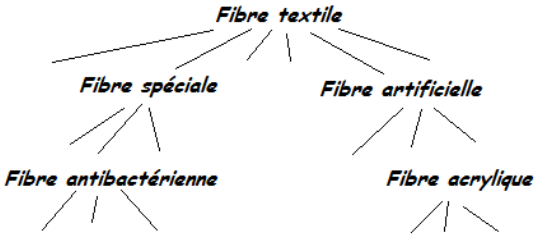


FIGURE 7.3 – Hiérarchie de concepts.

Le concept *Fibre acrylique* possède la relation *se dégrade thermiquement à*. Le concept *Fibre antibactérienne* possède la relation *est constitué de* ayant pour co-domaine *Fibre acrylique*, et la propriété *détruit les bactéries*.

7.4.1 Evaluation de la similarité

Nous calculons successivement chacune des composantes de notre mesure de similarité SEMIOSEM.

7.4.1.1 Composante intensionnelle

D'un point de vue *intensionnel*, dans un espace vectoriel à douze dimensions, les concepts *Fibre acrylique* et *Fibre antibactérienne* ont pour vecteurs prototypes respectifs :
$\vec{p_{c_{acr}}} = (0.17, 0.05, 0.06, 0.93, 0, 0, 0.67, 0.43, 0.32, 0.26, 0, 0.42)$
$\vec{p_{c_{antib}}} = (0.29, 0.09, 0.16, 0.08, 0, 0.8, 0.11, 0.09, 0, 0.85, 0.13, 0.22)$

La composante intensionnelle de SemioSem est donc égale à :

$$intens(c_{acr}, c_{antib}) = 1 - dist(\overrightarrow{p_{c_{acr}}}, \overrightarrow{p_{c_{antib}}}) = 1 - 0.27 = 0.73$$

7.4.1.2 Composante expressionnelle

Par comparaison exacte, il n'y a aucun terme en commun entre ces deux concepts. Par comparaison approximative, et d'après notre dictionnaire, les termes *fil* et *fibre* sont synonymes. Nous avons donc deux termes à placer dans l'intersection lexicale approximative. La composante expressionnelle de SemioSem est donc égale à :

$$express(c_{acr}, c_{antib}) = \frac{|terme_c(c_{acr}) \cap_{approx} terme_c(c_{antib})|}{|terme_c(c_{acr}) \cup terme_c(c_{antib})|} = \frac{2}{8} = 0.25$$

7.4.1.3 Composante extensionnelle

Comme nous pouvons le constater, sur ces deux ensembles, une seule instance est partagée par les deux concepts. La composante extensionnelle de SemioSem est donc égale à :

$$extens(c_{acr}, c_{antib}) = \frac{|\sigma(c_{acr}) \cap \sigma(c_{antib})|}{Moyenne(|\sigma(c_{acr})|, |\sigma(c_{antib})|)} = \frac{1}{6{,}5} = 0.1538$$

7.4.1.4 Formule globale

Etudions maintenant, d'un point de vue *global*, la similarité entre le concept *Fibre acrylique* et le concept *Fibre antibactérienne*. Pour ce faire, une évaluation est menée dans quatre contextes différents, définis par quatre triplets (pondérations de chaque composante) représentant quatre coordonnées cognitives distinctes :

1. contexte 1 : $(\alpha = 0,33,\ \beta = 0,33,\ \gamma = 0,33)$;
2. contexte 2 : $(\alpha = 0,75,\ \beta = 0,125,\ \gamma = 0,125)$;
3. contexte 3 : $(\alpha = 0,125,\ \beta = 0,75,\ \gamma = 0,125)$;
4. contexte 4 : $(\alpha = 0,125,\ \beta = 0,125,\ \gamma = 0,75)$;

Pour chacun de ces contextes, le tableau 7.1 donne les valeurs de SemioSem. Quelque soient les coordonnées cognitives, les concepts *Fibre acrylique* et *Fibre antibactérienne* ont une similarité de valeur inférieure ou égale à 0,5. Suivant la nature du contexte, un des aspects de la conceptualisation peut être privilégié par rapport aux autres. Ainsi, dans notre exemple, seul le contexte où nous privilégions l'aspect intensionnel nous donne une similarité un peu plus importante entre les deux concepts. Notre mesure de similarité nous permet ainsi de refléter différents points du vue, en fonction des coordonnées cognitives de l'utilisateur et de la représentation de son univers, au travers de ses instances et de ses termes propres.

7.4.2 Evaluation de la proximité

Nous calculons successivement chacune des composantes de notre mesure de proximité ProxSem. Nous supposons que le corpus contient quatre documents doc_1, doc_2, doc_3 et doc_4.

Contexte	*SemioSem*
#1	**0,373**
#2	**0,598**
#3	**0,298**
#4	**0,235**

TABLE 7.1 – Similarité entre le concept *Fibre acrylique* et le concept *Fibre antibactérienne*.

7.4.2.1 Composante intensionnelle

D'un point de vue intensionnel, les concepts *Fibre acrylique* et *Fibre antibactérienne* sont liés par une seule relation : *est constitué de*. La composante intensionnelle de PROXSEM est donc égale à :

$$intens_{prox}(c_{acr}, c_{antib}) = \frac{1}{1-\log\left(\frac{1}{1+2}\right)} = 0,677$$

7.4.2.2 Composante expressionnelle

D'un point de vue expressionnel, le tableau 7.4 donne les valeurs *tf-idf* pour chaque terme et chaque document.

terme et document	doc_1	doc_2	doc_3	doc_4
fil	0.005	0.002	0	0.001
fibre acrylique	0.002	0	0	0.009
acrylique	0.0075	0	0	0
polyacrylonitrile	0.0056	0.005	0	0
PAN	0.0025	0.006	0	0
fibre	0.003	0.003	0.003	0
fibre antibactérienne	0.002	0.001	0.021	0
fibre antibactérienne de soin et de santé	0	0.004	0	0

TABLE 7.2 – Valeurs de *tf-idf* pour les termes dénotant les concepts *Fibre acrylique* et *Fibre antibactérienne*.

Calculons le score pour chaque document :

- $\zeta_t(doc_1) = 0.005 + 0.002 + 0.0075 + 0.0056 + 0.0025 + 0.003 + 0.002 = 0.0276$
- $\zeta_t(doc_2) = 0.002 + 0.005 + 0.006 + 0.003 + 0.001 + 0.004 = 0.021$
- $\zeta_t(doc_3) = 0.003 + 0.021 = 0.024$
- $\zeta_t(doc_4) = 0.001 + 0.009 = 0.01$

La fonction $nbDocPond_t(c_{acr}, c_{antib})$ est égale à $nbDocPond_t(c_{acr}, c_{antib}) = \zeta_t(doc_1) + \zeta_t(doc_2) = 0.0276 + 0.021 = 0.0486$

La fonction $nbDoc_t(c_{acr}, c_{antib})$ est égale à $nbDoc_t(c_{acr}, c_{antib}) = \zeta_t(doc_1) + \zeta_t(doc_2) +$

$\zeta_t(doc_3) + \zeta_t(doc_4) = 0.0276 + 0.021 + 0.024 + 0.01 = 0.0826$

La composante expressionnelle de ProxSem est donc égale à :

$$express_{prox}(c_{acr}, c_{antib}) = \frac{nbDocPond_t(c_{acr},c_{antib})}{nbDoc_t(c_{acr},c_{antib})} = \frac{0.0486}{0.0826} = 0.589$$

D'un point de vue extensionnel, le tableau 7.4 donne les valeurs *tf-idf* pour chaque terme et chaque document.

terme et document	doc_1	doc_2	doc_3	doc_4
fil	0.005	0.002	0	0.001
fibre acrylique	0.002	0	0	0.009
acrylique	0.0075	0	0	0
polyacrylonitrile	0.0056	0.005	0	0
PAN	0.0025	0.006	0	0
fibre	0.003	0.003	0.003	0
fibre antibactérienne	0.002	0.001	0.021	0
fibre antibactérienne de soin et de santé	0	0.004	0	0

Table 7.3 – Valeurs de *tf-idf* pour les termes dénotant les concepts *Fibre acrylique* et *Fibre antibactérienne*.

Calculons le score pour chaque document :
- $\zeta_t(doc_1) = 0.005 + 0.002 + 0.0075 + 0.0056 + 0.0025 + 0.003 + 0.002 = 0.0276$
- $\zeta_t(doc_2) = 0.002 + 0.005 + 0.006 + 0.003 + 0.001 + 0.004 = 0.021$
- $\zeta_t(doc_3) = 0.003 + 0.021 = 0.024$
- $\zeta_t(doc_4) = 0.001 + 0.009 = 0.01$

La fonction $nbDocPond_t(c_{acr}, c_{antib})$ est égale à $nbDocPond_t(c_{acr}, c_{antib}) = \zeta_t(doc_1) + \zeta_t(doc_2) = 0.0276 + 0.021 = 0.0486$

La fonction $nbDoc_t(c_{acr}, c_{antib})$ est égale à $nbDoc_t(c_{acr}, c_{antib}) = \zeta(doc_1) + \zeta(doc_2) + \zeta(doc_3) + \zeta(doc_4) = 0.0276 + 0.021 + 0.024 + 0.01 = 0.0826$

La composante expressionnelle de ProxSem est donc égale à :

$$express_{prox}(c_{acr}, c_{antib}) = \frac{nbDocPond_t(c_{acr},c_{antib})}{nbDoc_t(c_{acr},c_{antib})} = \frac{0.0486}{0.0826} = 0.589$$

7.4.2.3 Composante extensionnelle

D'un point de vue extensionnel, le tableau 7.4 donne les valeurs *tf-idf* pour chaque terme et chaque document.
Calculons le score pour chaque document :
- $\zeta_i(doc_1) = 0.0021 + 0.005 + 0.001 + 0.003 + 0.004 + 0.0012 = 0.0163$
- $\zeta_i(doc_2) = 0.002 + 0.0012 + 0.0043 = 0.0075$
- $\zeta_i(doc_3) = 0.001$

terme et document	doc_1	doc_2	doc_3	doc_4
Amicor (acrylique)	0.0021	0.002	0	0.002
Acrilan	0.005	0.012	0	0
Courtelle	0.001	0	0	0
Orlon	0.003	0	0	0
SilverRStat	0.004	0.0043	0.001	0
Amicor (antibactérienne)	0.0012	0	0	0

TABLE 7.4 – Valeurs de *tf-idf* pour les termes dénotant les instances des concepts *Fibre acrylique* et *Fibre antibactérienne*.

- $\zeta_i(doc_4) = 0.002$

La fonction $nbDocPond_i(c_{acr}, c_{antib})$ est égale à $nbDocPond_i(c_{acr}, c_{antib}) = \zeta_i(doc_1) + \zeta_i(doc_2) = 0.0163 + 0.0075 = 0.0238$

La fonction $nbDoc_i(c_{acr}, c_{antib})$ est égale à $nbDoc_i(c_{acr}, c_{antib}) = \zeta_i(doc_1) + \zeta_i(doc_2) + \zeta_i(doc_3) + \zeta_i(doc_4) = 0.0163 + 0.0075 + 0.001 + 0.002 = 0.0268$

La composante extensionnelle de PROXSEM est donc égale à :

$$exten_{prox}(c_{acr}, c_{antib}) = \frac{nbDocPond_i(c_{acr}, c_{antib})}{nbDoc_i(c_{acr}, c_{antib})} = \frac{0.0238}{0.0268} = 0.89$$

7.4.2.4 Formule globale

Etudions maintenant, d'un point de vue *global*, la proximité entre le concept *Fibre acrylique* et le concept *Fibre antibactérienne*. Pour ce faire, nous effectuons nos calculs dans quatre contextes différents, définis par quatre triplets (pondérations de chaque composante) représentant quatre coordonnées cognitives distinctes :

1. $\alpha = 0,33$, $\beta = 0,33$, $\gamma = 0,33$;
2. $\alpha = 0,75$, $\beta = 0,125$, $\gamma = 0,125$;
3. $\alpha = 0,125$, $\beta = 0,75$, $\gamma = 0,125$;
4. $\alpha = 0,125$, $\beta = 0,125$, $\gamma = 0,75$.

Pour chacun de ces contextes, le tableau 7.5 donne les valeurs de PROXSEM. Quelques soient les coordonnées cognitives, les concepts *Fibre acrylique* et *Fibre antibactérienne* ont une proximité supérieure ou égale à 0,6. Ainsi, dans notre exemple, seul le contexte où nous privilégions l'aspect extensionnel nous donne une proximité plus importante entre les deux concepts. Notre mesure de proximité nous permet ainsi de refléter différents points du vue, en fonction des coordonnées cognitives de l'utilisateur et de la représentation de son univers, au travers de ses instances et de ses termes propres.

Contexte	*ProxSem*
#1	**0,711**
#2	**0,693**
#3	**0,638**
#4	**0,826**

TABLE 7.5 – Proximité entre le concept *Fibre acrylique* et le concept *Fibre antibactérienne*.

7.5 Conclusion

La finalité de ce chapitre était de distinguer concrètement les notions de proximité et de similarité, et de proposer des mesures formelles de ces propriétés conceptuelles.

Notre position est que similarité et proximité sont deux concepts proches mais pas similaires. La proximité est un lien unissant deux éléments qui apparaissent souvent ensembles, qui sont proches dans une même zone perceptive. La similarité est un lien unissant deux éléments qui paraissent semblables. Il peut s'agir de similitudes perceptibles ou fonctionnelles.

Nous avons par conséquent bâti deux mesures distinctes pour évaluer chacune de ces notions. Ces deux mesures s'inscrivent dans le cadre d'une conceptualisation sémiotique, en tenant compte tant de l'intension que de l'expression et de l'extension des concepts. Afin de valider ces mesures, nous les comparerons aux proximités et similarités mesurées expérimentalement sur des sujets.

Chapitre 8

Expérimentations

8.1 Analyse distributionnelle des gradients de prototypicalité

8.1.1 Jeux de tests

Afin d'évaluer l'analyse distributionnelle des valeurs du gradient *spg* pour une ontologie (appelée O dans la suite) sur différents types de structures hiérarchiques (dont le nombre de fils est variable pour chaque concept), nous avons développé un prototype spécifique dont les paramètres sont : N le nombre de concepts de O, H la profondeur maximale de O, et W la largeur maximale de O.

Les résultats présentés sur la figure 8.1 ont été calculés dans le contexte suivant :

- une hiérarchie de concepts O_1 avec une structure d'arbre décrite par ($N = 800$, $H = 9$, $W = 100$) ;
- une hiérarchie de concepts O_2 avec une structure de treillis et une densité [1] de 0.5 décrite par ($N = 800$, $H = 9$, $W = 100$) ;
- $\alpha = 0.33$ - pondération de la composante intensionnelle, $\beta = 0.33$ - pondération de la composante expressionnelle, $\gamma = 0.33$ - pondération de la composante extensionnelle et $\delta = 1$ - pondération de la composante émotionnelle.

Une hiérarchie de concepts en treillis présente de nombreuses valeurs de *spg* très faible, alors qu'une hiérarchie de concepts en arbre a des valeurs de *spg* plus équilibrées. Ces résultats confortent le fait que le multi-héritage conduit à une dilution de la typicalité. En effet, la notion de typicalité au sein des gradients de prototypicalité - qu'elle soit conceptuelle ou lexicale - est essentiellement une évaluation de la proportion d'information partagée entre deux concepts liés hiérarchiquement [2]. Plus il y a d'ascendance ou de descendance, plus l'apport d'un concept particulier est dilué dans la masse.

Les résultats présentés dans la figure 8.2 ont été calculés dans le contexte suivant :

1. La *densité* du graphe est ici le rapport entre le nombre d'arcs et le nombre de sommets.
2. Seule la composante intensionnelle est une distance entre le concept et le prototype.

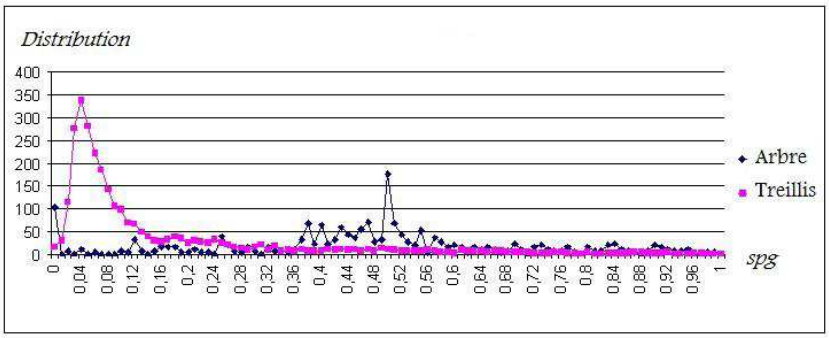


FIGURE 8.1 – Distribution des valeurs de *spg* sur les arcs.

- une hiérarchie de concepts O_1 avec une structure d'arbre décrite par ($N = 800$, $H = 9$, $W = 100$) ;
- une hiérarchie de concepts O_2 avec une structure d'arbre décrite par ($N = 50$, $H = 2$, $W = 30$) ;
- $\alpha = 0.33$, $\beta = 0.33$, $\gamma = 0.33$ et $\delta = 1$.

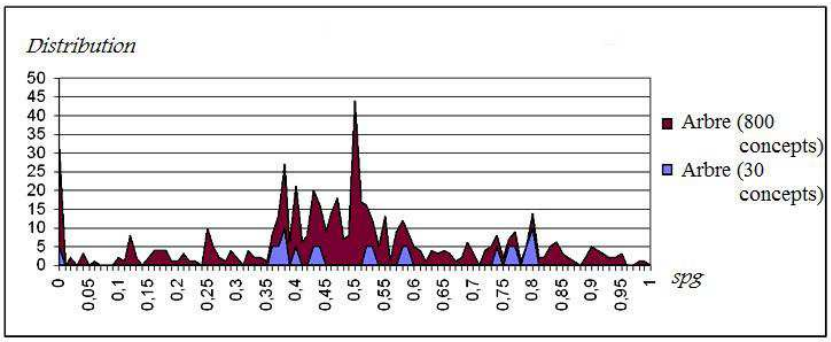


FIGURE 8.2 – Distribution des valeurs de spg dans des arbres.

Ces résultats montrent l'influence du nombre de concepts au sein d'un arbre. Ils indiquent une relative stabilité de la distribution des valeurs du *spg*, proportionnellement au volume des hiérarchies de concepts, et ce pour une même densité de graphes.

8.1.2 Application 1 : domaine de l'Agriculture

TOOPRAG a été utilisé dans le cadre d'un projet dédié à l'analyse de textes décrivant la Politique Agricole Commune (PAC) de l'Union Européenne. Dans ce projet, une ontologie spécifique a été définie à partir du thésaurus multilingue Eurovoc (*http ://europa.eu/eurovoc/*). Ce thésaurus, qui existe en 21 langues officielles de l'Union Européenne, couvre plusieurs thématiques (par exemple politique, éducation et communication, science, environnement, agriculture, énergie, etc.). Il fournit

une base pour permettre aux utilisateurs d'indexer les documents du système de documentation des institutions européennes. À partir des thématiques du thésaurus dédiées au domaine de l'agriculture, nous avons défini une première hiérarchie de concepts en utilisant la relation d'hyponymie/hyperonymie (identifiée comme telle par le lien *Broader Term* dans Eurovoc) et les relations de synonymie (identifiée par le lien *Used For* dans Eurovoc). Ensuite, cette hiérarchie a été modifiée puis validée par un expert du domaine agricole et forestier.

Dans sa version courante, cette ontologie inclut une hiérarchie de concepts ayant une structure d'arbre de 283 concepts (profondeur=4 et largeur maximale=11). Le lexique de cette ontologie est composé de 597 termes. En moyenne, chaque concept est associé à 2,1 termes (min=1 et max=11). La figure 8.3 présente un extrait de cette hiérarchie chargée dans Protégé [3].

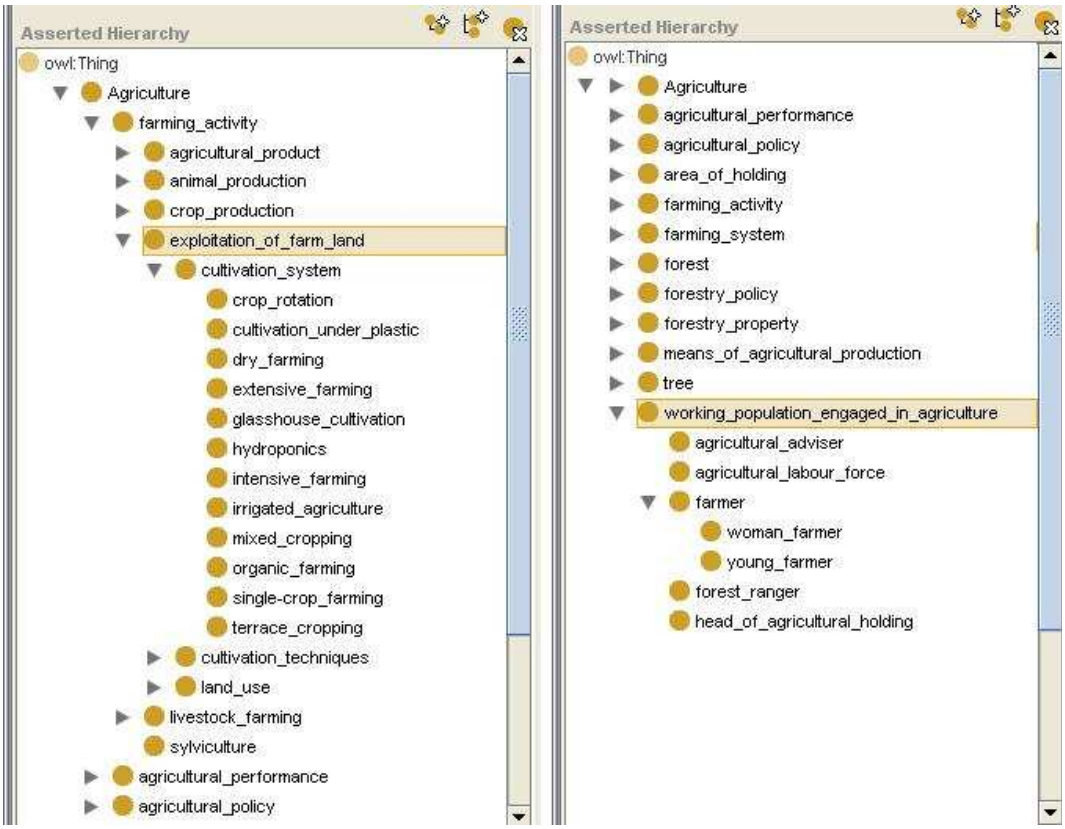


FIGURE 8.3 – Extrait de la hiérarchie de concepts d'une ontologie dédiée au domaine de l'agriculture.

Le corpus utilisé pour cette expérimentation est composé de 55 textes publiés dans le Journal Officiel de l'Union Européenne [4] depuis 2005, soit 43 règlements, 1 directive,

3. http ://protege.stanford.edu
4. http ://eur-lex.europa.eu

8 décrets et 3 avis (avec un total d'environ 1.360.000 mots). Le calcul des gradients de prototypicalité *spg* sur l'ontologie EuroVoc à partir de ce corpus donne les résultats suivants :

- 23% des valeurs de spg non nulles ;
- 9% des valeurs de spg sont égales à 1 ;
- 3% des valeurs de spg sont dans l'intervalle [0.75 , 1[;
- 2% des valeurs de spg dans l'intervalle [0.50 , 0.75[;
- 7% des valeurs de spg dans l'intervalle [0.25 , 0.50[;
- 5% des valeurs de spg dans l'intervalle [0,125 , 0,25[;
- 19% des valeurs de spg dans l'intervalle [0,01 , 0,125[;
- 55% des valeurs de spg dans l'intervalle]0 , 0,01[.

La valeur moyenne des spg est de 0,143. En ce qui concerne la représentation des concepts dans le corpus :

- 61 concepts sont utilisés directement dans le corpus par au moins un terme ;
- 37 le sont indirectement au moyen d'un terme dénotant un concept de leur descendance.

Cette représentation des concepts au sein du corpus donne un taux de couverture (quantité de concepts présents dans le corpus) d'environ 34,63%.

En ce qui concerne les termes :

- le taux de termes de l'ontologie présents dans le corpus est de 34% environ ;
- près de 26.000 termes de l'ontologie ne sont pas présents dans le corpus ;
- l'index du corpus comprend 79.782 entrées.

Bien que l'ontologie considérée dans ce projet n'inclut pas de propriétés (*i.e.* attributs et relations de domaine), les résultats fournis par TooPrag pour ce corpus spécifique sont intéressants car où ils apportent aux experts du monde agricole une aide à l'analyse sémantique des textes réglementaires de la PAC de l'Union Européenne. Par exemple, les valeurs de spg soulignent le fait que, depuis 2005, le système de culture encouragé par la PAC est orienté vers l'agriculture biologique. De façon similaire, les valeurs de spg montrent que les *cols bleus* du secteur agricole (fermiers, paysans, etc.) sont plus représentés que les *cols blancs* (conseillers agricoles, chefs d'exploitation).

8.1.3 Application 2 : domaine HSE

Le calcul des gradients de prototypicalité sur l'ontologie HSE-Tennaxia, à partir d'un corpus composé de 1.100 textes réglementaires dans ce domaine, donne les résultats suivants :

- 30% des valeurs de spg non nulles ;
- 4% des valeurs de spg sont égales à 1 ;
- 6% des valeurs de spg sont dans l'intervalle [0.75 , 1[;

- 3% des valeurs de spg dans l'intervalle [0.50 , 0.75[;
- 5% des valeurs de spg dans l'intervalle [0.25 , 0.50[;
- 5% des valeurs de spg dans l'intervalle [0,125 , 0,25[;
- 16% des valeurs de spg dans l'intervalle [0,01 , 0,125[;
- 63% des valeurs de spg dans l'intervalle]0 , 0,01[.

La valeur moyenne des spg est de 0,128. En ce qui concerne la représentation des concepts dans le corpus :

- 1383 concepts sont utilisés directement dans le corpus par au moins un terme ;
- 90 le sont par héritage, soit un taux de couverture d'environ 15%.

En ce qui concerne les termes :

- le taux de termes de l'ontologie présents dans le corpus est de 13% environ ;
- près de 17.000 termes ne sont pas présents dans le corpus ;
- l'index du corpus comprend 97.632 entrées ;
- en ce qui concerne les substances chimiques, plus de 60% sont représentées dans le corpus au moyen de références issues de nomenclatures et non par leur nom usuel.

8.2 Comparaisons gradients de prototypicalité / jugement humain

Une expérimentation a été réalisée sur un groupe de 19 individus d'une promotion d'élèves ingénieurs en 5 année d'étude pouvant être sur ce critère considérée comme un endogroupe. Cette expérimentation a porté sur deux aspects : (1) la construction d'une ontologie à partir de textes, et (2) la personnalisation de cette ontologie par le calcul de gradients de prototypicalité.

L'objectif de cette expérimentation était de tester une méthodologie collaborative de construction d'une OVD, et de valider les valeurs de prototypicalité conceptuelle avec un jugement humain dans le cadre d'une OPVD.

8.2.1 Construction de l'ontologie

La première phase de l'expérimentation a consisté à construire, en un temps limité et collectivement, une ontologie ayant trait au Grenelle de l'environnement [5]. Pour ce faire, un corpus de textes représentatifs du domaine a été constitué à partir de sites web, à raison d'environ 15 textes par site (partis politiques, associations militantes, etc.). Les sites web retenus sont :

- Chasse-Pêche-Nature-Tradition (*http ://www.cpnt.asso.fr*)

5. Ensemble de rencontres citoyennes organisées depuis octobre 2007, dont l'objectif est de prendre des décisions à long terme en matière d'environnement et de développement durable.

- Front National (*http ://www.frontnational.com*)
- Les Verts (*http ://www.lesverts.fr*)
- Mouvement Démocrate (*http ://www.mouvementdemocrate.fr*)
- Mouvement Pour la France (*http ://www.pourlafrance.fr*)
- Nouveau Parti Anti-capitaliste (*http ://www.npa2009.org*)
- Parti Communiste (*http ://www.pcf.fr*)
- Parti Socialiste (*http ://www.parti-socialiste.fr*)
- Union pour un Mouvement Populaire (*http ://www.u-m-p.org*)
- Palais de l'Elysée (*http ://www.elysee.fr*)
- Fondation Nicolas Hulot (*http ://www.fondation-nicolas-hulot.org*)
- Greenpeace (*http ://www.greenpeace.org/france*)

Chaque étudiant a traité un sous-ensemble du corpus obtenu. L'objectif, pour chacun d'eux, était d'extraire de chaque texte entre 10 et 15 termes qu'ils jugeaient les plus pertinents. Ce choix s'est fait soit de manière manuelle, soit au moyen de KGen [6], un module complémentaire du navigateur FireFox qui analyse les mots contenus dans une page et détermine ceux qui, pour les moteurs de recherche, ont le plus d'importance.

Après consolidation (*i.e.* suppression des doublons, des verbes, etc.) des différentes listes de termes données par l'ensemble des étudiants, près de 350 termes composés de un à deux mots ont été retenus. Ces syntagmes nominaux ont tous un rapport avec le Grenelle de l'environnement, au travers de sous-domaines comme l'environnement, les déchets, les ressources, les structures sociales, etc. Cette consolidation a été effectuée de manière consensuelle par l'ensemble des étudiants composant l'endogroupe, dans le sens où chaque terme présent dans cette liste a fait l'objet préalablement de discussions et d'un accord sémantique par tous. Nous y trouvons, par exemple, les termes *environnement*, *engagement*, *discours*, *message*, *fondation*, *jeunesse*, *république*, *nucléaire*, *biodiversité*, etc.

Lors de la phase de conceptualisation, nous avons choisi de créer un concept par terme (avec suppression des termes synonymes). Cette phase a consisté à rechercher des relations hiérarchiques entre concepts. Par exemple, *Déforestation* et *Consommation* sont considérés, par les étudiants, comme des sous-concepts de *Action*. Cependant, la conceptualisation adoptée a demandé dans certains cas de prendre partie, le choix de la sémantique reflétant également une opinion. Par exemple, le concept *Nucléaire* est à rapprocher du concept *Ressources*, mais doit-il être rattaché au concept *Polluant* ? D'autres questions d'ordre plus encyclopédique se sont posées : l'Europe est-elle une *Communauté*, un *Continent*, ou une *Organisation* ?

Au final, sur la liste initiale, 47 concepts ont été éliminés. Ils ne pouvaient ni être rattachés à un concept existant, ni former des concepts entre eux. Il en ressort une

6. http ://kgen.elitwork.com/accueil.html

hiérarchie conceptuelle (dont un extrait est présenté par la figure 8.4) :

- composée de 176 concepts (un terme par concept) ;
- sur une profondeur maximale de 3 niveaux ;
- avec une largeur moyenne de 9 sous-concepts par concept.

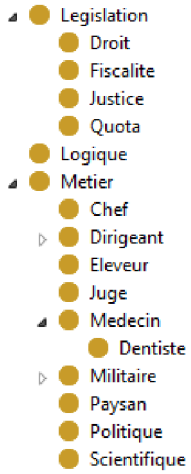


FIGURE 8.4 – Extrait de l'ontologie produite, visualisation au moyen du logiciel TOPBRAID.

8.2.2 Personnalisation de l'ontologie

L'objectif de cette seconde phase a été, à partir de l'ontologie précédemment créée,- de calculer, pour chaque étudiant, les valeurs des gradients de prototypicalité conceptuelle sur les liens hiérarchiques, puis de comparer les résultats obtenus avec son jugement personnel pour une catégorie donnée.

Chaque étudiant a constitué son corpus de référence, corpus composé d'une dizaine de documents, suffisamment longs et pertinents, et surtout en adéquation avec sa conception du Grenelle de l'environnement. Ces documents peuvent être de natures différentes (discours, textes encyclopédiques, articles de presse, etc.) et provenir de multiples sources sur le Web (blog, réseaux sociaux, sites institutionnels, ...).

L'ontologie construite ne comporte ni propriétés, ni instances, il s'agit d'une hiérarchie conceptuelle (avec héritage multiple).

Nous avons ensuite demandé à chaque étudiant d'ordonner par représentativité décroissante (suivant leur jugement) la liste des sous-concepts de chaque concept. Chaque étudiant a calculé les gradients de prototypicalité conceptuelle (composante expressionnelle) pour l'ensemble des liens hiérarchiques, et ce en fonction de son propre corpus.

La dernière phase a consisté, pour chaque étudiant, à comparer les valeurs de *spg* calculées et le classement qu'il a effectué préalablement (*cf.* table 8.1).

cpt	$spg(ressource, cpt)$	jugement humain
énergie	0.93	1^{er}
eau	0.41	2^{e}
argent	0.01	3^{e}
matière	0.03	4^{e}

TABLE 8.1 – Valeurs de *spg*, pour un étudiant, avec le concept *Ressource*

Il ressort de cette expérimentation que 89% (soit 17 étudiants) ont obtenu des résultats similaires ou tout au moins très proches de leur opinion, et 11% (soit 2 étudiants) ont obtenu des résultats différents. Ces différences peuvent s'expliquer par le fait que, dans le second cas, leur corpus personnel ne correspondait pas réellement à leur vision du domaine.

8.2.3 Analyse des résultats

Cette expérimentation répondait à un objectif : tester la correspondance entre les valeurs des gradients de prototypicalité et un jugement humain.

La seconde partie de l'expérimentation montre que la qualité du processus de personnalisation est très dépendante, dans sa composante expressionnelle, de la composition du corpus personnel. Les seuls cas de dissonances entre les résultats calculés et le jugement humain furent relevés lorsque ces corpus ne reflétaient pas exactement les opinions personnelles. Il se pose dès lors la question de la prise en compte des documents sélectionnés par l'utilisateur, tels que ses mails (rédigés ou reçus), ses textes (rédigés ou consultés), ses sites Web et blogs (personnels ou consultés). Ceux-ci doivent refléter la perception que peut avoir l'individu de son univers. Le but de cette opération est d'établir une sorte de profil de l'individu à partir d'un corpus. Mais la constitution d'un tel profil soulève plusieurs problèmes.
En premier lieu, celui de l'*étendue* des documents. Ce profil peut soit être global - pour une ontologie qu'il serait possible de qualifier de généraliste, soit spécifique - pour une ontologie de domaine. Cette distinction a une conséquence sur le choix des documents à indexer. Il se pose également la question de la *récence* des documents, l'appréhension d'un domaine donné évoluant au fil du temps. Il est également pertinent de s'intéresser à la *nature* des documents choisis.

8.3 Expérimentation à l'aide de THESEUS

8.3.1 Protocole

Une expérimentation est réalisée sur le corpus de textes réglementaires de la société Tennaxia à partir du prototype THESEUS développé au sein de ce projet.

Elle vise à évaluer plusieurs aspects : (1) l'intérêt de l'utilisation de l'ontologie HSE-Tennaxia pour l'extension des requêtes, et (2) l'intérêt de la personnalisation de cette ontologie pour la RI par le calcul des gradients de prototypicalité.
Cette évaluation porte sur la recherche de 20 termes, composés de un à six mots, dénotant 20 concepts de l'ontologie HSE-Tennaxia possédant différentes caractéristiques (feuille, descendance sur un à deux niveaux, sur-concepts multiples, sur-concept unique). La recherche s'effectue sur la base des textes réglementaires de la société Tennaxia comportant 2 483 documents.

8.3.2 Utilisation de l'ontologie HSE-Tennaxia

Il s'agit ici de lancer des requêtes sur le corpus de textes de Tennaxia et d'étendre ces requêtes au moyen d'une ontologie. Ces résultats ont été obtenus à partir de l'ontologie HSE-Tennaxia, sans tenir compte des gradients de prototypicalité conceptuelle et lexicale, et sur la base de textes réglementaires de la société Tennaxia. Le tableau 8.2 présente les résultats obtenus dans les deux cas suivants :

- *Fi* : le terme soumis seul au filtre du module *Veille et Conformité* ;
- *Th* : le terme soumis à Theseus avec une extension à toute la descendance du concept (soit nb_{term} termes).

Afin d'évaluer les résultats obtenus pour chaque terme *Th* soumis à Theseus, nous utilisons les indicateurs de précision[7] (p), de rappel[8] (r), de bruit[9] (b) et de silence[10] (s).

L'extension de requêtes offre à l'utilisateur une quantité de résultats plus importante qu'avec la simple utilisation du filtre. Il s'agit d'une amélioration notable par rapport au système existant. Cependant, lorsqu'un concept possède une descendance importante, ce type d'extension peut fournir un nombre de résultats trop grand, ce qui nuit à l'effet apporté en terme d'intérêt pour l'utilisateur.

THESEUS étant un outil d'extension de requêtes fondé sur une ontologie de domaine, les indicateurs du tableau 8.2 évaluent non pas le prototype mais le filtre du module *Veille et Conformité* auquel THESEUS soumet une requête étendue. Globalement, cet outil génère peu de bruit. Un cas a néanmoins été relevé : celui où les mots composant un terme sont contigus mais dans des cellules distinctes. Le terme *carbone 14* en est un parfait exemple. Le contenu de la dernière cellule d'une ligne i d'un tableau se termine par le mot *carbone*. Le contenu de la ligne $i+1$ débute par le chiffre 14. Le moteur d'indexation, aplanissant les structures des tableaux, détecte le terme *carbone 14* alors qu'en réalité les deux mots sont distincts et ne forment pas un terme. Le faible taux de bruit, et donc la bonne précision, de cet outil sont

7. $precision_i = \frac{Documents.correctement.attribues.a.la.classe.i}{Nombre.de.documents.attribues.a.la.classe.i}$
8. $rappel_i = \frac{documents.correctement.attribues.a.la.classe.i}{Nombre.de.documents.appartenant.a.la.classe.i}$
9. $bruit_i = 1 - precision_i$
10. $silence_i = 1 - rappel_i$

Terme	**Fi**	**Th**	nb_{term}	**p**	**r**	**b**	**s**
acétone	13	*13*	10	1	0,68	0	0,32
agrumes	2	*46*	16	1	1	0	0
aluminium	110	*139*	12	1	0,89	0	0,11
atteinte infectieuse	0	*1*	4	1	0,09	0	0,91
canalisation	335	*574*	7	1	1	0	0
carbone 14	4	*4*	2	0,75	1	0,25	0
contravention	115	*169*	16	1	0,92	0	0,08
déchets de jardins et de parcs	8	*441*	5	1	0,95	0	0,05
Escherichia coli	11	*11*	1	1	1	0	0
essence	77	*77*	6	1	0,94	0	0,06
furane	5	*6*	3	1	1	0	0
harnais	14	*14*	2	1	1	0	0
méthane	77	*82*	5	1	1	0	0
pile	75	*107*	9	1	0,88	0	0,12
pluie	78	*78*	1	1	1	0	0
rail	71	*71*	1	1	1	0	0
S14	2	*3*	8	1	0,33	0	0,67
solvants	168	*168*	3	1	0,98	0	0,02
solvants halogénés	25	*25*	1	1	0,81	0	0,19
tonne	259	*307*	3	1	1	0	0
vertige	4	*4*	1	1	1	0	0

TABLE 8.2 – Résultats obtenus par extension de requête avec THESEUS.

dûs au fait que les recherches sont effectuées en "mot exact" (avec recherche de la forme au singulier et au pluriel).

Si la précision de ce système est assez satisfaisante (proche ou égale à 1 dans 99% des cas), le rappel est en revanche nettement moins bon dès lors que la recherche s'effectue sur des termes composés de plusieurs mots. Si, par exemple, un texte contient la chaîne de caractères *atteintes aigues et sub-aigues*, cette dernière évoque bien deux concepts distincts : *atteintes aigues* et *atteintes sub-aigues*. Or, dans le cadre d'une recherche sur le terme *atteinte sub-aigue*, ce texte n'est pas extrait du fait que les deux mots n'y sont pas contigus. La recherche sur le terme *atteinte infectieuse* est étendue aux sous-concepts de la manière suivante : *atteinte infectieuse* OU *atteinte aigue* OU *atteinte chronique* OU *atteinte sub-aigue*. L'ensemble de ces termes est composé d'au moins deux mots souvent associés ensemble, d'où un taux de rappel relativement faible.

De même, l'extension de requête s'effectuant à partir des termes présents dans l'ontologie, si un terme synonyme dénotant un concept est présent dans les textes, mais absent de l'ontologie, la recherche ne peut s'effectuer sur ce terme et le texte concerné n'est pas extrait.

Le rappel et la précision, dans le cadre de THESEUS, sont fortement tributaires de la complétude de l'ontologie d'une part, et de l'outil d'indexation / recherche utilisé (ici, Lucène) d'autre part.

8.3.3 Utilisation des gradients de prototypicalité

Une seconde série de résultats a été obtenue à partir de l'ontologie HSE-Tennaxia, en tenant compte des gradients de prototypicalité conceptuelle et lexicale, et sur la base de textes réglementaires de la société Tennaxia. L'ontologie ne possédant pas de relations ni d'instances, seule la composante expressionnelle des gradients de prototypicalité conceptuelle a été utilisée.

Les gradients de prototypicalité ont été calculés pour un consultant de Tennaxia, à partir d'un sous-ensemble de 75 textes réglementaires sur lesquels il travaille. Le tableau 8.3 montre les résultats obtenus dans trois cas :

- le terme soumis seul au filtre du module *Veille et Conformité* ;
- *Th1* : le terme soumis à Theseus avec une extension à toute la descendance du concept (soit nb_{term1} termes) ;
- *Th2* : le terme soumis à Theseus avec une extension à toute la descendance du concept en tenant compte des gradients (soit nb_{term2} termes).

Afin d'évaluer les résultats obtenus, nous avons demandé à l'utilisateur de noter la pertinence des résultats par un chiffre compris entre 1 (*pas satisfaisant*) et 10 (*très satisfaisant*).

Terme	Filtre	Th1	nb_{term1}	Th2	nb_{term2}	Note
acétone	13	*13*	10	*13*	10	5
agrumes	2	*46*	16	*2*	1	3
aluminium	110	*139*	12	*120*	5	7
atteinte infectieuse	0	*1*	4	*1*	4	5
canalisation	335	*574*	7	*442*	3	3
carbone 14	4	*4*	2	4	1	5
contravention	115	*169*	16	*243*	2	5
déchets de jardins et de parcs	8	*441*	5	*8*	1	3
Escherichia coli	11	*11*	1	*11*	1	7
essence	77	*77*	6	*77*	6	3
furane	5	*6*	3	*6*	3	7
harnais	14	*14*	2	*14*	2	7
méthane	77	*82*	5	*82*	5	7
pile	75	*107*	9	*85*	3	5
pluie	78	*78*	1	*78*	1	3
rail	71	*71*	1	*71*	1	7
S14	2	*3*	8	*2*	2	3
solvants halogénés	25	*25*	1	*25*	1	3
tonne	259	*307*	3	*307*	3	3
vertige	4	*4*	1	*4*	1	3

TABLE 8.3 – Résultats obtenus par extension de requête et gradients avec THESEUS.

Les gradients de prototypicalité conceptuelle comme lexicale ont pour effet - pour chaque concept - de classer par ordre de typicalité l'ensemble des termes qui le dénotent d'une part, mais également l'ensemble de ses sous-concepts. Ces valeurs ont été calculées en fonction d'un corpus de référence composé de ces 75 textes réglementaires. Pour chaque concept, certains termes de l'ontologie HSE-Tennaxia possèdent - pour cet utilisateur - une valeur de prototypicalité nulle, tout comme certains sous-concepts. Cela signifie que ces concepts peuvent être présents directement (par les termes qui les dénotent) ou indirectement (par les termes qui dénotent leur descendance) dans l'ensemble de la base de textes réglementaires, mais que pour un utilisateur et par rapport à son profil, ils ont une représentativité nettement moindre. Nous pouvons également dire que, dans l'extension de la requête, les termes et sous-concepts de faible valeur de gradient de prototypicalité ne sont pas incorporés. La taille des requêtes, en nombre de termes, s'en trouve par conséquent réduite. La quantité de résultats se situe dès lors entre ce que retournerait le filtre (nombre minimal) et ce que retournerait THESEUS sans tenir compte des gradients de prototypicalité (nombre maximal). Si la précision demeure identique, le rappel diminue automatiquement de manière non négligeable, mais avec comme intérêt de proposer des textes correspondant aux concepts les plus pertinents pour

notre utilisateur.

Les notes attribuées par l'utilisateur possèdent un fort aspect subjectif. En le questionnant sur ses réponses, nous pouvons dire que le taux de satisfaction dépend d'au moins deux critères :

1. l'application ne doit pas retourner trop de résultats ;
2. l'application ne *devrait* pas retourner le même nombre de résultats que le filtre.

Si, globalement, il perçoit très nettement l'amélioration des résultats fournis par le moteur de recherche tant en quantité qu'en qualité, il existe néanmoins quelques souhaits par rapport à un tel produit.

En premier lieu, les utilisateurs souhaiteraient pouvoir étendre leur recherche à des concepts liés non plus seulement par une relation de type hiérarchique mais par tout type de relation. Par exemple, les peintures contiennent des composés organiques volatiles (COV) qui sont des substances dangereuses, et les contenants de peintures forment des déchets qu'il faut traiter. Dans le cadre d'un moteur de recherche sémantique, c'est à dire fondé sur une ontologie modélisant la connaissance du domaine HSE, une recherche sur le concept *Peinture* devrait non seulement retourner les documents traitant de peintures et autres sous-concepts, mais également ceux traitant de COV, de pots de peinture, de déchets issus de l'utilisation de peintures, etc.

Une telle extension provoque inévitablement une inflation sur la quantité de documents retournés. De ce fait, souhaiter que l'application retourne davantage de documents qu'en mode "filtre" peut devenir antinomique avec le souhait de ne pas avoir trop de résultats. Si le fait d'utiliser les gradients de prototypicalité permet de réduire le nombre de documents retournés (en élaguant la requête étendue aux concepts et termes les plus prototypiques), il reste néanmoins à exploiter plus en avant ces données pour affiner encore les résultats. Actuellement, THESEUS ne gère que l'extension des requêtes et pas les résultats fournis à l'utilisateur. Il pourrait être pertinent de se servir également des gradients pour ordonner ces résultats.

8.4 Expérimentations sur les mesures de similarité et proximité

Cette section présente des évaluations expérimentales des mesures PROXSEM et SEMIOSEM. Ces évaluations sont basées sur une double comparaison avec des jugements humains et avec d'autres mesures.

8.4.1 Protocole expérimental

Notre évaluation a consisté à évaluer l'adéquation pour une liste de couples de mots, entre un jugement humain et le résultat produit par le calcul. La liste des

30 mots du test de [Miller 1991] fait encore référence en la matière. Nous avons également utilisé le test WordSimilarity-353 [11] de [Finkelstein 2002] [12], qui contient 353 couples de mots reliés essentiellement par des liens fonctionnels et non plus uniquement hiérarchiques dont nous avons sélectionné 31 couples. Cependant, la lecture des résultats *humains* de ce test montre une confusion entre les notions de similarité et de proximité. En effet, si nous prenons le cas du couple *Tasse* - *Café*, le test WordSimilarity-353 donne une valeur de près de 65% de similarité. Or, si les concepts de *Tasse* et de *Café* sont peu dissociables, ils ne sont pas similaires ! Ce test, s'il fait référence, ne permet pas d'évaluer de manière satisfaisante similarité et proximité. Afin de palier ce problème, nous avons soumis notre sélection à un panel de 32 individus, en suivant une approche similaire à [Finkelstein 2002], mais dans lequel nous avons posé deux questions aux individus : une sur l'évaluation de la similarité et une seconde sur l'évaluation de la proximité. Les résultats obtenus sont donnés par le tableau 8.4. Il montre quelques différences notoires entre les valeurs obtenues par notre enquête et celles du test de Finkelstein, ce qui tend à confirmer la différence entre similarité et proximité.

Afin de calculer les valeurs de similarité et de proximité, nous prenons comme ontologie la partie nom de WordNet 3.0. Nous effectuons une comparaison avec les principales mesures de similarité : Jiang et Conrath ([Jiang 1997]), Wu et Palmer ([Wu 1994]), Resnik ([Resnik 1999]), Hirst et St-Onge ([Hirst 1998]) et Lin ([Lin 1998]). Pour ce faire, nous utilisons les résultats fournis par *WordNet Similarity*, un outil de calcul de similarités fondé sur WordNet 3.0 et développé en Java par Siddharth Patwardhan et Ted Pedersen [13]([Patwardhan 2003]).

En matière de corpus, nous utilisons le Corpus of Contemporary American English [14]. Il s'agit d'un corpus fondé sur des magazines, journaux, romans et textes académiques (de 1920 à 2010) et comportant près de 410 millions de termes.

Notre objectif final est triple, il vise à évaluer l'adéquation :

1. de ces mesures avec la référence humaine du test de Finkelstein ;

2. de la mesure de proximité avec un jugement humain de même nature ;

3. de la mesure de similarité avec un jugement humain de même nature.

11. http ://www.cs.technion.ac.il/ gabr/resources/data/wordsim353/wordsim353.html
12. Les couples de mots en français ont été soumis à des individus vivant en France. Chaque personne évaluait son jugement au moyen d'une notation de 1 à 10 pour chaque couple, la valeur 1 correspondant à une proximité ou similarité nulle, la valeur 10 correspondant à une forte proximité ou similarité.
13. http ://www.d.umn.edu/ tpederse/similarity.html
14. http ://www.americancorpus.org/

8.4.2 Jugements humains de proximité et de similarité

Le test WordSimilarity-353 publie une liste de 353 paires de termes avec, pour chacun, la moyenne des notes de similarité sur une échelle de 0 à 10 données par un ensemble de 16 individus. Cette liste est beaucoup plus récente et complète que [Miller 1991], de par sa taille mais également la nature des liens entre les deux concepts dénotés par ces termes.
Sur ces 353 paires, nous avons sélectionné 31 paires répartis en trois groupes :

- un groupe de 10 paires de concepts estimés peu ou pas similaires, avec une note faible (entre 0 et 1) ;
- un groupe de 10 paires de concepts estimés très similaires, avec une note élevée (entre 8 et 10) ;
- un groupe de 11 paires de concepts estimés plus ou moins similaires, avec une note intermédiaire (entre 1 et 8).

Notre liste, classée aléatoirement, a été soumise à un ensemble de 22 personnes. Il leur a été ensuite demandé de donner une note de proximité et une note de similarité sur la même échelle que le test WordSimilarity-353, sans que les valeurs obtenues à ce test ne leur soit communiquées. Les résultats sont donnés dans le tableau 8.4.

8.4.3 Comparaisons avec le test WordSimilarity-353

Dans la suite de cette section, nous mentionnons par les termes suivants :

- *WS*, les notes obtenues par le test WordSimilarity-353 ;
- *prox*, les notes obtenues par notre test de proximité ;
- et *sim*, les notes obtenues par notre test de similarité.

8.4.3.1 Jugement de proximité vs WordSimilarity-353

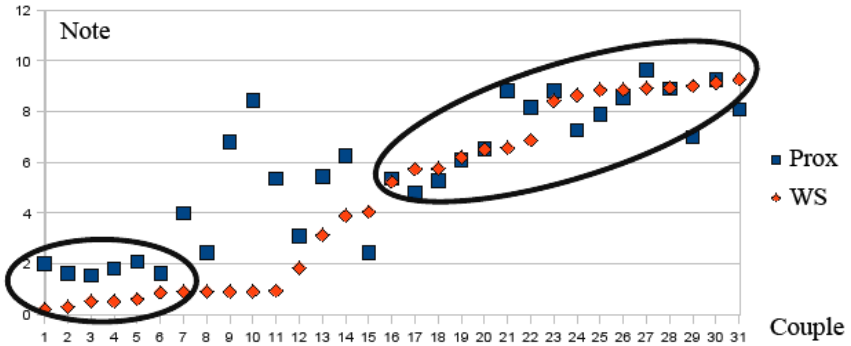


FIGURE 8.5 – Jugement de proximité vs WordSimilarity-353.

Le coefficient de corrélation linéaire entre ces deux ensembles de valeurs (Jugement de proximité et WordSimilarity-353) est de 0,8. Nous pouvons distinguer sur la figure 8.5 trois classes distinctes :

Mot 1	*Mot 2*	*WordSimilarity-353*	*Similarité*	*Proximité*
journey	voyage	*9,29*	*7,55*	*8,0909*
money	cash	*9,15*	*8,73*	*9,2727*
football	soccer	*9,03*	*7,55*	*7,0000*
magician	wizard	*9,02*	*8,64*	*8,9091*
gem	jewel	*8,96*	*9,82*	*9,6364*
car	automobile	*8,94*	*8,36*	*8,5455*
street	avenue	*8,88*	*8,55*	*7,9091*
asylum	madhouse	*8,87*	*7*	*7,2727*
mile	kilometer	*8,66*	*7,91*	*8,8182*
calculation	computation	*8,44*	*5,09*	*8,1818*
street	block	*6,88*	*2,09*	*8,8182*
cup	coffee	*6,58*	*5*	*6,5455*
space	world	*6,53*	*2,36*	*6,0909*
skin	eye	*6,22*	*3,64*	*5,2727*
plane	car	*5,77*	*1,73*	*4,8182*
planet	people	*5,75*	*5,36*	*5,3636*
man	governor	*5,25*	*1,82*	*2,4545*
focus	life	*4,06*	*4,18*	*6,2727*
theater	history	*3,91*	*2,55*	*5,4545*
coast	forest	*3,15*	*2,36*	*3,0909*
forest	graveyard	*1,85*	*4,82*	*5,3636*
lad	wizard	*0,92*	*8,27*	*8,4545*
monk	slave	*0,92*	*3*	*6,8182*
stock	jaguar	*0,92*	*2,55*	*4,0000*
stock	life	*0,92*	*5,18*	*2,4545*
sugar	approach	*0,88*	*1*	*1,6364*
rooster	voyage	*0,62*	*1,18*	*2,0909*
chord	smile	*0,54*	*1,09*	*1,8182*
noon	string	*0,54*	*1*	*1,5455*
professor	cucumber	*0,31*	*1*	*1,6364*
king	cabbage	*0,23*	*1,36*	*2,0000*

TABLE 8.4 – Jugements humains de proximité et de similarité

- une classe rassemblant les paires de concepts avec un *prox* faible et un *WS* faible, où les deux jugements convergent de la paire 1 à la paire 6 ;
- une classe rassemblant les paires de concepts avec un *prox* fort et un *WS* fort, où les deux jugements convergent de la paire 22 à la paire 31 ;
- une classe de notes intermédiaires où les deux jugements divergent parfois fortement ; les concepts évoqués sont estimés avec une proximité (*prox*) supérieure à leur similarité (*WS*) pour les paires 7 à 16.

8.4.3.2 Jugement de similarité vs WordSimilarity-353

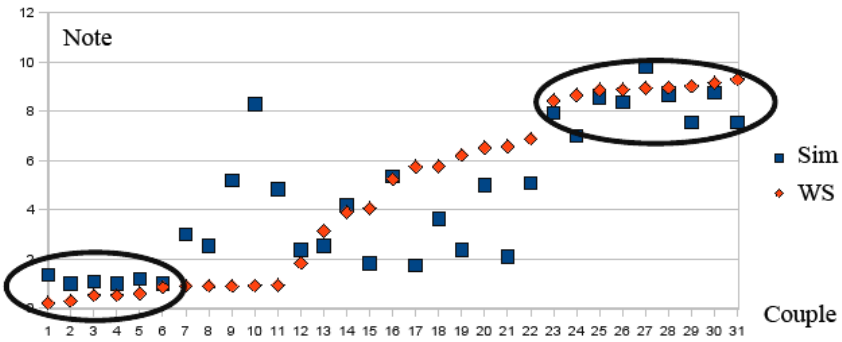


FIGURE 8.6 – Jugement de similarité vs WordSimilarity-353.

Le coefficient de corrélation linéaire entre ces deux ensembles de valeurs (similarité et WordSimilarity-353) est de 0,65. Nous pouvons distinguer sur la figure 8.6 trois classes distinctes :

- une classe rassemblant les paires de concepts avec un *sim* faible et un *WS* faible, où les deux jugements convergent pour les mêmes paires que précédemment ;
- une classe rassemblant les paires de concepts avec un *sim* fort et un *WS* fort, où les deux jugements convergent pour les paires 22 à 31 ;
- une classe de notes intermédiaires où les deux jugements divergent parfoiss fortement. Les concepts évoqués sont estimés avec une similarité (*sim*) supérieure à leur similarité (*WS*), et ce pour les paires 7 à 14 (quasiment même intervalle que ci-dessus). Les concepts évoqués sont estimés avec une similarité *sim*) inférieure à leur similarité (*WS*), et pour les paires 15 à 21.

8.4.3.3 Jugement de similarité vs proximité

Le coefficient de corrélation linéaire entre ces deux ensembles de valeurs (Jugement de similarité et Jugement de proximité) est de 0,82. Nous pouvons distinguer sur la figure 8.7 trois classes distinctes :

- une classe rassemblant les paires de concepts avec un *sim* faible et un *prox* faible, où les deux jugements convergent pour les mêmes paires que précédemment ;

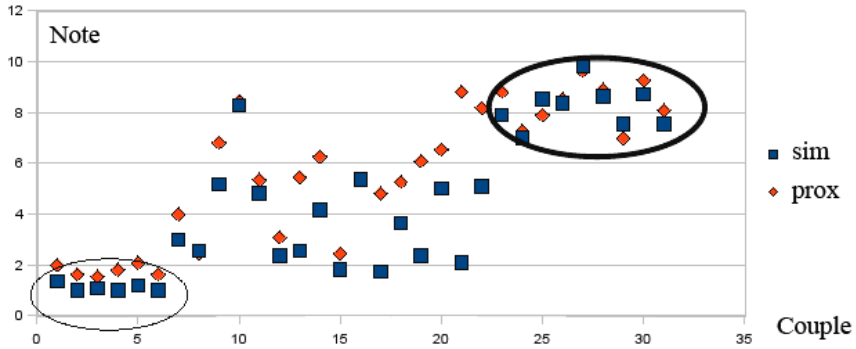


FIGURE 8.7 – Jugement de similarité vs proximité.

- une classe rassemblant les paires de concepts avec un *sim* fort et un *prox* fort, où les deux jugements convergent pour les paires 22 à 31 ;
- une classe de notes intermédiaires où les deux jugements divergent parfois fortement, pour les paires 7 à 21.

8.4.3.4 Moyenne de similarité-proximité vs WordSimilarity-353

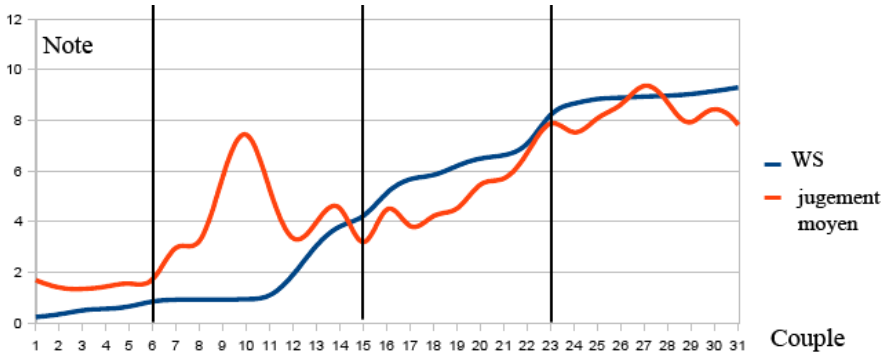


FIGURE 8.8 – Moyenne de similarité-proximité vs WordSimilarity-353.

Nous pouvons distinguer sur la figure 8.8 quatre classes distincts :

- une classe rassemblant les paires de concepts avec un *WS* faible et une moyenne *prox-sim* faible, où les deux jugements convergent pour les mêmes paires que précédemment.
- une classe rassemblant les paires de concepts avec un *WS* fort et une moyenne *prox-sim* forte, où les deux jugements convergent pour les mêmes paires que précédemment.
- une première classe de notes intermédiaires où les deux jugements divergent parfois fortement, pour les paires 7 à 15 (moyenne $>$ *WS*) ;
- une seconde classe de notes intermédiaires où les deux jugements divergent

parfois fortement, pour les paires 16 à 21 (moyenne $<$ *WS*).

8.4.3.5 Différence entre similarité et proximité

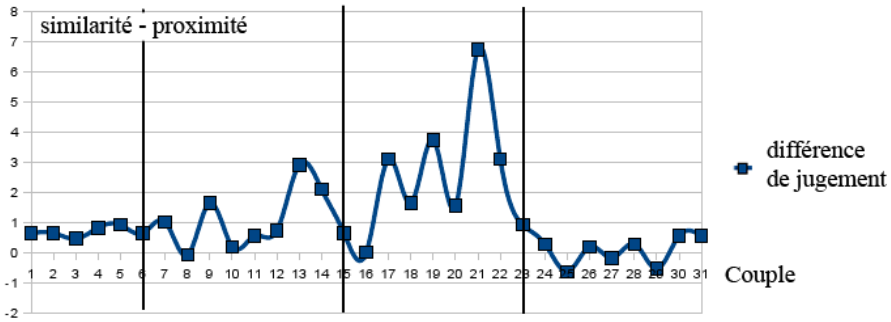


FIGURE 8.9 – Différence entre similarité et proximité.

Nous pouvons distinguer sur la figure 8.9 quatre classes distinctes :
- une première classe avec une faible différence, pour les paires 1 à 7 ;
- une deuxième classe avec une différence plus élevée, pour les paires 8 à 15 ;
- une troisième classe avec une différence plus élevée, pour les paires 16 à 23 ;
- une quatrième classe avec une faible différence, pour les paires 24 à 31.

8.4.3.6 Discussion

L'analyse détaillée des résultats de WordSimilarity-353 montre que les notions de similarité et de proximité peuvent faire l'objet d'une certaine confusion. Il s'avère donc plus pertinent de distinguer ces deux notions et de les évaluer séparément. D'après le calcul des corrélations, il semblerait que les valeurs données par WordSimilarity-353 soient plus proches de la proximité que de la similarité.Il semblerait également que la similarité soit un jugement plus binaire (un concept est ou n'est pas similaire à un autre), alors que la proximité serait une notion plus modulée qui peut prendre de nombreuses valeurs.

Sur les 31 paires de notre test, nous pouvons opérer différents regroupements.
Tout d'abord, deux classes sont corrélées : celles rassemblant les paires de concepts avec un *WS* très faible (paires 1 à 7), et à l'inverse celles rassemblant les paires de concepts avec un *WS* très élevé (paires 25 à 31). Autrement dit, ce qui apparaît comme étant fortement similaire (automobile / voiture) ou au contraire pas du tout (roi et chou) au sein de WordSimilarity-353, l'est également en terme de jugement tant de proximité que de similarité.
Ensuite, un premier bloc de divergence de jugement (paires 8 à 16) apparaît. Il s'agit d'un bloc d'écart important entre proximité et similarité, bloc où (1) la moyenne *prox-sim* est *supérieure* à *WS*, (2) *prox* et *sim* sont supérieurs à *WS*.

Il y un second bloc de divergence de jugement (paires 17 à 24), bloc d'écart important entre proximité et similarité où (1) la moyenne *prox-sim* est *inférieure* à *WS*, (2) seule *prox* est supérieure à Finkelstein, et (3) *sim* et *WS* convergent.

En résumé, nous pouvons en déduire que :
- ce qui est jugé similaire dans WordSimilarity-353 est similaire et proche ;
- ce qui est jugé non-similaire dans WordSimilarity-353 n'est ni similaire ni proche ;
- ce qui est jugé intermédiaire dans WordSimilarity-353 est similaire ou proche.

8.4.4 Comparaison entre les mesures existantes et WordSimilarity-353

Nous évaluons dans un premier temps les principales mesures de similarité utilisées dans la littérature : Jiang et Conrath ([Jiang 1997]), Wu et Palmer ([Wu 1994]), Resnik ([Resnik 1999]), Hirst et St-Onge ([Hirst 1998]) et Lin ([Lin 1998]). Pour ce faire, nous utilisons les résultats fournis par *WordNet Similarity*. Le tableau 8.5 donne les valeurs de corrélation obtenues entre ces mesures - ainsi que les mesures de proximité et de similarité - et les valeurs obtenues avec WordSimilarity-353. Les valeurs de proximité et de similarité ont été calculées à partir de WordNet 3.0 et sur la base du Corpus of Contemporary American English.

Mesure	*Corrélation linéaire*
Jiang & Conrath	0,61
Lin	0,7
Resnik	0,78
Wu & Palmer	0,75
Hirst & St Onge	0,64
PROXEM	0,45
SEMIOSEM	0,71

TABLE 8.5 – Coefficients de corrélation linéaire WordSimilarity-353 / mesures.

Nous pouvons constater que l'ensemble des mesures existantes ont un coefficient de corrélation dans l'intervalle $[0.6, 0.8]$. Ces mesures possèdent une valeur de coefficient de corrélation moins élevée. Ceci est principalement dû à la confusion faite par le test WordSimilarity-353 entre proximité et similarité. Deux concepts similaires dans ce test, par exemple *Cup* et *Coffee*, possédent une plus forte valeur de proximité et une valeur de similarité quasi nulle.
La figure 8.10 représente le classement par ordre croissant des valeurs de similarité pour le test de WordSimilarity-353, en tenant compte des valeurs calculées avec PROXEM et SEMIOSEM. Elle montre que :
- deux concepts jugés non similaires dans WordSimilarity-353 possèdent également une valeur faible avec SEMIOSEM (par exemple corde et sourire, en 28^e

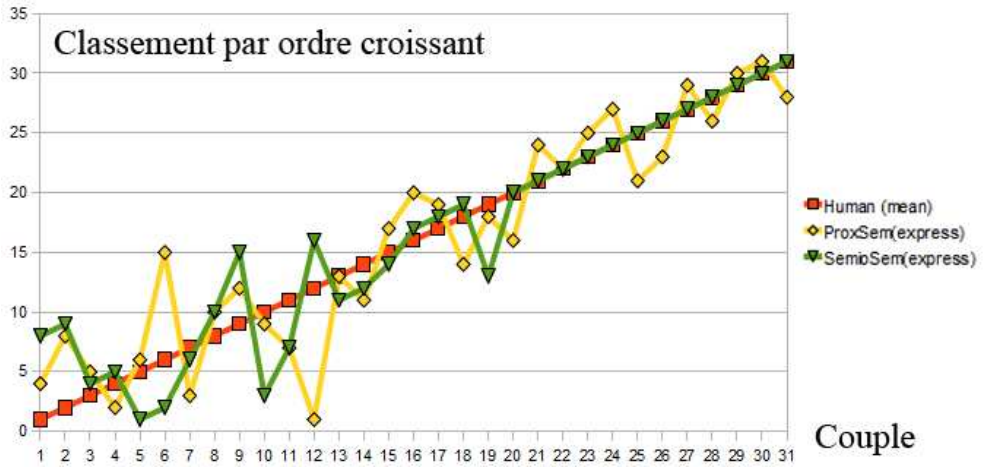


FIGURE 8.10 – WordSimilarity-353, PROXEM et SEMIOSEM.

place dans les deux classements) ;
- deux concepts jugés similaires dans WordSimilarity-353 ne sont pas forcément proches d'un point de vue PROXEM (par exemple, le paire *trajet / voyage*, premier d'un point de vue WordSimilarity-353, et 5^e d'un point de vue PROXEM) ;
- deux concepts mesurés proches avec PROXEM ne sont pas forcémment similaires dans WordSimilarity-353 (par exemple, le paire *tasse / café*, premier d'un point de vue PROXEM, et 12^e d'un point de vue WordSimilarity-353).

Bien que sur les données expérimentales, les deux mesures ont une distribution proche des résultats de WordSimilarity-353, il apparaît plus pertinent - dans le détail - de les évaluer avec des résultats de jugement de similarité et de proximité.

8.4.5 Evaluation de SEMIOSEM avec les jugements de similarité

Mot 1	*Mot 2*	*Similarité)*	SEMIOSEM(**US**)	SEMIOSEM(**FR**)
pierre précieuse	joyau	*9,82*	1,00	1,00
voiture	automobile	*8,36*	1,00	1,00
rue	avenue	*8,88*	0,50	0,75
avion	voiture	*1,73*	0,14	0,13
mile	kilomètre	*7,91*	0,000	0,25
tasse	café	*5*	0,00	0,00

TABLE 8.6 – Extrait des valeurs obtenues avec SEMIOSEM.

Les valeurs de SEMIOSEM, dont un extrait est donné dans le tableau 8.6, ont été calculées uniquement à partir de la composante expressionnelle qui évalue le ratio entre le nombre de termes communs ou synonymes entre les deux concepts et le

nombre total de termes pour désigner ces concepts. La valeur de cette composante, liée au lexique, est fortement dépendante - pour une même ontologie de domaine - de l'endogroupe pour lequel elle est calculée. Alors que le WordSimilarity-353 a été effectué auprès d'une population d'individus anglo-saxons, nos valeurs de références en terme de jugement humain (proximité comme similarité) ont été obtenues à partir d'un ensemble d'individus de nationalité française. Pour être pertinent, la comparaison doit donc se faire à endogroupe équivalent. Pour preuve, le coefficient de corrélation entre les valeurs de SemioSem et celles du jugement humain de similarité est de 0,57 lorsque nous le calculons sur WordNet 3.0, et il monte à 0.69 pour un lexique français.

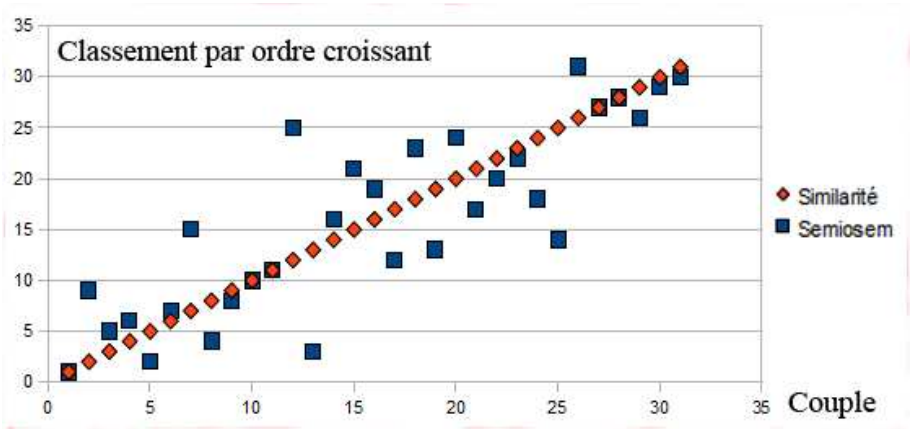


Figure 8.11 – semiosem et similarité.

La figure 8.11 montre la différence de classement (1 pour la paire la plus similaire, 31 pour la moins similaire) entre jugements de similarité et valeurs de semiosem. Avec le lexique français, nous obtenons un coefficient de corrélation de 0,69 et une certaine homogénéité entre les deux types de résultats. Néanmoins, nous n'avons pas calculé SemioSem sur les autres composantes (intensionnelle et extensionnelle), totalement indépendantes du lexique. Il est cependant possible que ces composantes sont également dépendantes de l'endogroupe pour lesquelles nous les calculons

8.4.6 Evaluation de proxsem avec les jugements de proximité

Les valeurs de proxsem, dont un extrait est donné par les tableaux 8.7 et 8.8, ont été calculées à partir de la composante expressionnelle et de la composante intensionnelle qui est égale au ratio entre le nombre de documents contenant les deux termes et le nombre de documents contenant au moins l'un des deux termes. La composante intensionnelle de proxsem est égale au ratio entre le nombre de relations liant les deux concepts dénotés par ces termes et le nombre de relations ayant pour domaine au moins l'un des deux concepts dénotés par ces termes.

Afin de tenir compte de la nature de notre endogroupe de test, nous avons calculé

Mot 1	*Mot 2*	*Proximité*	PROXEM
pierre précieuse	joyau	*9,6364*	*0,50730*
monnaie	argent	*9,2727*	*0,53953*
rue	avenue	*7,9091*	*0,45517*
tasse	café	*6,5455*	*0,41793*
professeur	concombre	*1,6364*	*0,00001*
corde	sourire	*1,5455*	*0,00613*

TABLE 8.7 – Extrait des valeurs obtenues avec PROXSEM.

Mot 1	*Mot 2*	**express**	**intens**	PROXEM
pierre précieuse	joyau	0,000026	0,76862	*0,50730*
monnaie	argent	0,097708	0,76862	*0,53953*
rue	avenue	0,025329	0,67699	*0,45517*
tasse	café	0,018050	0,62420	*0,41793*
professeur	concombre	0,000029	0,00000	*0,00001*
corde	sourire	0,018561	0,00000	*0,00613*

TABLE 8.8 – Valeurs des composantes de PROXSEM.

la composante expressionnelle de notre mesure à partir de l'ensemble des ouvrages français numérisés par Google. La composante intensionnelle a été calculée à partir d'un ensemble de relations impliquant les termes de la liste, et fournie par un second groupe d'individus (de même nature que celui ayant servi pour les valeurs de références). Nous avons également cherché, de manière empirique, à évaluer les pondérations de ces composantes offrant le meilleur taux de corrélation entre les valeurs de PROXSEM et celles du jugement humain de proximité. Nous avons au final obtenu le paire $\alpha = 0.66$ (pondération de la composante intensionnelle), $\beta = 0.33$ (pondération de la composante expressionnelle), et $\gamma = 0.0$ (pondération de la composante extensionnelle).
La figure 8.12 montre la différence de classement (1 pour la paire le plus proche, 31 pour la moins proche) entre jugement de proximité et valeurs de PROXSEM. Avec les valeurs de α et β retenues, nous obtenons un coefficient de corrélation de 0,81 et une certaine homogénéité entre les deux types de résultats.

8.4.7 Discussion

Ces deux évaluations montrent qu'afin de pouvoir évaluer au mieux des mesures de proximité ou de similarité en comparant les résultats obtenus avec des jugements humains, il est nécessaire d'avoir une adéquation entre l'écosystème[15] du groupe de test et le domaine choisi pour le calcul des mesures. Ce groupe d'individus choisi

15. On entend par écosystème l'ensemble formé par une communauté d'êtres vivants et son environnement.

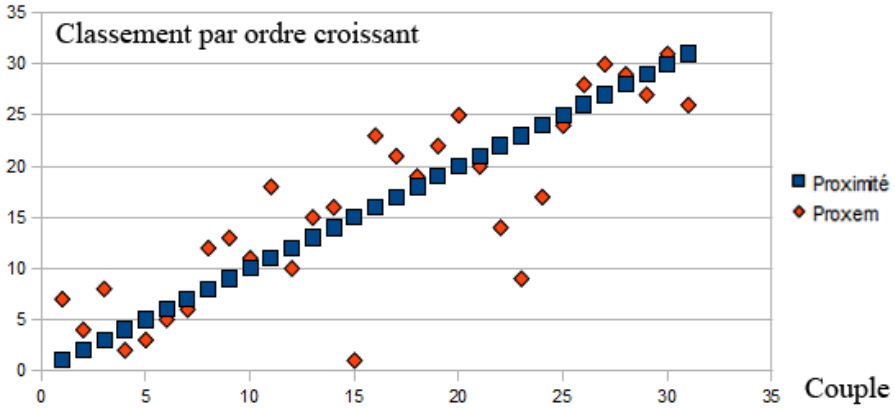


FIGURE 8.12 – PROXEM et proximité.

ici pour les expérimentations était francophone et généraliste (plusieurs spécialités professionnelles étaient représentées), pour pouvoir effectuer une comparaison, notre domaine de modélisation se devait de l'être également. Néanmoins, d'un point de vue expressionnel, une différence a été relevée entre les calculs effectués à partir d'un corpus généraliste anglo-saxon et ceux effectués à partir des ouvrages francophones indexés par Google. Cette différence est d'autant plus sensible avec cette composante, qu'elle est fondée - pour SEMIOSEM - sur le lexique. Or, suivant les cultures, certains concepts sont dénotés par plus de termes que dans d'autres. La valeur de similarité s'en trouve par conséquent fortement dépendante.

Il est également intéressant de noter que les notions de similarité et de proximité ne s'appuient pas identiquement sur les trois dimensions d'une conceptualisation. En ce qui concerne la similarité, nous avons fondé notre expérimentation uniquement sur la composante expressionnelle, avec en résultat un coefficient de corrélation moyen. On peut émettre l'hypothèse que, psychologiquement parlant, le poids des propriétés est prépondérant dans le processus de jugement de similarité, la partie extensionnelle faisant surtout appel à la mémoire épisodique, à la comparaison avec des éléments déjà rencontrés. Deux concepts sont d'autant plus similaires qu'ils possèdent des propriétés communes. Ainsi, un bol et une tasse sont beaucoup plus similaires (ce sont tous deux des récipients ronds destinés à la boisson) qu'une tasse et un café. Quant à la mesure de proximité, la dimension intensionnelle (avec une pondération de 2/3) joue un rôle non négligeable dans ce type de perception. Deux concepts sont d'autant plus proches qu'ils sont reliés par des relations d'une part, et qu'ils sont présents ensemble dans notre univers cognitif (soit sous la forme de termes, soit sous la forme d'instances).

Partie III

Contributions industrielles

Cette thèse ayant été financée par un contrat CIFRE, notre contribution scientifique, *i.e.* l'introduction de la notion de prototypicalité en ingénierie des ontologie, s'est accompagnée d'une contribution industrielle. L'objectif industriel de cette thèse est la réalisation d'une ontologie du domaine Hygiène-Sécurité-Environnement (HSE), et son exploitation dans le cadre d'une recherche d'information au sein d'une base de textes réglementaires. L'ontologie développée dans le cadre de cette thèse, plus orientée Environnement que Hygiène et Sécurité, comporte environ 10 000 concepts dénotés par près de 17 000 termes. Elle couvre les domaines des substances dangereuses, des activités et produits d'activité, des risques, des pathologies et maladies professionnelles, des équipements et installations. Cette ontologie est utilisée dans un processus de recherche d'information sémantique, processus supporté par le prototype Theseus. Ce prototype prend un entrée un terme saisi, et étend la requête au moyen de termes dénotant la descendance du concept recherché. Cette extension est fondée sur le parcours de l'ontologie de domaine personnalisée pour un individu donné. Cette requête étendue est ensuite soumise au moteur de recherche de l'outil Veille et Conformité de la suite logicielle Tennaxia.

—

Chapitre 9 : *Suite logicielle de Tennaxia et Ontologie HSE-Tennaxia*
Chapitre 10 : *Applications*

Chapitre 9

Suite logicielleTennaxia & Ontologie HSE-Tennaxia

9.1 Introduction

L'entreprise Tennaxia[1], dans laquelle s'est effectuée cette thèse, propose une gestion optimisée de l'environnement et de la sécurité, par le biais (1) d'une suite intégrée de logiciels de gestion *Hygiène / Sécurité / Environnement* (HSE) et développement durable en mode *Software as a Service* (SaaS), et (2) de conseils métier.

La fonction HSE, longtemps maintenue dans un rôle purement opérationnel et local, est aujourd'hui appelée à contribuer à la pérennité des entreprises, et ce à différents niveaux :

- maîtrise des risques et des impacts potentiels ;
- optimisation des coûts ;
- santé des personnes ;
- innovation sur de nouveaux procédés ;
- participation à la gestion de l'image de marque et au développement de nouveaux produits...

Pour faire face à ces nouveaux défis, l'entreprise Tennaxia a développé une suite logicielle qui vise à aider les entreprises à maîtriser leurs risques HSE, réduire leurs coûts HSE, et anticiper les évolutions (réglementaires, attentes des différentes parties prenantes).

Cette suite logicielle est accessible sous la forme d'un logiciel de services (*Software as a Service*, SaaS) et commercialisée en trois éditions : Standard, Entreprise et Illimitée. Chaque édition (*cf.* figure 9.1) a été conçue pour regrouper les fonctionnalités nécessaires à la gestion HSE et Développement Durable dans les fonctions :

1. Société Anonyme avec Directoire et Conseil de Surveillance au capital de 77 498 euros. Siège social : 3 rue Léonard de Vinci, 53000 Changé. http ://www.tennaxia.com

1. **réglementaire** : veille, conformité, situation administrative, reportings réglementaires (DRIRE, GEREP, etc.) ;
2. **opérationnelle** : suivi opérationnel des politiques DD-HSE, des rejets et des consommations de ressources, suivi des coûts ;
3. **management** : stratégie développement durable et performance environnementale, gouvernance et communication interne/externe, mise en œuvre d'une approche de type système de management.

FIGURE 9.1 – Suite logicielle Tennaxia

Bien qu'étroitement liés, ces trois aspects de la fonction HSE sont souvent gérés dans les groupes et les entreprises de façon disjointe, que ce soit au niveau de l'organisation interne, des outils ou des systèmes d'information. Les conséquences de cette approche sont :

- une maîtrise imparfaite des risques par manque de vision globale des processus, et difficulté à gérer l'information pertinente au bon moment et au bon endroit ;
- une moindre efficacité de la fonction DD-HSE du fait qu'une proportion importante du temps est consacré à des tâches de faible valeur ajoutée : saisie et contrôle de données hétérogènes, production de rapports à partir de sources disparates, gestion manuelle de plans d'action.

En intégrant toutes les activités de la fonction HSE-DD dans un seul système breveté [2], la suite logicielle Tennaxia permet d'améliorer significativement la maîtrise des risques et la performance. La conformité réglementaire, le suivi des émissions

2. brevet n° FR 2806182 - B1

polluantes et des déchets, la gestion des plans d'action et des politiques environnementale et de sécurité, la production des reportings réglementaires, de développement durable ou de gestion : tout est relié dans un seul système cohérent délivrant une information fiable et libérant du temps pour les activités à plus forte valeur ajoutée.

Le projet développé dans cette thèse devant intégrer le module *Veille et Conformité*, nous ne détaillons pas ici les autres modules de la suite logicielle.

9.2 Veille et Conformité

Face à un nombre croissant de textes réglementaires, il devient de plus en plus difficile pour un industriel de gérer au mieux la conformité réglementaire HSE de ses installations et unités de travail. Les exigences applicables sur certains sites industriels se comptent désormais en milliers ; ce qui entraîne des difficultés de gestion opérationnelles de la réglementation.

Le module *Veille et Conformité réglementaire HSE* de la suite logicielle Tennaxia permet d'assurer une traçabilité complète du processus depuis la veille réglementaire jusqu'aux audits de conformité réglementaire. Il s'agit en premier lieu d'une offre de conseils qui consiste à identifier les textes et les exigences applicables afin d'anticiper les évolutions induites par certaines nouveautés réglementaires. Ces conseils s'appliquent de manière graduée en fonction des besoins des industriels :

- aux *textes* : identification et analyse des textes applicables au(x) site(s) industriel(s) ;
- aux *exigences* : identification dans les textes des exigences réglementaires applicables au(x) site(s) industriel(s) ;
- à la *conformité* : affectation de chaque exigence réglementaire aux installations et préparation des fiches de conformité nécessaires à l'évaluation de la situation de conformité des sites industriels.

D'un point de vue opérationnel, la procédure mise en place est la suivante. Le client reçoit un bulletin de veille réglementaire HSE par email selon la fréquence qu'il a choisie. Un consultant Tennaxia l'appelle pour le commenter et détailler les conditions d'applicabilité des nouveautés réglementaires qu'il lui propose. L'ensemble des textes analysés en exigences réglementaires est mis à sa disposition dans une base de données accessible via internet lui donnant ainsi un accès instantané à une information réglementaire pertinente pour l'évaluation de sa conformité réglementaire. Le processus de gestion des textes est le suivant :

1. les textes réglementaires sont copiés dans leur format original et classés par domaines et thèmes, le tout avec une possibilité de gérer les textes locaux (arrêtés préfectoraux, conventions de raccordement, etc.) ;
2. les textes réglementaires sont analysés manuellement par les consultants et

ventilés en *exigences réglementaires* ;

3. les textes réglementaires sont complétés par des notes d'analyse ;
4. toute modification de textes réglementaires est gérée.

Dans une approche groupe (c'est à dire un ensemble de sites distincts appartenant à un même client), des fonctions de diffusion de la veille réglementaire HSE et de pilotage groupe de la conformité réglementaire sont également disponibles (affectation des textes aux sites concernés, notifications de mise à jour, tableaux de synthèse de conformité Groupe). Pour les clients, les avantages sont multiples. Cette approche permet de :

- *prévenir* les risques de non conformité à partir d'une information pertinente et actionnable ;
- *capitaliser* les connaissances dans une base documentaire structurée et en ligne ;
- *éviter* les analyses redondantes ;
- *aider* chaque site en focalisant leur attention sur la conformité.

Les risques de non-conformité se traduisent en risques d'accidents, en risques de non assurance, en risques de pénalités financières, en risques juridiques et pénaux pour la direction du site, et également en risques de dégradation de l'image de l'entreprise et de difficultés de renouvellement des certifications des systèmes de management. Des difficultés administratives en cas d'évolution souhaitée du site (par exemple, nécessité de retarder un agrandissement et de le subordonner à telle ou telle mise en conformité) peuvent également être dues à une non conformité des installations. La conformité HSE, c'est :

- la conformité *aux exigences réglementaires* applicables ;
- la conformité *administrative* (adéquation entre la réalité des installations du site et leur statut vis à vis des régimes de la nomenclature ICPE ou des statuts Seveso le cas échéant) ;
- la conformité *normative* pour les sites qui sont engagés dans des démarches systèmes de management Santé Sécurité au Travail et/ou Environnement ;
- la conformité *opérationnelle* : traçabilité des déchets au quotidien, Bordereau de Suivi des Déchets (BSD) conformes, déclarations administratives conformes, gestion au quotidien du transport de matières dangereuses...

Les principales fonctionnalités du module *Veille et Conformité* sont :

- la définition des exigences réglementaires à partir de la veille réglementaire ;
- la définition des installations et unités de travail d'un site ;
- la création des fiches de conformité exigence/installation et exigence/unité de travail ;
- la définition de plans d'action de mise en conformité ;
- le suivi détaillé des plans d'action ;
- la prise en compte des coûts de mise en conformité.

9.2.1 Processus de veille réglementaire au moyen de *Veille & Conformité*

Les textes - arrêtés, décret, etc. - sont téléchargés en règle générale sur le site web de l'INERIS[3] à raison d'environ dix par mois. L'expert au sein d'un groupe industriel, ou d'un site, enregistre chaque texte[4] avec un certain nombre de méta-données (au plus 12, nombre paramétrable suivant le client). De ce fait (*cf.* figure 9.2) un texte est défini par :

- son thème - liste fermée (par exemple *environnement : installations classées*) ;
- sa nature - liste fermée (par exemple *arrêté, décret, directive, loi, etc.*) ;
- sa référence - identifiant unique (par exemple *2005-635*) ;
- son libellé (par exemple *Texte relatif au contrôle des circuits de traitement des déchets*) ;
- sa date de création (par exemple *30 mai 2005*) ;
- sa date de parution au JO/BO (par exemple *31 mai 2005*) ;
- sa date de modification (par exemple *15 octobre 2006*) ;
- un titre complet incluant la date et le type de document (par exemple *Décret n° 2005-635 du 30 mai 2005 relatif au contrôle des circuits de traitement des déchets - JO du 31 mai 2005*) ;
- s'il s'agit d'un projet de loi ou d'un texte abrogé ;
- un commentaire au niveau groupe ;
- les conditions d'applicabilité (date, installations visées...) ;
- et son contenu en texte intégral.

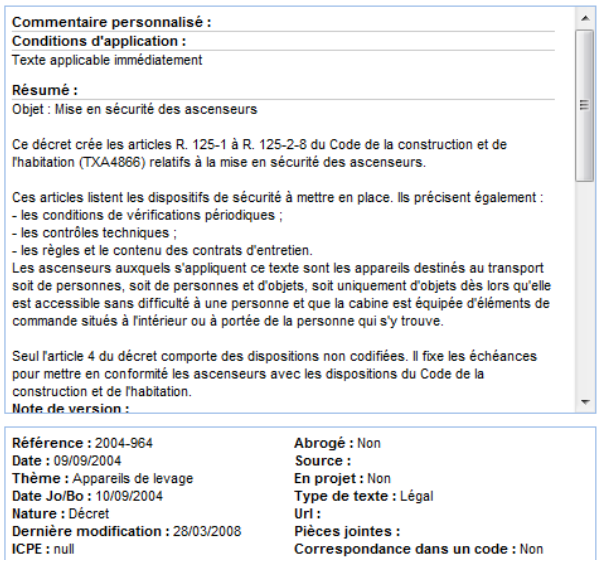


FIGURE 9.2 – Méta-données sur les textes.

3. Institut National de l'Environnement Industriel et des Risques, http ://aida.ineris.fr/
4. Chacun possède un indentifiant unique.

Une recherche de ces textes peut être effectuée au moyen de filtrages sur ces champs avec une visualisation de ces textes, ou une modification des données afférentes.

N° exigence	Titre exigence	Catégorie	Zones surlignées	N° texte	Thème texte
TXA02425	Interdiction de rejet pour les substances figurant en	T: Interdiction	Article 2 Sans préjudice de textes plus contraignants applicables à différentes	TXA3355	Eau
TXA02426	Interdiction de rejeter certaines substances	T: Interdiction	Article 2 (article modifié par l'arrêté du 26 avril 1993 et l'arrêté du 13 juin 2005)	TXA3355	Eau
TXA02427	Eaux de ruissellement	T: Utilisation / Exploitation	Le préfet tient à jour un inventaire des rejets existants mentionnés à l'article 3, avec l'indication de la nature et des quantités de	TXA3355	Eau
TXA02428	Dérogation spécifique	T: Utilisation / Exploitation	- si nécessaire, les mesures permettant la surveillance des eaux souterraines, et en particulier de leur qualité. "	TXA3355	Eau

FIGURE 9.3 – Définition des exigences.

L'utilisateur définit ensuite des "exigences" par surlignage de parties de texte (paragraphes au sens HTML, *i.e.* du texte placé entre des balises $<P>$) (*cf.* figure 9.3) pouvant concerner soit le groupe, soit son site, soit un ou plusieurs sites du groupe. D'après [Després 2007], une exigence est *une portion de texte homogène portant sur un contexte industriel donné et sur un type d'obligation ou d'interdiction donné.* Chacune de ces exigences possède un certain nombre d'informations complémentaires comme un identifiant unique, un titre, un commentaire, une catégorie (*mesure, consigne d'utilisation, etc.*) et une couleur (en sachant qu'à l'heure actuelle, il n'existe aucune sémantique associée à ces couleurs).

Comme pour les textes, l'utilisateur peut effectuer une recherche au moyen de filtrages sur les champs avec pour objectif soit la visualisation de ces exigences, soit la modification des données afférentes. Le responsable de groupe peut dès lors définir une association entre un texte et un ou plusieurs sites (statut de disponibilité) avec un niveau d'applicabilité (binaire : *oui ou non*). Il reste alors aux personnes concernées (averties par un bulletin de veille) à établir des fiches de conformité. Pour cette activité d'expertise, les groupes et les sites travaillent de manière autonome. Il n'y a communication d'information que s'il y a autorisation de partage.
De son côté, un utilisateur rattaché à un site peut :

- définir ses propres exigences (en plus de celles définies par le groupe) ;
- définir la structure des sites (pour le moment, il s'agit plus d'une description d'ordre structurel sous forme arborescente, nous devrions avoir à l'avenir une version centrée sur l'organisation des activités, *cf.* figure 9.4) avec la possibilité de définir une Installation Classée pour la Protection de l'Environnement (ICPE) [5] ;
- sélectionner, accepter et annoter les textes et exigences le concernant en provenance du groupe ;

5. Les ICPE sont des installations exploitées ou détenues par toute personne physique ou morale, publique ou privée, qui peuvent présenter des dangers ou des inconvénients pour la commodité du voisinage, la santé, la sécurité, la salubrité publique, l'agriculture, la protection de la nature et de l'environnement, la conservation des sites et des monuments. http ://www.ineris.fr/aida/files/aida/file/nomenclature.pdf

FIGURE 9.4 – Définition de la structure d'un site.

– établir des fiches de conformité.

Les fiches de conformité (*cf.* figure 9.5) sont des associations installation(s)-exigence. Pour chaque fiche sont également précisés le nom du responsable (le vérificateur de la conformité), le taux de conformité de l'installation (pourcentage), le niveau de criticité (*i.e.* l'impact environnemental - faible à fort), les dates du dernier et du prochain contrôle ainsi qu'un commentaire de conformité.

9.2.2 Moteur de recherche Lucène

La base de textes réglementaires de Tennaxia est composé (1) des textes concernant l'ensemble des clients, et (2) de toutes les méta-données sur ces textes (y compris le découpage en exigences et les méta-données sur ces exigences). Cette base est entièrement indexée au moyen du moteur d'indexation *Lucène*[6].

Le moteur de recherche du module *Veille et Conformité* est uniquement un filtre sur cette base de textes : il sélectionne les documents suivant des critères saisis par l'utilisateur, sans fournir en retour un tri des documents par pertinence par rapport à la requête. Les critères peuvent ainsi porter sur :

– le numéro du texte (par exemple *TXA2356*) ;
– un texte dans le bulletin de veille (par exemple *un texte modifié paru dans le bulletin du 06/07/2010*) ;
– le thème du texte (par exemple *pollution atmosphérique)* ;
– la nature du texte (par exemple *arrêté préfectoral*) ;
– la référence (par exemple *CdT Art. R. 234-1 à 23*) ;
– le titre du texte (sélection sur un des termes, sur tous les termes ou sur la

6. http ://lucene.apache.org/java/docs/index.html

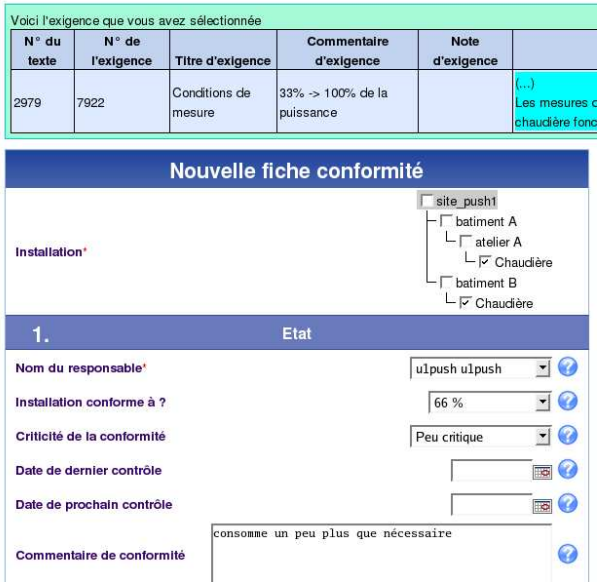


FIGURE 9.5 – Fiche de conformité.

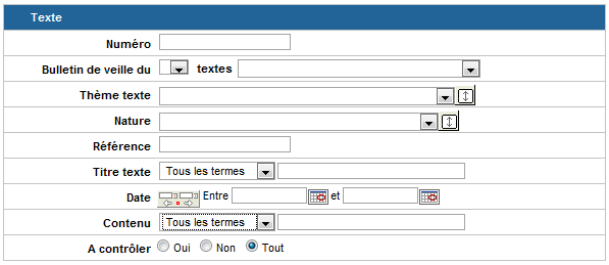


FIGURE 9.6 – Moteur de recherche.

phrase exacte) ;
- la date de parution du texte ;
- le contenu du texte (sélection sur un des termes, sur tous les termes ou sur la phrase exacte) ;

9.3 Ontologie HSE-Tennaxia

L'objectif industriel de nos travaux a été de concevoir, développer et intégrer, pour le compte de la société Tennaxia, un moteur de recherche sémantique au sein d'une base de textes réglementaires. Afin d'apporter des éléments de réponse aux insuffisances relevées au chapitre précédent, il a été décidé de développer ce moteur (1) en s'appuyant sur une ontologie du domaine HSE, et (2) en s'appuyant sur la notion de prototypicalité introduite dans la partie II et en intégrant les techniques de personnalisation en vue d'adapter les recherches aux types d'utilisateurs. Une première ontologie HSE (plus orientée *Environnement* que *Hygiène et Sécurité*) a été construite lors d'une étude préalable aux travaux de cette thèse par [Desprès 2007]. Au sein de cette ontologie HSE v.1.0, construite manuellement par un groupe composé d'experts du domaine et d'experts en IC, quatre principaux concepts apparaissent :

- *les objets d'activité (entités produites ou consommées par une activité)* ;
- *les activités* ;
- *les installations et équipements* ;
- *les processus dommageables.*

Ces quatre concepts généraux subsument bien entendu un ensemble de concepts (voir figure 9.7). Par exemple, un *Risque* est considéré comme étant un *Processus dommageable*, une *Substance dangereuse* comme une notion de *Objet d'Activité.*

Pour les *Activités*, l'analyse des textes, des interviews et des nomenclatures a permis de dresser la typologie suivante :

- *production* : une activité de production a pour but de fabriquer un type d'objets (y compris des substances) en utilisant d'autres objets ou substances et en rejetant éventuellement des objets qui sont des déchets. La production inclut la transformation d'objets (conditionnement, traitement) ;
- *logistique* : une activité logistique a pour but de modifier la localisation d'un objet ou d'un ensemble d'objets ;
- *contrôle* : une activité de contrôle a pour but de produire des mesures et des rapports sur une autre activité ;
- *élimination* : une activité d'élimination a pour but de détruire un objet ou d'assurer son stockage définitif ;
- *valorisation* : une activité de valorisation a pour but de transformer un déchet en un objet pouvant servir à une activité ;
- *commerce* : une activité de commerce modifie le propriétaire d'un objet.

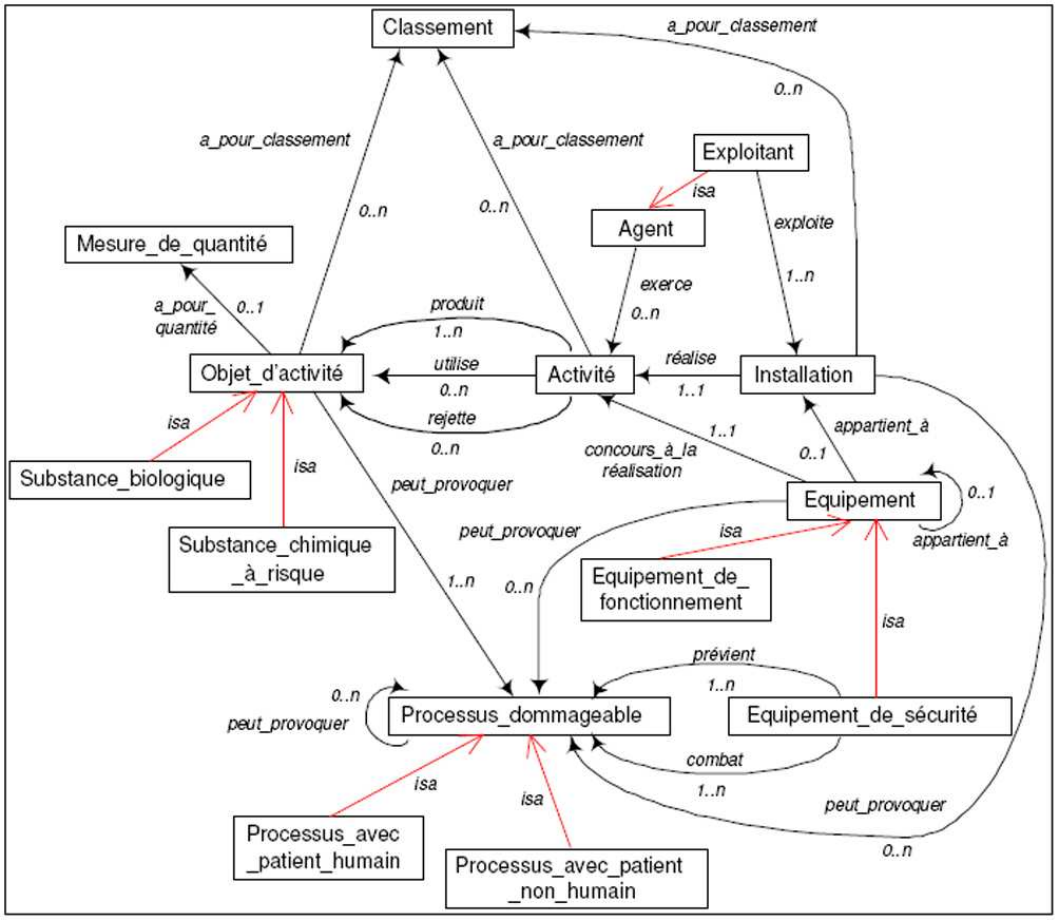


FIGURE 9.7 – Structure globale de l'ontologie.

Au cœur de la réglementation se trouvent les installations où se déroulent les activités. Les *installations et équipements* (industriels comme agricoles) sont définis par les objets physiques qui concourent à la réalisation d'une activité. Une installation et ses équipements sont donc fonction de l'activité, mais aussi des produits utilisés ou générés.

La sémantique du domaine est exprimée par les relations, lesquelles sont données avec des cardinalités qui expriment le nombre potentiel de relations pouvant exister entre instances des concepts liés (ainsi, une activité donnée produit au minimum un objet d'activité mais peut ne pas rejeter d'objet d'activité, *i.e.* de déchets). De manière synthétique :

- un *objet d'activité*, une *installation* ou un *équipement* peuvent provoquer un *processus dommageable* ;
- une *activité* produit, utilise et/ou rejette un ou plusieurs *objets d'activité* ;
- une *installation* réalise une *activité* et un *équipement* participe à cette réalisation. De plus, un *équipement* est rattaché à une *installation*. Dans les *équipements*, on distingue les *équipements de sécurité* qui sont ceux qui participent à la prévention ou à la lutte contre un *processus dommageable*.

Une ontologie plus conséquente en terme de concepts, extension de la version 1.0 et à visée opérationnelle et industrielle, a par la suite été développée dans le cadre de nos travaux. Elle couvre les domaines suivants :

- les *substances dangereuses* avec environ 3.800 concepts ;
- les *activités* avec environ 1.400 concepts ;
- les *produits d'activité* avec environ 1.600 concepts ;
- les *risques* avec environ 800 concepts ;
- les *pathologies et maladies professionnelles* avec environ 900 concepts ;
- les *équipements et installations* avec environ 400 concepts.

Au total, l'ontologie HSE v.2.0 que nous avons développée englobe 9.536 concepts [7], dénotés par 16.268 termes (soit environ deux termes par concepts en moyenne). Cette hiérarchie conceptuelle est d'une profondeur de 12.

À partir de textes réglementaires, de nomenclatures, d'interviews des consultants de la société Tennaxia, la structure générale de l'ontologie a pu être fixée et enrichie des concepts apparaissant dans le corpus de textes réglementaires suivant une méthodologie de type *middle-out*. Les concepts ont été soit extraits manuellement dans leur majeure partie, ou de manière automatique à partir de documents structurés (tableaux) ou à l'aide de SYNTEX. Cette ontologie a été développée sous Protégé, puis TopBraid Composer. Cette ontologie bénéficie d'une double protection légale : (1) protection de la structure auprès de l'Institut National de Propriété Industrielle

7. Il n'y a aucune relation dans cette version.

(INPI)[8] au moyen d'une enveloppe Soleau, et (2) protection du contenu avec le dépôt à l'association « *Scam Vélasquez* »[9] d'un CD de données. Cette ontologie de domaine, nommée *Ontologie HSE-Tennaxia*, est la propriété de la société Tennaxia.

9.3.1 Le corpus

L'ontologie HSE-Tennaxia se fonde sur un corpus constitué uniquement de textes réglementaires, afin de disposer de sources les plus fiables possibles. Il peut exister, en effet, pour certains domaines des dissonances entre les sources, par exemple pour les risques liés aux substances dangereuses.

L'ontologie HSE v.1.0 a été réalisée à partir d'un corpus, comprenant au total 34.021 mots, composé de cinq textes juridiques (des décrets et arrêtés[10]) choisis par des consultants de Tennaxia pour leur représentativité en matière de droit environnemental :

1. décret n°77-1133 du 21 septembre 1977 pris pour l'application de la loi n°76-663 du 19 juillet 1976 relative aux Installations Classées pour la Protection de l'Environnement (*14.772 mots*) ;
2. décret n°99-374 du 12 mai 1999 relatif à la mise sur le marché des piles et accumulateurs et à leur élimination (*2.752 mots*) ;
3. arrêté du 29 juin 2004 relatif au bilan de fonctionnement prévu par le décret n°77-1133 du 21 septembre 1977 modifié (*2.792 mots*) ;
4. arrêté type - Rubrique n°2910 : Combustion (*11.783 mots*) ;
5. décret n°2005-635 du 30 mai 2005 relatif au contrôle des circuits de traitement des déchets (*1.922 mots*).

Cette première ontologie s'est avérée assez rapidement insuffisante pour être utilisée dans une phase pré-industrielle, que ce soit comme base de connaissances ou pour être intégrée dans un module de recherche d'information. Ainsi, plusieurs parties spécialisées de cette ontologie ont été étendues, fondées chacune sur un corpus spécialisé :

- Pour la partie *Substances dangereuses* :
  - annexe I de la version consolidée de la Directive 67/548/CEE ;
  - classification de Mendeleïv ;
  - thésaurus Eurovoc ;
  - résultat d'analyse linguistique de la base de textes réglementaires de la société Tennaxia.
- Pour la partie *Activités* :

8. http ://www.inpi.fr - dépôt enveloppe Soleau N°322.408 en date du 13 juin 2008.
9. http ://www.scam.fr/ - dépôt N°2008090075 en date du 16 septembre 2008.
10. Les directives et lois ne sont pas directement applicables et n'ont donc pas à être prises en compte par les entreprises, sauf - éventuellement - en cas de doute sur l'interprétation d'un décret où il faut peut être remonter à la loi pour voir dans quel esprit elle a été écrite.

  - Code des Douanes, Art. 266 sexies à 266 quindecies : Taxes - TGAP ;
  - Nomenclature des Activités Françaises - révision 2008 - INSEE ;
  - Nomenclature ICPE = article R. 511-9, décret du 12 octobre 2007 ;
  - Eurovoc ;
  - résultat d'analyse linguistique de la base de textes réglementaires de la société Tennaxia.
- Pour la partie *Produits* :
  - décret n°2007-1888 du 26 décembre 2007 portant approbation des nomenclatures d'activités et de produits français ;
  - résultat d'analyse linguistique par SYNTEX de la base de textes réglementaires de la société Tennaxia.
- Pour la partie *Risques* :
  - résultat d'analyse linguistique de la base de textes réglementaires de la société Tennaxia.
- Pour la partie *Pathologies et maladies professionnelles* :
  - arrêté du 18 juillet 1994 fixant la liste des agents biologiques pathogènes - JO du 30-07-1994 ;
  - recommandation 2003/670/CE - JO L238 du 25 septembre 2003 concernant la liste européenne des maladies professionnelles ;
  - Code de la Sécurité Sociale (titre VI du livre IV) ;
  - Tableaux des maladies professionnelles du régime général (1998) ;
  - résultat d'analyse linguistique de la base de textes réglementaires de la société Tennaxia.
- Pour la partie *Équipements et installations* :
  - résultat d'analyse linguistique de la base de textes réglementaires de la société Tennaxia.

Le détail de ces ressources est donné dans l'annexe A.

9.4 Conclusion

Le système de recherche d'information au sein de l'outil *Veille et Conformité* est perfectible. Aujourd'hui, il est de nature purement syntaxique. Une des applications de ce travail de thèse est d'enrichir les fonctionnalités de recherche de ce moteur afin que cet outil ne se limite plus à la fonction de moteur de recherche syntaxique. Sans prétendre devenir exhaustif, il doit fournir un panel de textes réglementaires plus large qui celui proposé actuellement, et ce à l'aide d'une ontologie du domaine HSE. Il est, en effet, souhaitable de pouvoir non plus saisir un mot mais un concept ou une chaîne de concepts par le biais de termes les dénotant. Par exemple, la requête *incendie* doit aussi bien retourner des textes évoquant des *ventilateurs à dépression positive* que des *tubes à mousse* sans que ces mêmes textes ou leurs méta-données comportent le mot *incendie*. Il existe un autre point venant enrichir ces fonctionnalités : la notion de typicalité entre concepts où - par exemple - une *chaudière* est davantage considérée comme un *appareil à combustion* (95% par exemple) qu'un

four (60% par exemple).

L'objectif de la construction de l'ontologie HSE-Tennaxia est de pouvoir l'utiliser comme base d'un moteur de recherche sémantique sur une bibliothèque de textes réglementaires de la société Tennaxia. Il s'avérait donc capital de disposer d'une ontologie ayant à la fois (1) une quantité suffisante de concepts pour pouvoir offrir une réponse à la demande d'un utilisateur, (2) une conceptualisation de qualité pour pouvoir étendre une requête de manière pertinente. Il n'existe pas, à ce jour, d'ontologie véritablement du domaine HSE. D'autre part, il était souhaité de la part des dirigeants de l'entreprise, de disposer d'une ontologie reflétant parfaitement leur vision de ce domaine, et donc construite uniquement à partir de leur sélection de textes réglementaires.

Chapitre 10

Réalisations techniques

10.1 Introduction

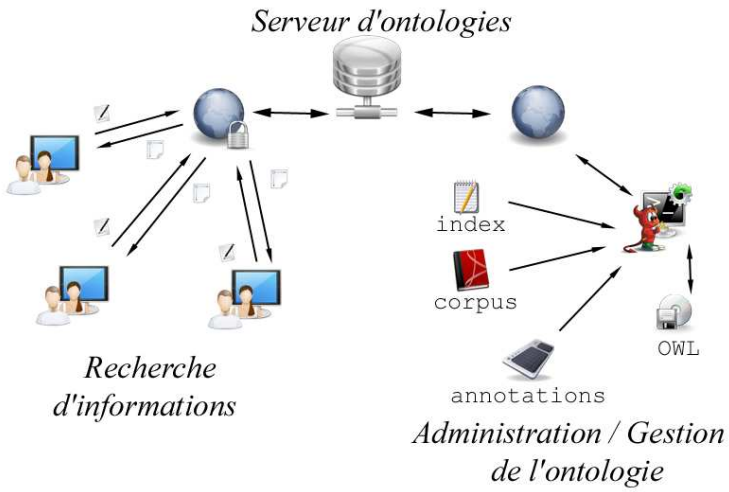


Figure 10.1 – Architecture.

De manière générique, la figure 10.1 représente l'architecture globale du système. L'architecture se structure en trois composantes principales :

- Une *ontologie de domaine* stockée sur un serveur dédié. Dans le cas présent, elle est stockée sous la forme d'une base de données MySQL, mais il est possible d'envisager d'autres formes de stockage en s'inspirant des technologies existantes consacrées aux grands entrepôts de données comme DBPedia, par exemple.
- Une partie *administration à distance de l'ontologie*. Il s'agit d'une application permettant de mettre à jour l'ontologie de domaine, mais également d'assurer sa personnalisation (calculs des gradients de prototypicalité) à partir de corpus de référence. Dans le cadre de ce projet au sein de la société Tennaxia, TooPrag occupe cette fonction.

– Une partie *recherche d'information sémantique* fondée sur l'ontologie de domaine. Il s'agit de moteurs de recherche sémantiques qui étendent leurs requêtes, et ce au moyen de termes dénotant des concepts présents dans une OPVD pour un utilisateur donné. Ce processus d'extension, toujours dans le cadre de ce projet pour la société Tennaxia, est supporté par le prototype THESEUS.

10.2 TOOPRAG

TOOPRAG (*A Tool dedicated to the Pragmatics of Ontology*) est un outil dédié au calcul automatique des gradients de prototypicalité , et à l'administration/gestion de l'ontologie HSE-Tennaxia. Cet outil, implémenté en Java 1.5, s'appuie sur les bibliothèques Lucène [1] et Jena [2].

10.2.1 Principe

TOOPRAG prend en entrée une ontologie représentée en OWL 1.0, où chaque concept est associé à un ensemble de termes définis via la primitive `rdfs:label` et un corpus composé de fichiers au format texte (auquel cas il est indexé à l'aide de Lucène) ou un index. TOOPRAG calcule les valeurs de SPG des liens *is-a* entre concepts, ainsi que les valeurs des LPG de tous les termes utilisés pour dénoter les concepts.

```
…
<owl:Class rdf:ID="agricultural labour force">
  <rdfs:label xml:lang="EN" xml:lpg=0.7>farm worker</rdfs:label>
  <rdfs:label xml:lang="EN" xml:lpg=0.3>agricultural labour force</rdfs:label>
  <rdfs:subClassOf rdf:resource="#working population engaged in agriculture" xml:cpg=0.0074/>
</owl:Class>

<owl:Class rdf:ID="farmer">
  <rdfs:label xml:lang="EN" xml:lpg=0.375>grower</rdfs:label>
  <rdfs:label xml:lang="EN" xml:lpg=0.0>peasant</rdfs:label>
  <rdfs:label xml:lang="EN" xml:lpg=0.0>raiser</rdfs:label>
  <rdfs:label xml:lang="EN" xml:lpg=0.625>farmer</rdfs:label>
  <rdfs:subClassOf rdf:resource="#working population engaged in agriculture" xml:cpg=0.9841/>
</owl:Class>

<owl:Class rdf:ID="forest ranger">
  <rdfs:label xml:lang="EN" xml:lpg=0.0>forest ranger</rdfs:label>
  <rdfs:subClassOf rdf:resource="#working population engaged in agriculture" xml:cpg=0.0/>
</owl:Class>

<owl:Class rdf:ID="agricultural adviser">
  <rdfs:label xml:lang="EN" xml:lpg=0.0>agricultural adviser</rdfs:label>
  <rdfs:subClassOf rdf:resource="#working population engaged in agriculture" xml:cpg=0.0/>
</owl:Class>
…
```

FIGURE 10.2 – Extrait d'un fichier OWL produit par TOOPRAG.

1. Lucène est une librairie Java très performante d'indexation et de recherche full-text. Lucène est disponible en open source : *http ://lucene.apache.org/*.

2. Jena est un framework Java pour la construction d'applications orientées Web Sémantique, permettant la prise en charge de RDF, RDFS, OWL et SPARQL et incluant un moteur d'inférence. Jena est disponible en open source : *http ://jena.sourceforge.net/*.

Il en résulte une OPVD, stockée dans un format OWL étendu par rapport aux spécifications de OWL 1.0. Comme le montre la figure 10.2, une valeur de LPG est représentée par un nouvel attribut *xml :lpg*, directement associé à la primitive *rdfs :label*. Par exemple, les valeurs de LPG des termes " *Grower* " et " *Farmer* ", utilisés pour dénoter le concept *Farmer*, sont respectivement de 0.375 et 0.625. Une valeur de SPG est représentée par un nouvel attribut *xml :spg*, directement associé à la primitive *rdfs :subClassOf*. Par exemple, les valeurs de SPG des liens *is-a* définis entre le concept *Working population engaged in agriculture* et ses sous-concepts *Farmer* et *Agricultural Labour Force* sont respectivement de 0.9841 et 0.0074.

10.2.2 Fonctionnalités

Pour exploiter des ontologies volumineuses (dans notre cas, plusieurs milliers de concepts), nous nous sommes trouvés limités par deux facteurs au début de cette thèse : le manque d'outils, et les limites du langage OWL. *Protégé* s'est avéré trop limité avec de nombreux problèmes de pertes de données et de lenteur d'exécution, et l'éditeur TopBraid, dont nous nous servons aujourd'hui, ne disposait pas encore d'une édition gratuite. Avec le langage OWL, nous ne pouvions pas stocker les valeurs dans des attributs respectant les normes du W3C afférentes à ce langage. Pour surmonter ces limitations, nous avons développé l'outil TooPrag en adoptant un stockage de l'ontologie dans une base de données relationnelles.

D'un point de vue fonctionnel, TooPrag permet de travailler sur :

- les *concepts* : ajout, suppression, modification ;
- les *termes* dénotant les concepts : ajout, suppression, modification, importation ;
- les *commentaires* affectés aux concepts : ajout, suppression, modification, importation, typage (définition, commentaire, *ToDo*) ;
- les *ressources* (document, expert...) d'où proviennent les concepts : ajout, suppression, modification, importation ;
- les *instances* affectés aux concepts : ajout, suppression, modification, gestion de leurs labels et commentaires respectifs.

La figure 10.3 présente une copie de l'un des écrans de TooPrag. La navigation au sein de la hiérarchie de concepts s'effectue :

- soit de manière graphique, avec pour chaque concept actif une vision locale du graphe (sur-concepts et sous-concepts) ;
- soit de manière tabulaire, avec pour chaque concept actif un tableau résumant l'ensemble des sur-concepts et sous-concepts.

Pour chaque concept actif, l'utilisateur peut accéder à l'aide d'onglets aux concepts de son ascendance comme de sa descendance directe, mais également gérer ses labels, ses commentaires, ses instances et les ressources à partir desquelles il a construit ce concept. Pour des liaisons hiérarchiques et les labels, l'utilisateur dispose également des valeurs de gradients de prototypicalité.

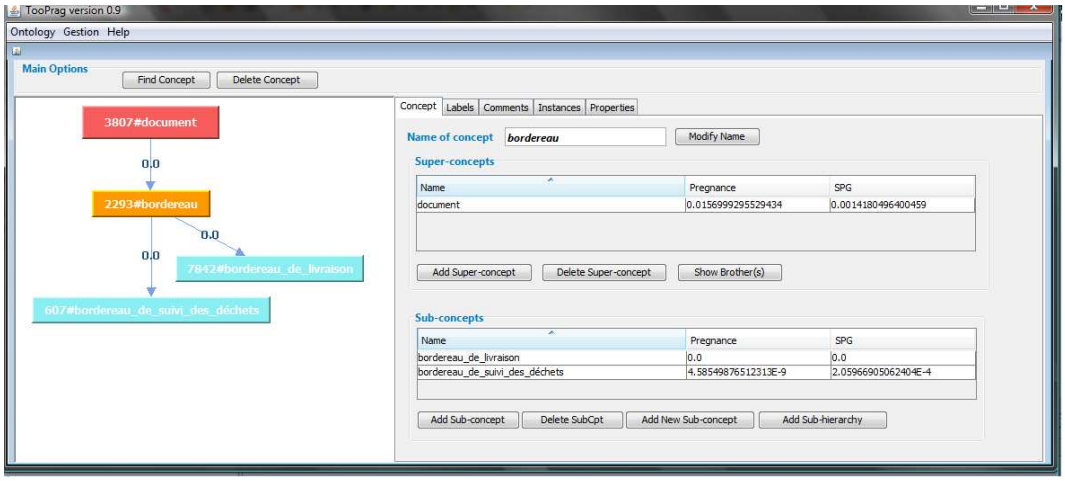


FIGURE 10.3 – Copie d'écran de TOOPRAG.

Pour le calcul de ces gradients, TOOPRAG dispose de deux modes :

- soit une importation d'un index *Lucène*, auquel cas l'utilisateur doit préciser ce qui relève du contenu indexé et ce qui relève de l'adressage des documents ;
- soit une indexation interne à l'application d'un répertoire contenant un corpus de textes.

La première phase du calcul des gradients consiste en une analyse du vocabulaire de l'ontologie. Pour chaque terme, TOOPRAG calcule à partir de l'index le nombre d'occurrences de chaque terme du vocabulaire et le nombre de textes comportant chacun de ces termes. La seconde phase effectue le calcul des gradients de prototypicalité conceptuels comme lexical. A la fin de ce processus, l'application fournit différentes statistiques à l'utilisateur comme le taux de couverture directe ou indirecte de l'ontologie sur un corpus de textes ; c'est à dire le nombre de concepts présents directement par les termes les dénotant dans le corpus, ou indirectement par des concepts de leur descendance. A partir de cette information, il est - par exemple - possible d'établir un nuage de concepts comme on établit un nuage de tags.

Afin de pouvoir stocker en mémoire les valeurs des gradients, et alléger en mémoire l'application dans le cas d'ontologies volumineuses, il a été choisi de travailler au moyen d'une base de données relationnelles dont le schéma est donné par la figure 10.4.
Ce mode de stockage de l'information offre de multiples avantages :

- une capacité importante en terme de volume ;
- l'ajout d'attributs non-présents dans les normes W3C d'OWL (comme les ressources, les gradients...) ;

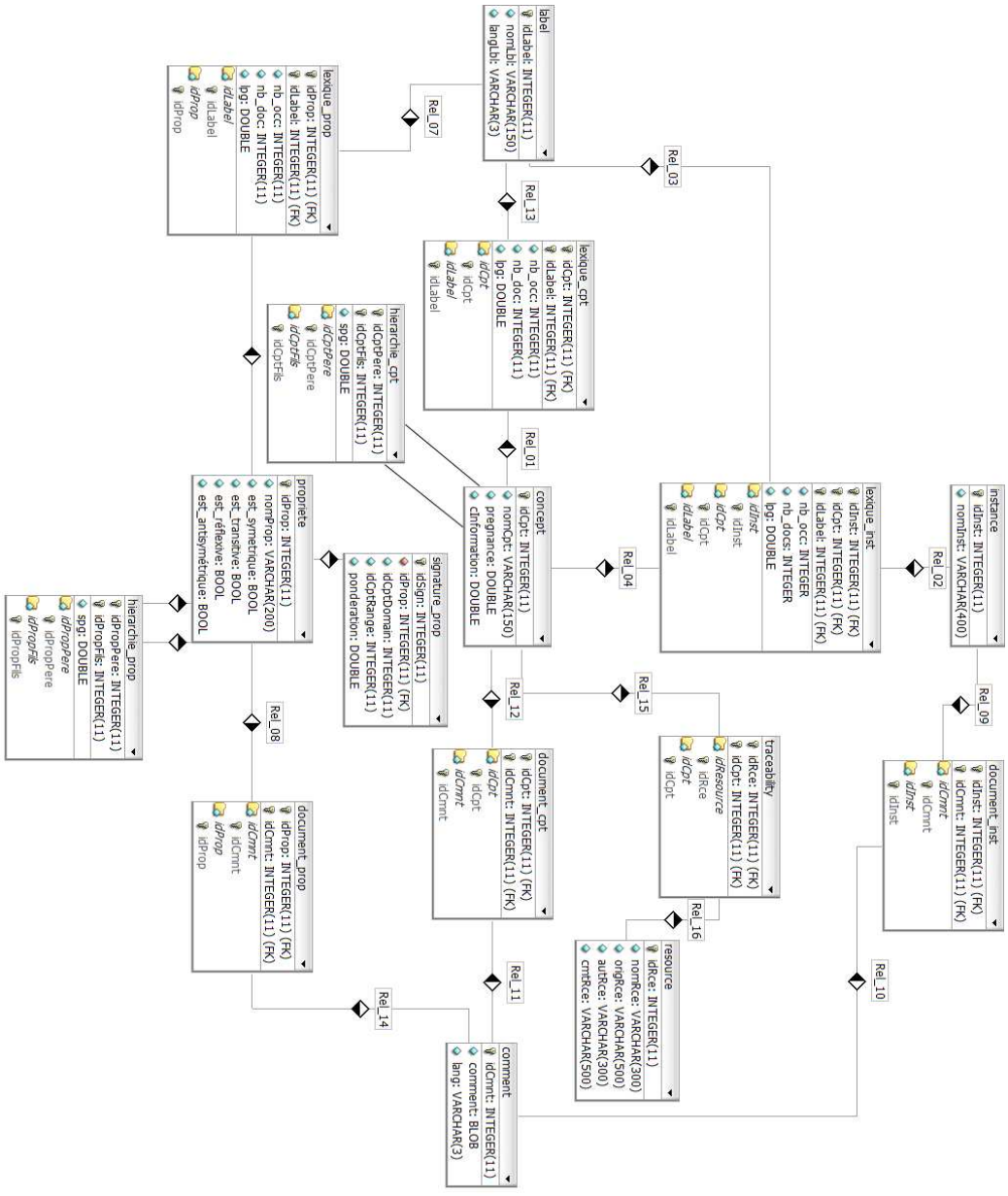


FIGURE 10.4 – Base de données relationnelle de TooPrag.

- le bénéfice de toutes les fonctionnalités de SQL ;
- un gain de temps au démarrage de l'application (l'ontologie n'est jamais en mémoire, l'ouverture de l'ontologie dans TooPrag est une simple connexion à la base).

Ce mode de stockage de l'information présente néanmoins quelques inconvénients :

- l'impossibilité d'utiliser un moteur d'inférence ;
- l'incompatibilité avec les normes W3C d'OWL (comme les ressources, les gradients...) ;
- la difficulté à modéliser certaines informations pour les relations.

TooPrag importe des ontologies au format OWL 1.0 et les stocke dans cette base de données. Il peut également exporter ces ontologies dans ce format, avec la perte de données non présentes dans les spécifications OWL.

10.3 Theseus

Theseus, au sein du module *Veille et Conformité*, est un outil d'extension de requêtes reposant sur l'ontologie HSE-Tennaxia. Ces requêtes étendues sont soumises au filtre Tennaxia, lequel assure la restitution des résultats. Theseus a été développé suivant l'architecture représentée en figure 10.5.

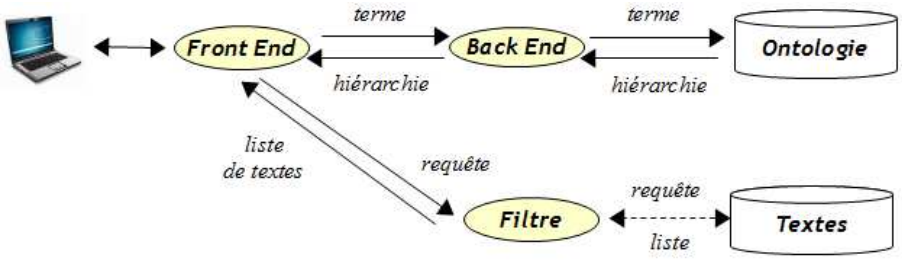


Figure 10.5 – Architecture de Theseus.

Le développement de Theseus a engendré quelques modifications sur le filtre du module *Veille et Conformité* de la suite logicielle de Tennaxia. En effet, afin de pouvoir adopter une approche sémantique, il était nécessaire de pouvoir rechercher des expressions exactes (encadrées par des guillemets, à la manière de Google) et d'étendre par défaut cette recherche tant à la forme singulier que pluriel des termes. Ainsi, par exemple, une recherche sur le terme *Pomme de terre* se fait sur les chaînes de caractères *Pommes de terre* et *Pomme de terre*.
Il a également été nécessaire de modifier le mode d'indexation des textes réglementaires. A l'origine, il était effectué par une indexation par racine qui permettait de retourner toutes les formes dérivées d'un terme. Par exemple, une recherche sur le terme *Halogène* retournait tous les textes contenant les termes *halogène(s)*, mais également *halogénure(s)* ou encore *halogené(e)(s)*. Or, dans le cas d'une extension de requête avec le terme *halogène*, les réponses contenant *halogénures* sont des faux

positifs d'un point de vue sémantique. Un halogène n'est pas un halogénure, et n'est pas non plus un composant halogené. Il a donc fallut utiliser un mode d'indexation par mots exacts, et développer un algorithme d'identification puis de transformation singulier / pluriel de termes (pouvant contenir plusieurs mots) afin d'obtenir avec une même requête des documents contenant les deux variations des termes recherchés.

10.3.1 Principe

Dans un premier temps, lorsqu'un terme t est recherché, l'application parcourt le vocabulaire de l'ontologie jusqu'à trouver une correspondance. Plusieurs cas sont possibles :

- il n'existe aucune correspondance dans l'ontologie, alors THESEUS retourne uniquement le terme saisi, ce qui revient à un usage classique du filtre ;
- il existe une correpondance dans l'ontologie (*i.e.* il existe un unique concept dénoté par au moins ce terme), alors le concept c dénoté par le terme t est associé à la recherche. L'extension est alors calculée sur l'ensemble des concepts de la descendance de c, mais également sur l'ensemble des concepts parents de c, avec pour chaque concept tous les termes les dénotant.
- il existe plusieurs correpondances dans l'ontologie (*i.e.* il existe plusieurs concepts dénotés par au moins ce terme), l'extension est alors calculée séparément sur chacune des correspondances.

L'utilisateur a ensuite la possibilité de sélectionner sur quels concepts il souhaite étendre ou non sa requête. Dans un second temps, une fois ce choix effectué, la requête finale est construite puis soumise au filtre du module *Veille et Conformité*, lequel se charge de la recherche des textes et de leur restitution. THESEUS est compatible avec tout type de moteur de recherche, car sa fonction première est de générer une requête fondée sur le parcours d'une ontologie personnalisée au moyen des gradients de prototypicalité.

Les gradients de prototypicalité interviennent à la fois d'un point de vue lexical et conceptuel. D'un point de vue lexical, ne sont proposés dans l'extension de requête que les termes disposant d'un gradient de prototypicalité lexical au dessus d'un seuil fixé. D'un point de vue conceptuel, la descendance proposée à l'utilisateur est calculée suivant un parcours en "profondeur d'abord" de l'ontologie, avec comme heuristique de parcours une valeur minimale de gradient. Seuls les sur-concepts dont le gradient de prototypicalité conceptuel dépasse un seuil fixé sont également proposés. Ainsi, l'utilisateur possède une vision personnelle de l'ontologie HSE-Tennaxia, qui dépend de l'indexation des textes appartenant à sa base. *De facto*, chaque utilisateur - pour une même requête - peut obtenir des résultats différents.

10.3.2 Frontend

Le frontend est la partie visible de THESEUS. Utilisant les technologies du Web comme les langages HTML, Php (Framework Symphony [3]) et Javascript (composants ExtJs [4]), le frontend assure toute la partie interface avec l'utilisateur. Comme l'illustre la copie d'écran de la figure 10.6, ses fonctions sont :

- zone de saisie, avec auto-complétion fondée sur le vocabulaire de l'ontologie **(1)** ;
- affichage de l'extrait de l'ontologie correspondant au terme recherché **(2)** ;
- affichage du filtre avec les résultats **(3)**.

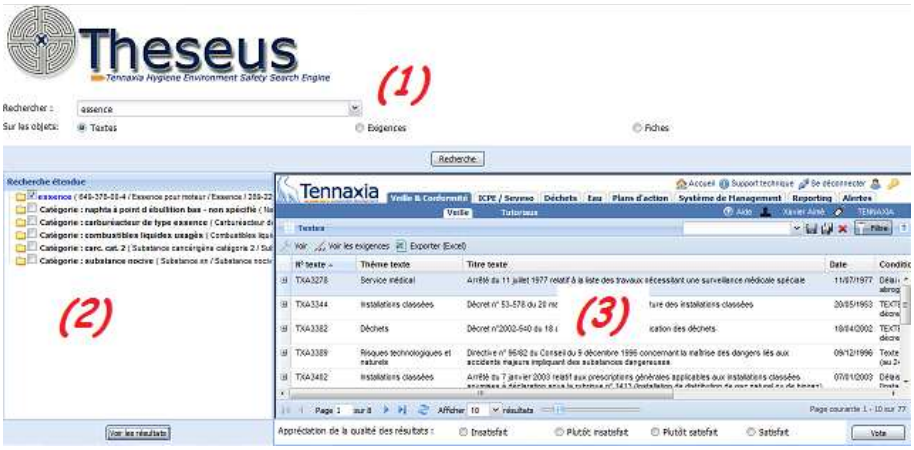


FIGURE 10.6 – Copie écran de Theseus.

Le frontend n'a aucun accès direct avec les données. Son rôle est uniquement dédié à l'affichage et au dialogue avec l'utilisateur, mais aussi à la soumission de la requête finale à un moteur de recherche (ici le filtre de *Veille et Conformité*, mais nous pouvons très bien la soumettre à d'autres moteurs de recherche). L'accès aux données et le parcours de l'ontologie sont réservés au back-end. Le dialogue entre ces deux entités s'établit avec des web-services.

10.3.3 Backend

Programmé en Java et fonctionnant sur un serveur Apache/Tomcat, le rôle du backend est de :

- fournir le vocabulaire de l'ontologie en vue de l'autocomplétion ;
- consulter l'ontologie pour vérifier s'il existe au moins un concept c dénoté par le terme t recherché ;
- retourner une hiérarchie extraite de l'ontologie, ayant pour racine le concept c, avec pour chaque concept l'ensemble des termes les dénotant ;

3. http ://www.symfony-project.org/
4. http ://www.sencha.com/products/js/

– retourner l'ensemble des sur-concepts du concept c.

La hiérarchie retournée est obtenue par un parcours en profondeur d'abord de l'ontologie, en ne prenant comme chemin entre deux concepts que les arcs possédant une valeur de prototypicalité conceptuelle supérieure à une valeur seuil. Il est également effectué un choix des termes en fonction de leur valeur de prototypicalité lexicale.

10.3.4 Résultats sur l'ontologie HSE

Pour illustrer le fonctionnement de Theseus, considérons une recherche sur le concept *Alcane* et comparons les différentes réponses du filtre actuel et des différentes versions de Theseus.

Thème texte	Titre texte
Risques chimiques / Substances particulières	Décret n° 92-1074 du 2 octobre 1992 relatif à la mise sur le marché, à l'utilisation et à l'élimination de certaines substances et préparations dangereuses
Déchets	Règlement du Conseil n° 259/93 du 1er février 1993 concernant la surveillance et le contrôle des transferts de déchets à l'entrée et à la sortie de la Communauté européenne
Déchets	Règlement (CE) n° 1013/2006 du Parlement européen et du Conseil du 14 juin 2006 concernant les transferts de déchets
Risques chimiques / Substances particulières	Arrêté du 7 août 1997 relatif aux limitations de mise sur le marché et d'emploi de certains produits contenant des substances dangereuses
Risques chimiques / Substances particulières	Code de l'Environnement - Articles R521-3 à R521-54 : Mise sur le marché et emploi de certains produits et substances
Déchets	Convention de Bâle sur le contrôle des mouvements transfrontières de déchets dangereux et de leur élimination
Principes généraux	Circulaire du 9 août 1978 relative à la révision du règlement sanitaire départemental type
Pollution atmosphérique	Protocole à la Convention sur la pollution atmosphérique transfrontière à longue distance, de 1979, relatif à la lutte contre les émissions des composés organiques volatils ou leurs flux transfrontières
Risques chimiques / Substances particulières	Règlement (CE) n o 466/2008 de la Commission du 28 mai 2008 imposant aux fabricants et aux importateurs de certaines substances prioritaires de fournir des informations et de procéder à des

Page courante 1 - 9 sur 9

Figure 10.7 – Recherche sur alcane à l'aide du filtre actuel.

A l'heure actuelle, une recherche au moyen du filtre sur le terme *alcane* retourne l'ensemble des textes contenant les mots *alcane* ou *alcanes*. La figure 10.7 indique que nous obtenons neuf textes.
Une recherche sur le terme *alcane*, au moyen de Theseus, retourne en premier un extrait de l'ontologie avec toute la descendance du concept *Alcane* et le sur-concept auquel il est rattaché. La figure 10.8 montre cet extrait de l'ontologie, avec entre parenthèses pour chaque concept la liste des termes dénotant le concept. L'extension de requête à l'ensemble de ces concepts retourne 378 textes (contre neuf dans le cas précédent).
Si nous effectuons la même recherche, en tenant compte des gradients calculés sur un corpus personnalisés de textes réglementaires, nous obtenons l'extrait de l'ontologie de la figure 10.9. Le nombre de concepts proposés est alors très nettement inférieur au cas précédent. De plus, ceux-ci sont classés pour chaque niveau par ordre décroissant de valeur de gradient de prototypicalité conceptuelle. L'extension de requête à l'ensemble de ces concepts retourne 131 textes (contre 378 dans le cas précédent, et 9 dans le premier cas).

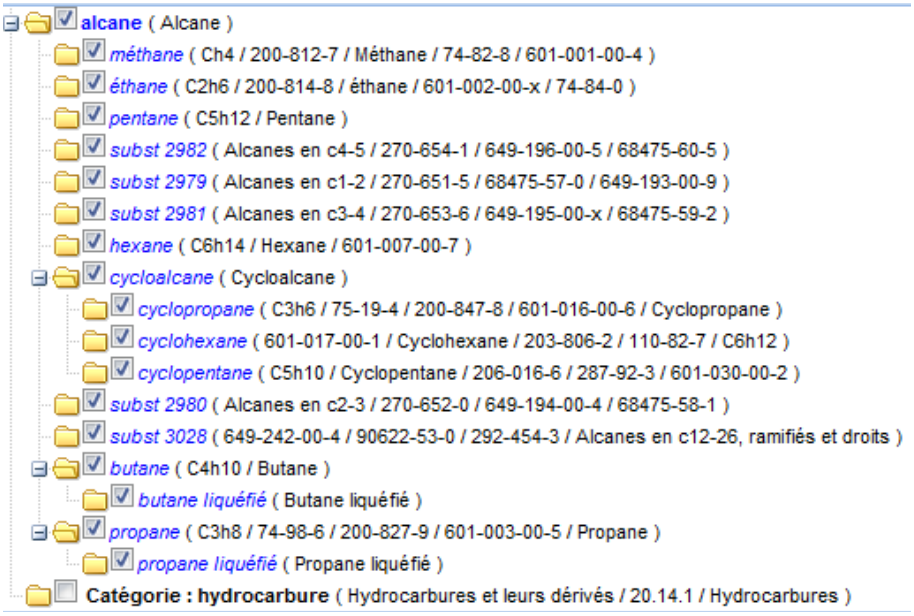


FIGURE 10.8 – Recherche sur alcane à l'aide de THESEUS, sans les gradients.

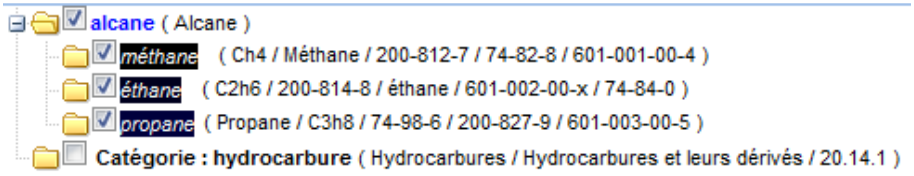


FIGURE 10.9 – Recherche sur alcane à l'aide de THESEUS, avec les gradients.

Ce test a été validé par un ensemble de consultants de la société Tennaxia. Par l'utilisation des gradients de prototypicalité, et d'un point de vue quantitatif, le nombre de documents est nettement supérieur à celui retourné par le simple filtre, mais moindre que si la requête avait été étendue avec l'ensemble des concepts de la descendance. D'un point de vue qualitatif, les utilisateurs disposent d'une quantité d'information gérable, qui répond à leur besoin et leur point de vue. Les principaux indicateurs de Recherche d'Information sont explicités dans la partie 8 de cette thèse.

10.4 Conclusion

Nous venons de présenter deux outils développés dans le cadre de ce projet : (1) TooPrag, destiné à la manipulation d'ontologies volumineuses et au calcul des gradients, (2) Theseus, destiné à l'extension de requêtes à partir d'un terme saisi et en se fondant sur l'ontologie personnalisée HSE-Tennaxia.

Durant les trois années de ce projet, les technologies du Web sémantique et des ontologies ont largement évoluées pour passer - pour certaines - du stade de projet de recherche à celui de véritable produit industriel. La taille des données à gérer a également explosé. Si les premières ontologies contenaient quelques dizaines de concepts, aujourd'hui il n'est pas rare que les entrepôts de données contiennent plusieurs dizaines de milliers de tripets RDF, voir pour certains quelques millions. Si les fonctionnalités de TooPrag pour la gestion d'ontologies volumineuses s'avéraient primordiales il y a encore deux ans, il est clair qu'aujourd'hui la version free de TopBraid, ou encore Neon Toolkit, remplissent pleinement (et même mieux) ce rôle. Ces produits se comportent très bien sur des ontologies, par exemple, comme Agrovoc (près de 50 000 concepts). Ils offrent tous deux, de plus, une certaine aisance tant dans la consultation que dans la gestion de telles bases de connaissances. *A contrario*, le calcul des gradients étant une spécificité de nos travaux, seul TooPrag est à même aujourd'hui de réaliser cette tâche. Il serait néanmoins envisageable, aujourd'hui, de réaliser un plugin Eclipse pour l'intégrer à de tels outils.
Concernant cette partie, la modification récente du mode d'indexation des textes réglementaires a eu une conséquence non négligeable sur l'aspect générique de notre outil. En effet, afin de répondre à certaines spécifications, les développeurs de Tennaxia ont été amené à développé des analyseurs syntaxiques propres à l'entreprise. Or, pour pouvoir consulter l'index Lucène généré et l'utiliser dans le calcul des gradients, TooPrag doit disposer des mêmes analyseurs. Ceci pose donc la question d'intégrer directement à TooPrag le processus d'indexation avec son propre jeu d'analyseurs.

Theseus est encore à l'état de prototype et doit - dans sa forme actuelle - d'avantage être considéré comme un système d'extension de requêtes fondée sur une ontologie personnalisée que comme un moteur de recherche. En effet, il a comme principale

fonctionnalité de composer une nouvelle requête à partir d'un terme saisi, puis de soumettre cette dernière à un moteur de recherche qui propose à l'utilisateur les documents indexés répondant à cette requête avec en plus un tri des résultats par pertinence. Son évolution industrielle, et sa mise en production, dépend de plusieurs facteurs. En premier lieu, le retour sur investissement. Comme évoqué dans le chapitre précédent, l'intérêt de THESEUS repose majoritairement - outre l'aspect personnalisation - sur les côtés qualitatif et quantitatif de l'ontologie utilisée. L'avenir industriel de THESEUS chez Tennaxia est donc fortement lié à la capacité financière et humaine à maintenir et faire évoluer l'ontologie. En second lieu, il y a la place à donner à THESEUS. Aujourd'hui système d'extension de requêtes, THESEUS - pour devenir moteur de recherche - doit gérer tout le processus, en partant de la saisie du terme jusqu'au classement des documents par pertinence. Ceci implique dès lors d'étudier non seulement différentes stratégies d'extensions, mais également l'utilisation des gradients dans la restitution des résultats.

Conclusion

Les ontologies, en tant que représentations consensuelles et formelles de connaissances selon la définition de [Gruber 1993], sont aujourd'hui au cœur de l'Ingénierie des Connaissances et du Web sémantique. Modélisations des connaissances sous la forme de hiérarchies de concepts et de relations, elles sont également le reflet de la perception d'un domaine par un ensemble d'individus autour desquels s'est établi ce consensus.

Contributions

Cette thèse vise à l'introduction de la prototypicalité en Ingénierie des Connaissances. Nous avons défini une nouvelle typologie des ontologies de domaine, et nous avons fourni un cadre sémiotique pour la définition de nouvelles mesures (gradient de prototypicalité et mesures sémantiques) en nous fondant sur les trois dimensions d'une conceptualisation : intension (les propriétés), expression (les termes) et extension (les instances). Nous avons également développé un outil de calcul des gradients de prototypicalité et un moteur de recherche sémantique fondé sur une ontologie de domaine pragmatisée.

La "théorie des prototypes", issue des travaux de psychologie cognitive d'E. Rosch [Rosch 1975], pose l'hypothèse que certaines catégories sont de meilleurs représentants de leur catégorie mère que d'autres (par exemple, *Moineau* est plus représentatif de la catégorie des oiseaux que *Poule*). Toutes les catégories d'une catégorie mère ne sont ainsi pas équivalentes. Partant de ce principe, le prototype pour chacune des catégories est défini comme étant le "meilleur" exemplaire, le plus représentatif, selon un consensus de l'endogroupe concerné. L'application de cette théorie, au sein d'une ontologie de domaine, nous a amené à créer plusieurs gradients de prototypicalité afin d'évaluer, et refléter pour un individu donné, une différence de typicalité entre un concept et différents éléments afférents : l'ensemble de ses sous-concepts (prototypicalité conceptuelle), l'ensemble des termes le dénotant (prototypicalité lexicale) et l'ensemble de ses instances (prototypicalité extensionnelle).

Selon [Koffka 1935], le tout est différent de la somme de ses parties. Tel est l'un des principes de la "théorie de la Gestalt", ou "théorie de la perception". De ce principe découlent six lois. Nous nous sommes intéressés particulièrement à la loi de similarité et à la loi de proximité. La loi de **similarité** permet de regrouper des éléments qui paraissent semblables de manière descriptives (perceptibles) ou fonctionnelles. La loi de **proximité** permet de regrouper des éléments qui apparaissent souvent ensemble, qui sont proches dans une même zone perceptive. À partir de ces deux lois, nous avons développé deux mesures sémantiques distinctes : une mesure

de similarité SemioSem et une mesure de proximité ProxSem. La première se fonde uniquement sur les caractéristiques des concepts (propriétés, termes, instances) indépendamment de la structure de l'ontologie et du corpus de textes. La seconde se fonde sur la représentation simultanée des deux concepts (relations, présence de termes dans un même document, présence simultanée des instances). Afin d'évaluer la validité de cette distinction et de nos mesures, nous avons procédé à une comparaison des valeurs obtenues par calcul avec un jugements humain de similarité et de proximité pour 31 paires de concepts. Ce test se fonde sur les données fournies par le test WordSimilarity-353 [Finkelstein 2002].

D'un point de vue applicatif, nous avons construit une ontologie du domaine HSE à partir de textes réglementaires issus de la base de textes de la société Tennaxia. Nous avons proposé un prototype d'outil de gestion d'ontologies, TooPrag, dont l'objectif est de permettre le calcul de nos gradients de prototypicalité, et manipuler des ontologies volumineuses. Nous avons également proposé un prototype de moteur de recherche sémantique, Theseus, fondé sur l'ontologie de domaine *Tennaxia-HSE* dotée des gradients de prototypicalité, et portant sur la base des textes réglementaires de la société Tennaxia.

Limites

Ontologie Tennaxia-HSE

L'une des contraintes de ce projet, pour la partie construction de l'ontologie, était de se fonder sur un corpus composé exclusivement des textes réglementaires de la base Tennaxia. Nous nous sommes cependant intéressés à d'autres sources (non réglementaires, mais en provenance d'institutions reconnues) afin de savoir s'il était possible de s'en servir comme ressources complémentaires (par exemple, les fiches toxicologiques de l'Institut National de Recherche et de Sécurité pour la prévention des accidents du travail et des maladies professionnelles (INRS) [1]. Plusieurs éléments mettent en cause l'utilisation de telles ressources : non-exhaustivité, problème de la récence des informations, textes non réglementaires.

L'ajout de données en provenance d'autres ontologies peut également être étudié. Ces ontologies sont réalisées par des organismes internationaux comme la Communauté Européenne [2] ou l'ONU [3]. Elles peuvent offrir une extension intéressante tant en terme de quantité (ontologies volumineuses couvrant de multiples domaines) que de qualité (sources réglementaires). Leur mise à jour régulière peut régler le problème de la récence des données. La masse d'information collectée par ces organismes assure une certaine exhaustivité. Enfin, pour la plupart, elles sont bâties à partir de

1. http ://www.inrs.fr/
2. Par exemple, *European Environment Information and Observation Network* : http ://www.eionet.europa.eu/
3. Par exemple, *Food and Agriculture Organization of the United Nations* : http ://www.fao.org/

nomenclatures internationales et autres textes réglementaires.

Deux autres questions se posent en ce qui concerne cette ontologie : (1) celle de sa possible exhaustivité et (2) celle de sa maintenance. Ces questions sont liées en particulier à celle de la nécessité de la présence de tous les concepts dans le corpus de textes réglementaires. En effet, en filigrane se dessine un unique problème : est-il est nécessaire d'avoir tous les concepts présents dans le corpus de textes réglementaires. Actuellement, ce n'est pas le cas, tous les concepts de cette ontologie ne sont pas présents dans le corpus étudié. Ils ont été créés pour répondre à des besoins de conceptualisation. Il pourrait par conséquent s'avérer pertinent de consulter les logs d'utilisation du filtre de l'outil *Veille et Conformité* afin de connaître un peu mieux les centres d'intérêt des utilisateurs. Cela reviendrait à opérer une complétion de l'ontologie guidée par les usages.

Gradients de prototypicalité et mesures sémantiques

Il existe principalement trois limitations à notre approche :

1. La première réside dans l'affectation de valeurs aux coordonnées cognitives (α, β, γ). Pour le moment, ces coordonnées sont attribuées de manière empirique. Il pourrait être envisagé de lier profil psychologique et coordonnées cognitives au moyen de questionnaires, et ce afin de déterminer si l'individu a une vision plus intensionnelle, extensionnelle ou expressionnelle.
2. La deuxième réside dans la pondération des propriétés pour le calcul des composantes intensionnelles. Il n'est, pour le moment, pas possible d'automatiser une telle tâche. Chronophage, il faut envisager cette affectation de poids de manière collaborative au niveau de l'endogroupe.
3. La troisième limitation concerne le calcul des composantes expressionnelles, qui repose sur la recherche d'occurrences de termes au sein de corpus. Ce type de recherche se heurte aux problèmes liés au traitement de la langue naturelle. Citons, entre autre, le cas des anaphores, des ambiguïtés, etc. Ainsi, un terme peut-être largement sous-estimé quantitativement.

Perspectives

Nos travaux, réalisés dans le cadre d'un contrat CIFRE, ont suivi deux directions complémentaires : (1) une direction de recherche avec la proposition de gradients de prototypicalité et de nouvelles mesures sémantiques, et (2) une direction applicative dans un contexte industriel avec le développement d'un prototype de moteurs de recherche sémantique fondée sur une ontologie de domaine. Notre réflexion sur la suite à donner à ces travaux peut se structurer autour de trois grandes problématiques :

1. l'amélioration de la caractérisation de la prototypicalité en tenant compte conjointement des relations concept←sous-concepts (prototypicalité concep-

tuelle descendante) et des relations concept→sur-concepts (prototypicalité conceptuelle ascendante) ;

2. l'intégration des gradients de prototypicalité dans le processus de personnalisation ;
3. l'intégration de ces gradients en recherche d'information.

Prototypicalités conceptuelles ascendante et descendante

Le gradient de prototypicalité conceptuelle que nous proposons est utilisé pour mesurer la typicalité d'un sous-concept par rapport à l'un de ses sur-concepts [Aimé 2010b, Aimé 2010c, Aimé 2009b]. Nous qualifions ce degré d'attachement de *gradient de prototypicalité descendante*. Il est également possible de mesurer la typicalité d'un concept par rapport à l'un de ses sous-concepts. Nous qualifions ce degré d'attachement de *gradient de prototypicalité ascendante*.

D'un point de vue intensionnel, plus un concept est proche en terme de distance du prototype de l'un de ses sur-concepts, plus ce sur-concept en est représentatif. Cela revient à dire que plus un élément partage de propriétés avec son sur-concept, plus ce dernier en est caractéristique. Ensuite, d'un point de vue expressionnel, au moyen des prégnances, plus un concept fournit de contenu en information à son sur-concept, plus il est proche de celui-ci. Cela revient à dire que, quantitativement au sein d'un corpus, plus un élément participe à la dénotation d'un concept, plus ce dernier en est caractéristique. Enfin, d'un point de vue extensionnel, plus un concept fournit d'instances à l'un de ses sur-concepts pères, plus il en est proche. Cela revient à dire que, quantitativement, plus un élément contribue à la représentation d'un sur-concept, plus ce dernier en est caractéristique.

Le fait que ces trois composantes mesurent un degré d'apport d'un concept à un sur-concept nous permettrait de pouvoir utiliser ce gradient de prototypicalité conceptuelle aussi bien entre un concept et ses sous-concepts qu'entre ce concept et les sur-concepts auxquels il est rattaché.

Personnalisation

Ce processus de personnalisation, qui à partir d'une même OVD peut conduire à plusieurs OPVD [Aimé 2009a, Aimé 2008a], est fondé sur l'apport de ressources supplémentaires :

- un ensemble d'**instances** supposées représentatives de l'univers cognitif de l'utilisateur (dans le cas d'un système d'information commercial, par exemple, ces instances seront les clients traités par l'utilisateur, les produits qu'il leur vend, etc) ;
- un **corpus** fourni par l'utilisateur et supposé représentatif de son univers cognitif (ce corpus peut, par exemple, être tiré de documents numériques écrits par l'utilisateur sur un blog ou un wiki sémantique) ;

– des **pondérations portant sur les propriétés** de chaque concept et qui expriment l'importance que l'utilisateur accorde à cette propriété dans la définition de ce concept. Ces pondérations sont fixées par l'utilisateur de la façon suivante : pour chaque propriété p de $\mathcal{P}$, l'utilisateur ordonne sur une échelle de 0 à 1 les concepts qui font partie du domaine de p, selon qu'il associe plus ou moins ce concept à cette propriété. Par exemple, pour la propriété "a un auteur", on mettra en premier le concept d'article scientifique, puis celui d'article de presse (pour lequel cette propriété est moins importante), puis celui de mode d'emploi d'aspirateur (pour lequel cette propriété est encore moins importante). D'une certaine façon, l'utilisateur classe les concepts en fonction de leur prototypicalité par rapport à une propriété qu'ils partagent : pour une propriété donnée, il s'agit de savoir quel concept cette propriété évoque le plus souvent.

Ces ressources de personnalisation ne sont pas forcément toujours disponibles, mais la méthode proposée fonctionne même si aucune autre ressource que l'ontologie n'est disponible (dans ce cas, le calcul des prototypicalités ne permet plus une personnalisation, mais constitue un enrichissement de l'ontologie). Si l'utilisateur ne souhaite pas pondérer les propriétés, les poids sont tous fixés à 1. Baser le processus sur trois ressources différentes offre ainsi plusieurs façons de personnaliser le système.

Recherche d'information

Aujourd'hui, la majeure partie des moteurs de recherche sémantique étendant leur requête à partir d'une ontologie fournit un résultat identique quelque soit l'utilisateur. La personnalisation d'une ontologie de domaine au moyen des gradients de prototypicalité est une première étape dans la personnalisation d'une recherche d'information ([Aimé 2008b]). Parallèlement, il est possible d'utiliser des ontologies de domaine pour étendre des requêtes au sein d'un système de recherche d'information sémantique (*e.g.* [Guelfi 2007, Messai 2006]). Il serait intéressant de développer différentes stratégies d'extension de requêtes, et donc différentes stratégies de parcours de l'ontologie suivant les valeurs de gradients de prototypicalité conceptuelle et lexicale [Aimé 2010a, Aimé 2008c, Trichet 2010].

ANNEXES

—

Annexe A : *Ontologie HSE-Tennaxia*

Annexe A

Ontologie HSE-Tennaxia

A.1 Ressources utilisées

A.1.1 EUROVOC

Eurovoc [1] est un thésaurus multilingue [2] couvrant tous les domaines de l'activité de l'Union Européenne. Ce thésaurus a pour objectif de permettre l'indexation des documents dans les systèmes documentaires des institutions européennes et de leurs utilisateurs. Eurovoc est actuellement utilisé par le Parlement Européen, l'Office des publications des Communautés européennes, les parlements nationaux et régionaux en Europe, des administrations nationales et par certaines organisations européennes. La version utilisée pour cette ontologie, avec une licence spéciale recherche, est la 4.2. Il est par ailleurs intéressant de noter que ce thésaurus a intégré une dimension collaborative dans sa construction, puisque l'Office des publications propose aux internautes de participer à son élaboration sous la forme de soumission de nouveaux termes, ou de correctifs.

Ce thésaurus couvre 21 domaines intéressant les activités des Institutions européennes : vie politique, relations internationales, communautés européennes, *droit*, vie économique, échanges économiques et commerciaux, finances, questions sociales, éducation et communication, *sciences*, entreprise et concurrence, emploi et travail, transports, *environnement*, agriculture, sylviculture et pêche, agro-alimentaire, production, technologie et recherche, *énergie*, *industrie*, géographie, et organisations internationales.

D'une manière détaillée, le thésaurus EUROVOC comporte 127 micro-thesaurii, 6.645 descripteurs (dont 519 top terms), 6.669 relations hiérarchiques réciproques (BT/NT) et 3.636 relations associatives réciproques (RT). Il est entendu par relation

1. http ://europa.eu.int/celex/eurovoc/
2. Eurovoc existe dans les 11 langues officielles de l'Union Européenne : allemand, anglais, danois, espagnol, finnois, français, grec, italien, néerlandais, portugais et suédois. En plus de ces versions, Eurovoc a été traduit par les parlements nationaux de plusieurs pays comme l'Albanie, la Croatie, la Lettonie, la Lituanie, la Pologne, République tchèque, la Roumanie, la Russie, la Slovaquie et la Slovénie. Soit un total de plus d'une vingtaine de langues.

d'association (RT) une relation de type causalité, instrumentation, concomitance, succession dans le temps ou dans l'espace, constituant, localisation, similarité ou antinomie (il n'y a aucune précision, tous ces types sont englobés dans ce type d'association). Un concept est nommé par au moins un descripteur (avec possibilités de synonymes).

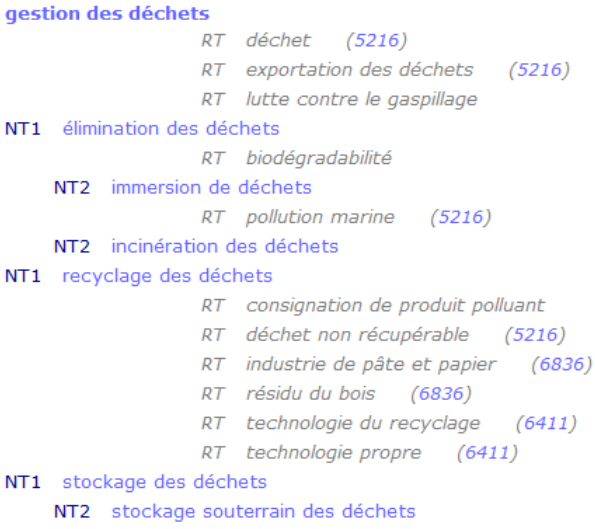


FIGURE A.1 – Extrait de la hiérarchie proposée par le thésaurus Eurovoc.

La figure A.1 présente un extrait de ce thésaurus. Nous y trouvons une hiérarchie conceptuelle en rapport avec la gestion des déchets. Nous avons ainsi trois top termes : *élimination* (avec comme sous-concepts *incinération* et *immersion*), *recyclage* et *stockage* (avec comme sous concept *stockage souterrain).* Chaque concept, ou presque, possède au moins une relation d'association avec un autre concept. Par exemple, *élimination des déchets* est lié à *biodégrabilité* (lien de similarité), ou encore *immersion des déchets* à *pollution marine* (lien de causalité).

Eurovoc est utilisé comme ressource pour la construction de notre ontologie, avec comme avantage la fourniture de synonymes issus des textes réglementaires de l'Union Européenne (qui plus est en multilingue), et la fourniture d'amorces ou d'extraits de hiérarchies conceptuelles. Par contre, il possède comme inconvénient le fait de ne pas contenir de liens hiérarchiques rigoureusement de type *is-a*, et de ne pas spécifier exactement le type de lien d'association (problème avec les RT), la sémantique des relations étant laissée à l'interprétation de l'utilisateur.

Ce thésaurus nous a servi essentiellement de jeux de tests, et de (re)source d'inspiration pour notre modélisation.

A.1.2 Nomenclature des Installations Industrielles Classées (IC)

En France, comme dans de nombreux pays industrialisés par ailleurs, le contrôle de la prévention des pollutions et risques industriels et agricoles est géré par l'Etat. L'introduction de la nomenclature IC [Min 2007] nous apporte quelques précisions : " *la nomenclature des installations classées, prévue par l'article L. 511-2 du code de l'environnement est fixée, en application de l'article 40 du décret du 21 septembre 1977, par le décret du 20 mai 1953 dans son annexe I. Celui-ci a été modifié à de nombreuses reprises, et notamment depuis 1992, date à laquelle une profonde refonte de la nomenclature a été entreprise, en introduisant de nouvelles rubriques (...).* "

Classement par activités	Classement par substances
21xx. Activités agricoles, animaux	10xx.Substances et préparations
22xx. Agroalimentaire	11xx. Toxiques
23xx. Textiles, cuirs et peaux	12xx. Susbtances comburantes
24xx. Bois, papier, carton, imprimerie	13xx. Explosifs et substances explosibles
25xx. Matériaux, minerais et métaux	14xx. Substances Inflammables
26xx. Chimie, parachimie, caoutchouc	15xx. Produits combustibles
27xx. Déchets	16xx. Corrosifs
29xx. Divers	17xx. Substances radioactives
	18xx. Réagissant avec l'eau

FIGURE A.2 – Nomenclature IC.

Le site web de l'INERIS[3] possède un service gratuit, AIDA, lequel recense en un même point l'ensemble des textes réglementaires relatifs aux installations industrielles classées. L'accès à ces textes peut se faire de deux manières, soit à l'aide de la nomenclature IC (rubriques liées aux substances et aux types d'activités industrielles, *cf.* figure A.2), soit par un classement transversal (air, eau, bruit, déchets, etc.)

La nomenclature IC est une classification hiérarchique sur quatre niveaux : le premier sépare activités et substances, le deuxième détaille les grands types d'activités et de substances, dans le troisième et le quatrième apparaissent les substances et activités elles-mêmes.

Les entrées dans la nomenclature IC - environ 200 - représentent autant de concepts utilisés par les experts pour catégoriser les textes et structurer le domaine, et doivent donc apparaître comme des notions centrales dans l'ontologie [Després 2007].

Nous utiliserons cette nomenclature conjointement à celle des activités et produits d'activités présentée dans la section suivante.

3. http ://www.ineris.fr/

A.1.3 Nomenclature des Activités et Produits d'activités

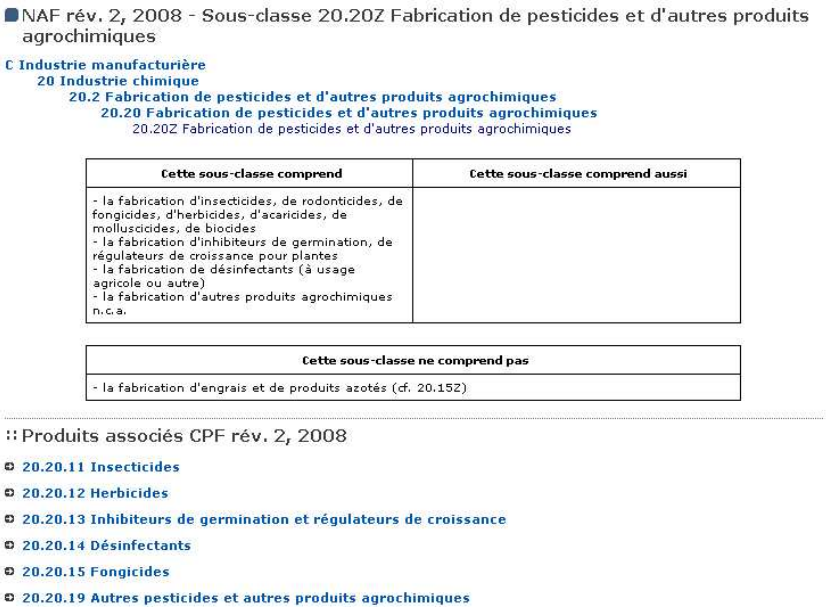
Il s'agit d'un document rédigé par l'Institut National de la Statistiques et des Études Économiques (INSEE)[4], sous l'égide de la Commission nationale des nomenclatures économiques et sociales (CNNES) du Conseil national de l'information statistique (CNIS). Ces nomenclatures, en vigueur depuis le 1er janvier 2003, ont fait l'objet d'une révision au 1er janvier 2008 (version utilisée pour l'élaboration de notre ontologie). Cette opération s'inscrit dans un processus de révision d'ensemble des nomenclatures d'activités et de produits aux niveaux mondial, européen et français. De ce fait, il existe aujourd'hui une cohérence garantie par l'emboitement des nomenclatures à ces trois échelles. Il a ainsi été créé deux nomenclatures *mères* au niveau international (Classification internationale type des industries (CITI) révision 4 et Classification Centrale des Produits (CPC) version 2), des nomenclatures *européennes* cohérentes avec les deux nomenclatures internationales (Nomenclature d'Activités des Communautés Européennes (NACE) révision 2 et Classification des Produits associée aux Activités (CPA) 2008 pour l'Europe) et des nomenclatures nationales (dans le cas de l'Union Européenne, ces dernières sont strictement emboîtées dans les nomenclatures européennes). Le dispositif central français comporte deux nomenclatures concernant respectivement les activités et les produits :

- la Nomenclature d'Activités Française (NAF) avec cinq niveaux (comportant respectivement 21 sections, 88 divisions, 272 groupes, 615 classes et 732 sous-classes). Son objectif est le classement des différentes activités économiques, *i.e.* les activités socialement organisées en vue de la production de biens ou de services[5]. Selon l'INSEE, il y a activité économique lorsque des ressources - telles que des biens d'équipement, de la main-d'oeuvre, des techniques de fabrication ou des produits intermédiaires - sont combinées pour produire des biens ou des services spécifiques. Toute activité est caractérisée par une entrée de ressources, un processus de production et une sortie de produits (biens ou services).
- la Classification des Produits Française (CPF) avec cinq niveaux (comportant respectivement 21 sections, 88 divisions, 261 groupes, 575 classes, 1342 catégories et 3142 sous-catégories). Son objectif est le classement des biens ou des services issus des activités économiques (ou dégradés lors de leur utilisation).

L'utilisation de ces nomenclatures est précisée dans l'article 4 - alinéa III du décret n°2007-1888 du 26 décembre 2007 portant approbation des nomenclatures d'activités et de produits : " *ces nomenclatures (et leurs adaptations éventuelles) sont utilisées dans les textes officiels, décisions, documents, travaux et études ainsi que dans les systèmes informatiques des administrations et établissements publics et dans les travaux effectués par des organismes privés à la demande des administrations.* "

4. http ://www.insee.fr

5. Ne sont donc pas concernés les actes économiques s'analysant comme un transfert de revenu (versement d'intérêt à un prêteur, par exemple) ou une opération financière (émission d'un emprunt par exemple) ni les actions qui ne relèvent pas de la sphère économique.



NAF rév. 2, 2008 - Sous-classe 20.20Z Fabrication de pesticides et d'autres produits agrochimiques

C Industrie manufacturière
20 Industrie chimique
20.2 Fabrication de pesticides et d'autres produits agrochimiques
20.20 Fabrication de pesticides et d'autres produits agrochimiques
20.20Z Fabrication de pesticides et d'autres produits agrochimiques

Cette sous-classe comprend	Cette sous-classe comprend aussi
- la fabrication d'insecticides, de rodonticides, de fongicides, d'herbicides, d'acaricides, de molluscicides, de biocides - la fabrication d'inhibiteurs de germination, de régulateurs de croissance pour plantes - la fabrication de désinfectants (à usage agricole ou autre) - la fabrication d'autres produits agrochimiques n.c.a.	

Cette sous-classe ne comprend pas
- la fabrication d'engrais et de produits azotés (cf. 20.15Z)

:: Produits associés CPF rév. 2, 2008

- 20.20.11 Insecticides
- 20.20.12 Herbicides
- 20.20.13 Inhibiteurs de germination et régulateurs de croissance
- 20.20.14 Désinfectants
- 20.20.15 Fongicides
- 20.20.19 Autres pesticides et autres produits agrochimiques

FIGURE A.3 – Extrait de la nomenclature NAFE - Sous-classe 20.20Z Fabrication de pesticides.

La figure A.3 donne un exemple de la nomenclature NAFE. Nous avons en sous-classe l'activité *Fabrication de pesticides et d'autres produits agrochimiques* avec comme précision :

- cette sous-classe est fille de *Industrie chimique*, elle-même sous-classe de *Industrie manufacturière*, etc. Nous pouvons utiliser cette information dans la construction de l'ontologie pour bâtir une hiérarchie des activités.
- cette sous-classe *comprend également* la fabrication d'insecticides, de rotondicides, d'herbicides, etc. Nous pouvons utiliser cette information dans la construction de l'ontologie (1) soit pour créer un niveau supplémentaire à raison d'un concept-fils par libellé, (2) soit pour compléter la liste des termes usités pour dénoter ce concept.
- cette sous-classe *ne comprend pas* la fabrication d'engrais et de produits azotés (20.15). Nous utiliserons cette information pour établir des disjonctions entre les concepts.
- l'activité décrite par cette sous-classe *produit* des insecticides, ainsi que des opérations de sous-traitance dans l'élaboration de pesticides. Nous pouvons utiliser cette information dans la construction de l'ontologie pour créer une relation entre l'activité et le(s) produit(s) généré(s).

La figure A.4 donne un exemple de la nomenclature CPF. Nous avons en sous-classe l'activité *Fongicides* avec comme précision :

- cette sous-classe est fille de *Pesticides et autres produits chimiques*, elle-même sous-classe de *Produits chimiques*, etc. Nous pouvons utiliser cette informa-

FIGURE A.4 – Extrait de la nomenclature CPF - Sous-catégorie 20.20.15 Fongicides.

tion dans la construction de l'ontologie pour bâtir une hiérarchie des produits d'activité.

– le produit décrit par cette sous-classe *est produite par* l'activité *Fabrication des pesticides et autres produits agrochimiques*. Nous pouvons utiliser cette information dans la construction de l'ontologie pour créer une relation entre un produit et l'activité qui le génère.

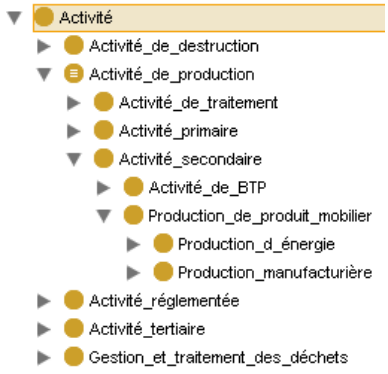


FIGURE A.5 – Extrait de la hiérarchie des activités.

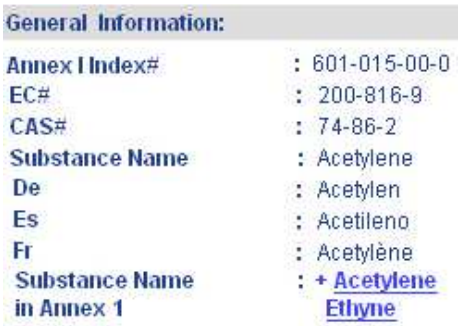
De manière générale, la NAF ne fait pas de distinction entre les activités marchandes et les activités non marchandes, telles qu'elles sont définies dans les systèmes de comptabilité nationale, même si la distinction est importante dans ces systèmes. De plus, les activité décrites dans la nomenclature ICPE possèdent une structure hiérarchique différente de celle donnée par la NAFE. Nous avons donc de bâtir une structure hiérarchique en combinant les connaissances exprimées dans ces deux sources, ce qui nous a conduit à créer 3.000 concepts dans l'ontologie HSE-Tennaxia (*cf.* figure A.5).

A.1.4 Annexe I de la version consolidée de la Directive 67/548/CEE

L'annexe I de la version consolidée de la Directive 67/548/CEE constitue, aujourd'hui, le répertoire des substances dangereuses pour lesquelles une classification

et un étiquetage harmonisés ont été convenus par l'ensemble des pays membres de la Communauté européenne. La directive *67/548/CEE* du Conseil de l'Europe a été modifiée neuf fois [6] et adaptée vingt-neuf fois au progrès technique [7]. À ce jour, il n'en existe pas de version consolidée. C'est pourquoi la Commission Européenne a élaboré, à titre officieux et aux seules fins d'information, une version consolidée de la directive 67/548/CEE comprenant la directive 92/32/CEE [8] et la vingt-huitième adaptation des annexes II à IX au progrès technique. Cette dernière comprend les annexes suivantes :

- **Annexe I** présentant la classification et l'étiquetage harmonisés des 8000 substances chimiques ;
- **Annexe II** décrivant les quinze symboles de danger servant à classer les substances dangereuses (explosif, très toxique, dangereux pour l'environnement, etc.) ;
- **Annexe III** présentant les phrases types (*phrases R*) indiquant la nature des risques particuliers des substances chimiques ;
- **Annexe IV** énumérant les phrases types indiquant les conseils de prudence (*phrases S*) en ce qui concerne la manipulation et l'utilisation de substances dangereuses ;
- **Annexe V** comportant les méthodes de détermination des propriétés dangereuses des substances ;
- **Annexe VI** présentant en détaille les critères permettant de choisir la catégorie de danger adéquate et d'affecter symboles de danger, phrases R et phrases S à une substance répertoriée ;
- **Annexe VII** et l'**annexe VIII** concernant la notification de nouvelles substances (informations devant figurer dans le dossier de notification) ;
- **Annexe IX** comportant les dispositions relatives aux fermetures de sécurité pour les enfants et les indications tactiles de danger, comme des emballages et des éléments d'étiquetage spécifiques.



General Information:	
Annex I Index#	: 601-015-00-0
EC#	: 200-816-9
CAS#	: 74-86-2
Substance Name	: Acetylene
De	: Acetylen
Es	: Acetileno
Fr	: Acetylène
Substance Name in Annex 1	: + Acetylene Ethyne

FIGURE A.6 – Extrait des informations concernant l'acétylène.

6. La neuvième modification est la directive 1999/33/CE
7. Journal Officiel L 216 du 16.6.2004, p. 3.
8. Il s'agit de la septième modification de la directive 67/548/CEE.

La figure A.6 nous indique les différents types d'entrée pour une substance. En premier lieu, nous avons un numéro d'index (*Annex I Index*) présenté sous la forme d'une séquence chiffrée du type *ABC-DEF-GH-Y*, où :

- *ABC* représente soit le numéro atomique de l'élément chimique le plus caractéristique (précédé d'un ou de deux zéros pour compléter la sous-séquence), soit le numéro conventionnel de la classification des substances organiques ;
- *DEF* représente le numéro progressif des substances considérées dans les séquences ABC ;
- *GH* représente la forme sous laquelle la substance est produite ou mise sur le marché ;
- et *Y* représente le chiffre de contrôle (check digit) calculé selon la méthode utilisée par l'ISBN (International Standard Book Number).

Un deuxième numéro (*EC*), sous la forme d'une suite de sept chiffres du type XXX-XXX-X, est également mentionné. Il s'agit :

- pour les substances dangereuses reprises dans l'*inventaire européen des produits chimiques commercialisés* (EINECS)[9], d'un numéro commençant par 200-001-8 ;
- pour les substances dangereuses répertoriées dans la *liste européenne des substances notifiées* (ELINCS), d'un numéro commençant par 400-010-9 ;
- pour les substances dangereuses figurant dans la liste des « *Ex-polymères* »[10], d'un numéro commençant par 500-001-0.

Une troisième référence est également mentionnée avec le numéro d'enregistrement unique de la substance auprès de la banque de données de *Chemical Abstracts Service* (CAS). Cette banque de données (contenant plus de 30 millions de références) recense et enregistre chaque produit chimique, mais également polymère, séquence biologique et alliage décrit dans la littérature. Ce numéro unique, attribué dans une ordre croissant et n'ayant pas de signification particulière, facilite l'identification de l'entrée. Les numéros CAS se composent de trois parties séparées par un tiret. Il est à noter que ces numéros référencent de manière très précise les substances. En effet, le numéro Einecs désigne à la fois la forme anhydre et les formes hydratées d'une substance alors qu'il existe des numéros CAS différents pour chacune des dites formes.

Chaque substance est référencée avec son nom dans les différentes langues usitées au sein des différents pays membres de la Commaunauté européenne, mais également les différentes dénominations possibles dans une même langue (comme indiqué dans la figure A.6, *Acétylène* et *Ethyne* pour la molécule de formule C_2H_2).
Chaque substance est placée dans une ou plusieurs catégories de danger (*cf.* figure A.7) se présentant sous la forme d'une abréviation représentant la catégorie

9. Journal Officiel n°C 146 A du 15.6.1990.
10. No-longer polymers - Document de l'Office des publications officielles des Communautés européennes, 1997, ISBN 92-827-8995-0.

FIGURE A.7 – Extrait des risques et recommandations concernant l'acétylène.

de danger, accompagnée d'une ou plusieurs phrase(s) de risque (phrases R). Les abréviations utilisées dans les différentes catégories de danger sont les suivantes :

- *explosif* **E** ;
- *comburant* **O** ;
- *extrêmement inflammable* **F+** ;
- *facilement inflammable* **F** ;
- *inflammable* **R 10** ;
- *très toxique* **T+** ;
- *toxique* **T** ;
- *nocif* **Xn** ;
- *corrosif* **C** ;
- *irritant* **Xi** ;
- *sensibilisant* **R 42 et/ou R 43** ;
- *cancérogène* **Carc. Cat. (1)** ;
- *mutagène* **Muta. Cat. (1)** ;
- *toxique pour la reproduction* **Repr. Cat. (1)** ;
- *dangereux pour l'environnement* **N** et/ou **R 52, R 53, R 59**.

Chaque substance possède également des conseils de prudence (phrases S), désignés par une série de chiffres précédés de la lettre S indiquant les précautions d'emploi.

Dénomination chimique	Numéro CE	Numéro CAS	Classification	Etiquetage	Limites de concentration
phosgène	200-870-3	75-44-5	T+; R26 C; R34	T+ R: 26-34 S: (1/2-)9-26-36/37/39-45	C ≥ 5 %: T+; R26-34 1 % ≤ C < 5 %: T+; R26-36/37/38 0,5 % ≤ C < 1 %: T; R23-36/37/38 0,2 % ≤ C < 0,5 %: T; R23 0,02 % ≤ C < 0,2 %: Xn; R20

FIGURE A.8 – Extrait de l'annexe I pour la substance phosgène.

Chaque substance possède des indications concernant les limites de concentrations (*cf.* figure A.8). Ces limites de concentration sont des pourcentages en poids de la substance calculés par rapport au poids total de la préparation. Lorsque aucune li-

mite de concentration n'est indiquée, les limites à utiliser pour appliquer la méthode conventionnelle d'évaluation des dangers pour la santé sont celles figurant à l'annexe II et les limites à utiliser pour appliquer la méthode conventionnelle d'évaluation des dangers pour l'environnement sont celles figurant à l'annexe III de la directive 1999/45/CE.

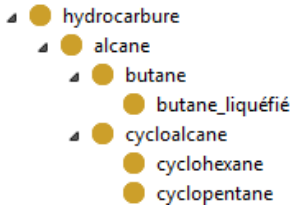


FIGURE A.9 – Extrait de la hiérarchie des substances.

La partie de l'ontologie HSE-Tennaxia relative aux substances, dont l'extrait apparaît figure A.9, a été élaborée essentiellement à partir des données contenues dans l'annexe I de la directive 67/548/CEE. Trois sous-hiérarchies sont créées :

- phrases de recommandations (correspondant aux phrases S) ;
- phrases de risques (correspondant aux phrases R) ;
- substances chimiques.

FIGURE A.10 – Extrait de la hiérarchie des substances chimiques.

Les substances chimiques (*cf.* figure A.10) se divisent en trois groupes :

- les *atomes*, éléments simples obtenus à partir de la classification périodique des éléments de Mendeleïev ;
- les *composés chimiques*, composés à base d'éléments simples et classés suivant l'élément chimique le plus caractéristique ou par famille de substances organiques (conformément à la version consolidée de l'annexe I de la directive 67/548/CEE) ;
- les *substances dangereuses*, classés par type de risque (conformément à la version consolidée de l'annexe I de la directive 67/548/CEE).

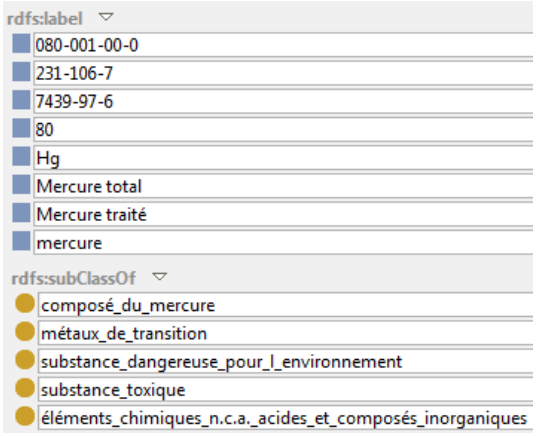


FIGURE A.11 – Extrait d'informations concernant le Mercure.

Pour une substance donnée, par exemple le mercure (figure A.11), nous avons les informations suivantes :

- une liste de labels pour dénoter le concept comprenant les références aux différents classements (CAS, EINECS, etc.), les multiples dénominations en langue française, le symbole chimique et son numéro de classification dans le cas d'une élément simple ;
- les différentes catégories de rattachement (avec héritage multiple : type de composé, type de substance, familles de risque, etc.).

Cette ressource, de par sa nature semi-structurée, a nécessité assez peu de post-traitements. Elle fournit environ 4.000 concepts, soit un peu moins de 50% de notre ontologie sur une profondeur de 7 et une largeur maximale d'environ 150.

A.1.5 Liste des pathologies et éléments pathogènes

La notion de maladie professionnelle[11] remonte à 1919, avec la loi du 25 octobre de cette année-là, loi fixant le fait qu'une maladie peut être reconnue comme maladie professionnelle si elle figure sur l'un des tableaux annexés au Code de la

11. Selon [Abadia 2007], une maladie est dite *professionnelle* si elle est la conséquence directe de l'exposition d'un travailleur à un risque physique, chimique, biologique, ou résulte des conditions dans lesquelles il exerce son activités professionnelle.

Sécurité sociale. Ces tableaux sont créés et modifiés par décret au fur et à mesure de l'évolution des techniques et des progrès des connaissances médicales. A ce jour, ce code compte 117 tableaux du régime général et 65 du régime agricole. La nature structurée d'une telle ressource n'engendre quasimment aucun post-traitement et peut être intégrée telle quelle. Elle génère environ un millier de concept sur une profondeur de 5 et une largeur maximale de 200.

Chacun de ces tableaux comporte de manière limitative les éléments suivants :
- les **symptômes** ou **lésions pathologiques** ;
- le délai de prise en charge (délai maximal entre la date à laquelle le travailleur a cessé d'être exposé au risque et la constatation de l'affection) ;
- les causes susceptibles de provoquer l'affection.

Le Code de la Sécurité Sociale recense ainsi 13 types de pathologies :

1. les pathologies bronco-pulmonaires ;
2. les pathologies cardiaques et vasculaires ;
3. les pathologies cutanées et muqueuses ;
4. les pathologies digestives, gastro-intestinales et hépatiques ;
5. les maladies infectieuses et parasitaires ;
6. les intoxications aiguës ;
7. les pathologiques neurologiques, musculaires et psychiatriques ;
8. les pathologies de l'oeil et de la vision ;
9. les pathologies ORL et stomatologique ;
10. les pathologies osseuses, articulaires et périarticulaires ;
11. les pathologies rénales, vésicales et génitales ;
12. les pathologies du sang et des organes hématopoïétiques ;
13. les cancers.

Chaque type de pathologie est ensuite décomposée sur au plus trois niveaux de hiérarchie (*cf.* l'exemple donnée par la figure A.12).

A.2 Processus de construction de l'ontologie

Deux types de ressources ont été utilisées pour la construction de cette ontologie : les nomenclatures et les textes réglementaires.

Les nomenclatures ont comme avantage de contenir déjà en leur sein un semblant de structuration. Cette structuration arborescente apparaît la plupart du temps sous

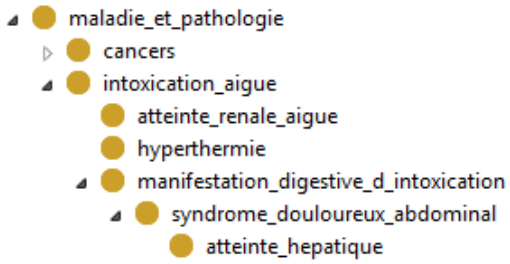


FIGURE A.12 – Extrait de l'ontologie HSE-Tennnaxia sur les intoxications aiguës.

le biais des systèmes de référence de chaque élément. Un concept ayant pour référence *A* a une série de sous-concepts notés *A.1, A.2*, etc. La structure y est alors facilement repérable.
De plus, tout y est concept ou terme candidat, et les représentations tabulaires permettent d'extraire commentaires et synonymes. Dans certains cas, il est même possible d'extraire des relations. Par exemple, dans le cas des substances dangereuses, la colonne *phrase de risque* nous permet non seulement de créer les sous-concepts de *Risque* mais également de les relier aux substances concernées.
Par contre, bon nombre de nomenclatures souffrent de légèreté sémantique. En effet, si les liens sont de type hiérarchique dans près de 75% des cas, il n'est pas rare de trouver des liens de composition ou de tout autre nature. Sans oublier tous les concepts *Autres...*, véritables concepts poubelles des nomenclatures, qu'il est très difficile d'insérer correctement dans une ontologie digne de ce nom. L'insertion d'une nomenclature dans une ontologie demande donc tout un travail d'analyse en amont, voire de refonte dans le pire des cas.
Nous sommes partis de la première version de l'ontologie, dont nous avons modifié la structure en nous inspirant du thésaurus EUROVOC. Nous avons ensuite inséré, après analyse lexicale et sémantique, la nomenclature des installations classées, puis celle des activités et des produits. Nous avons ajouté les substances dangereuses en provenance de l'annexe I de la directive 67/548/CEE. En volume, l'ensemble des nomenclatures a fourni près des trois quarts de l'ontologie HSE-Tennaxia livrée à la fin de ce projet.
Le corpus de texte demande un pré-traitement avant de pouvoir être exploité pour son insertion dans l'ontologie. L'ensemble des textes réglementaires a été soumis au début de ce projet, en 2007 soit près d'un millier de textes, à l'analyseur syntaxique SYNTEX [Bourigault 2000a, Bourigault 2005], lequel en a extrait une liste d'environ 600.000 termes candidats. Le temps imparti, les moyens techniques comme humains, ne permettant pas de traiter l'intégralité de cette liste, il a été décidé de ne pas les intégrer directement dans l'ontologie dans leur totalité et d'en faire une sélection.
Nous nous sommes inspirés pour ce faire de la méthode Terminae [Aussenac-Gilles 2003], laquelle distingue trois dimensions d'analyse :

1. suivant un axe *ascendant*, pour regrouper des concepts isolés ou des parties de

hiérarchies indépendantes ;

2. suivant un axe *descendant*, afin de détailler la hiérarchie en cherchant les enfants de chaque concept ;
3. suivant un axe *centrifuge*, dans le but de créer toutes les relations afférentes à un concept donné.

Cette sélection des termes s'est opérée en plusieurs temps.

La première phase est une phase dite de *structuration*. Il s'agit de repérer les *concepts centraux* puis d'établir une hiérarchie locale autour d'eux avec la relation *is_ a*. Cette opération se réalise à partir des syntagmes nominaux et de patrons lexicaux (un aspect plus délicat à gérer dans une phase automatique est la gestion des anaphores [12]). Par exemple, pour le concept *processus*, nous pouvons avoir *processus de xxx* mais également tous les *processus pour xxx* ainsi que les *xxx processus*, etc. Les verbes peuvent être utilisés pour bâtir l'ensemble des relations possibles. Par exemple, la relation *est requis* pour être liée à ses synonymes comme *est nécessaire, est indispensable, est prescrit, appelé à, demandé de, sollicité pour, etc.*. Afin d'alléger la tâche, nous avons décidé de ne pas prendre en compte les relations.

La seconde phase est une phase dite de *normalisation*. L'unicité des définitions est établie ainsi que l'homogénéité des points de vue et la cohérence des descriptions. Les concepts sont vérifiés (validation de l'étiquette, confirmation de la place dans la hiérarchie...). Une validation est opérée par une consultante de Tennaxia dédiée à cette tâche.

La conception d'une ontologie est un processus itératif. Une ontologie est un être vivant qui se nourrit d'éléments en provenance de l'univers dans lequel elle évolue. Deux facteurs principaux interviennent dans notre cas. L'un en provenance du domaine : l'évolution de la réglementation HSE. L'autre en provenance de la finalité de l'ontologie : les concepts recherchés par les utilisateurs du moteur de recherche de Tennaxia.
Pour le premier cas, durant la durée de ce projet, le domaine HSE a connu plusieurs évolutions massives en termes de réglementation avec d'une part le Grenelle de l'environnement, mais aussi la nouvelles nomenclature des substances dangereuses (SGH). Etant en phase de prototypage, il a été choisi de ne pas prendre en compte ces apports massifs de nouvelles connaissances.
Pour le second cas, l'ontologie construite ne couvre pas l'intégralité des concepts présents dans l'ensemble des textes réglementaires de Tennaxia. Il s'avère, par conséquent, intéressant d'analyser les requêtes faites dans l'outil en place au sein du

12. une anaphore est un mot ou un syntagme qui, dans un énoncé, assure une reprise sémantique d'un précédent segment appelé antécédent. Lorsque dans une telle relation, l'ordre d'apparition de ces deux éléments est inversé, le représentant prend alors le nom de cataphore, et le représenté, celui de conséquent.

module *Veille et Conformité* pour connaître les sous-domaines vers lesquels les utilisateurs orientent leur recherche. Ce relevé, croisé avec les termes candidats, permet de compléter l'ontologie au fur et à mesure de manière pertinente dans une approche client. Afin de couvrir certains sous-domaines plus spécifiques (foudre, équipement sous-pression...), des corpus spécialisés de quelques textes ont également été traités manuellement afin d'en extraire les concepts pertinents et les insérer dans l'ontologie.

Ce processus comporte encore une grande part d'humain, dans le sens où elle nécessite une double intervention de la part des experts du domaine : en premier lieu dans un but de validation, en un second pour compléter par une vue métier les connaissances modélisées. A ce titre, nous pouvons davantage classer notre travail dans le cadre des ontologies de métier que dans celui des ontologies du domaine juridique.

Le but de la construction de cette ontologie du domaine HSE est de fournir au prototype de moteur de recherche sémantique, et indirectement aux utilisateurs, une couverture de leur domaine suffisamment large pour pouvoir étendre n'importe quelle requête. De ce fait, l'un des critères de réussite de cette ontologie HSE-Tennaxia réside véritablement dans la proportion de termes saisis par les utilisateurs et possédant au moins une correspondance dans l'ontologie.

Bibliographie

[Abadia 2007] G. Abadia, C. Gayet, B. Delemotte, A. Delépine, A. Leprince, F. Michiels et D. Payan. Les maladies professionnelles - guide d'accès au tableau du régime général et du régime agricole de la sécurité sociale, septembre 2007. 201

[Aimé 2008a] X. Aimé, F. Fürst, P. Kuntz et F. Trichet. *Gradients de prototypicalité conceptuelle et lexicale.* In Revue des Nouvelles Technologies de l'Information - Extraction et Gestion des Connaissances (EGC'2008), volume 1 (11), pages 127–132. Cépaduès, 2008. ISBN 978.2.85428.819.3. 104

[Aimé 2008b] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *Conceptual and Lexical Prototypicality Gradients Dedicated to Ontology Personalisation.* In Proceedings of the 7th International Conference on Ontologies Databases and Applications of Semantics (ODBASE'2008 - Monterrey, Mexique), volume 5332, pages 1423–1439. Lecture Notes in Computer Science (LNCS). Springer Verlag / Heidelberg, 2008. ISBN 978-3-540-88872-7. 188

[Aimé 2008c] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *REDENE - Recherche documentaire assistée par ontologies de domaine adaptatives.* In Actes de la 5ième Conférence en Recherche d'Information et Applications (CORIA'2008), Trégastel, pages 467–474, 2008. 189

[Aimé 2008d] X. Aimé et F. Trichet. *Semantic Information Retrieval dedicated to Multimedia Systems : a platform based on Conceptual Graphs.* In Proceedings of the International Symposium on Intelligent Interactive Multimedia Systems and Services (Athens, Greece), Studies in Computational Intelligence, volume 42, pages 201–210. Springer Verlag / Heidelberg, 2008. ISBN 978-3-540-68126-7. 189

[Aimé 2009a] X. Aimé, F. Fürst, P. Kuntz et F. Trichet. *Gradients de prototypicalité appliqués à la personnalisation d'ontologies.* In F. L. Gandon, editeur, IC 2009 : Actes des 20es Journées Francophones d'Ingénierie des Connaissances IC 2009, pages 241–252. PUG, 2009. ISBN 978-2-7061-1538-7. Papier primé. 119

[Aimé 2009b] X. Aimé, F. Fürst, P. Kuntz et F. Trichet. *SEMIOSEM : une mesure de similarité conceptuelle fondée sur une approche sémiotique.* In F. L. Gandon, editeur, IC 2009 : Actes des 20es Journées Francophones d'Ingénierie des Connaissances IC 2009, pages 229–240. PUG, 2009. ISBN 978-2-7061-1538-7. 117

[Aimé 2009c] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *Ontology Personalization : an approach based on Conceptual Prototypicality.* In Proceedings of the International Workshop on Aspects in Evaluating Holistic Quality of Ontology-based Information Retrieval. Suzhou, China., 2009. 188

[Aimé 2009d] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *SemioSem : A Semiotic-Based Similarity Measure.* In Robert Meersman, Pilar Herrero et Tharam Dillon, editeurs, On the Move to Meaningful Internet Systems : OTM 2009 Workshops, volume 5872 of *Lecture Notes in Computer Science*. Springer, 2009. ISBN 978-3-642-05289-7. 117

[Aimé 2009e] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *Semiotic-based Prototypicality Gradient.* In Proceedings of the 11th International Conference on Informatics and Semiotics in Organisations, Beijing, China., pages 239–246. Ausino Academic Publishing House, 2009. ISBN : 978-0-9806057-2-3. 188

[Aimé 2010a] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées.* In Actes de l'atelier Personnalisation du Web, 10ième Journées francophones d'Extraction et de Gestion de Connaissances (EGC'2010), Hammamet, 2010. 189

[Aimé 2010b] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *Improving the efficiency of ontology engineering by introducing prototypicality.* In T. Agotnes, editeur, Proceedings of the 5th European Starting AI Researcher Symposium . Lisbon, Portugal. IOPress, 2010. ISBN 1607506750. 188

[Aimé 2010c] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *Improving the efficiency of ontology engineering by introducing prototypicality.* In R. Studer H. Coelho et M. Wolldridge, editeurs, Proceedings of the 19th European Conference on Artificial Intelligence. Lisbon, Portugal, pages 1081–1082. IOPress, 2010. ISBN 978-1-60750-605-8. 188

[Aimé 2010d] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *Prototypicality Gradient and Similarity Measure : a Semiotic-based Approach dedicated to Ontology Personalization.* Journal of Intelligent Information Management. Scientific Research, vol. 2, no. 2, pages 65–79, Feb. 2010. ISSN : 2150-8194. 119

[Amrani 2005] A. Amrani et O. Matte-Tailliez. *Extraction d'information sur la maladie de Crohn par visualisation interactive.* In Journées Ouvertes Biologie Informatique Mathématiques, 2005. 40

[Arucri 1986] L. Arucri et V. Girotto. *Norme di tipica per sei categorie naturali, Uno studio evolutivo.* Giornale Italiano di Psicologia, no. 3, pages 409–443, 1986. 78

[Au Yeung 2006] C. M. Au Yeung et H. F. Leung. *Ontology with Likeliness and Typicality of Objects in Concepts.* In Springer Berlin / Heidelberg, editeur, Proceedings of the 25th International Conference on Conceptual Modeling - ER 2006, volume 4215/2006, 2006. ISSN 0302-9743 (Print). 104

[Aussenac-Gilles 2000] N. Aussenac-Gilles, B. Biébow et S. Szulman. *Revisiting ontology design : a method based on corpus analysis.* In 12th European Knowledge Acquisition Workshop (EKAW'00), pages 172–188. R. Dieng, O. Corby (Eds.), 2000. 55

[Aussenac-Gilles 2003] N. Aussenac-Gilles, B. Biebow et S. Szulman. *D'une méthode à un guide pratique de modélisation de connaissances à partir de textes.* Actes

des 5e rencontres Terminologie et IA (TIA 2003), vol. 2003, pages 41–53, 2003. 203

[Aussenac-Gilles 2005] N. Aussenac-Gilles et D. Sörgel. *Text analysis for ontology and terminology.* Applied Ontology, vol. 1, pages 35–46, 2005. 55

[Bachimont 2000] B. Bachimont. Ingénierie des connaissances, évolutions récentes et nouveaux défis, chapitre Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances, pages 305–323. Eyrolles, 2000. 58, 59, 60

[Barsalou 1983] L.W. Barsalou. *Ad Hoc Categories.* Memory and Cognition, vol. 11, no. 3, pages 211–227, 1983. 3, 72, 76

[Barsalou 1985] L.W. Barsalou. *Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories.* Journal of Experimental Psychology : Learning, Memory, and Cognition., vol. 11, pages 629–654, 1985. 68

[Barsalou 1991] L.W. Barsalou. Deriving categories to achieve goals, volume 27, pages 1–64. CA : Academic Press, 1991. G.H. Bower (Ed.). 67

[Berger 2000] C. Berger et F. Bonthoux. *Influence de la nature de la tâche et des connaissances sur la catégorisation du jeune enfant : accès aux catégories par les propriétés.* Psychologie Française, vol. 45, pages 123–130, 2000. 70

[Berners-Lee 2001] T. Berners-Lee, J. Handler et O. Lassila. *The Semantic Web.* Scientific American, vol. 284, no. 5, pages 34–43, 2001. 3, 11, 12

[Berscheid 1998] E. Berscheid et H.T. Reis. Handbook of social psychology, chapitre Attraction and close relationships, pages 193–281. Oxford University Press, USA ; 4th edition, 1998. ISBN-10 : 0195213769. 108

[Blanc 2006] N. Blanc, A. Sysseau et D. Brouillet. Emotion et cognition. quand l'émotion parle à la cognition. Concept-Psy, septembre 2006. ISBN 2-84835-104-7. 78

[Blanchard 2008] E. Blanchard. *Exploitation d'une hiérarchie de subsomption par le biais de mesures sémantiques.* PhD thesis, Ecole Polytechnique de l'Université de Nantes, 2008. 4

[Boster 1988] J.S. Boster. *Natural sources of internal category structure : typicality, familiraity, and similarity of birds.* Memory and cognition, vol. 16(3), pages 258–270, 1988. 78

[Bourigault 2000a] D. Bourigault et C. Fabre. *Approche linguistique pour l'analyse syntaxique de corpus.* Cahiers de grammaire, vol. 25, pages 131–151, 2000. 56, 203

[Bourigault 2000b] D. Bourigault et C. Jaquemin. Construction de ressources terminologiques, pages 215–233. Jean-Marie Pierrel, 2000. 54, 55, 57

[Bourigault 2003] D. Bourigault et N. Aussenac-Gilles. *Construction d'ontologies à partir de textes.* In TALN 2003, 2003. 54, 55

[Bourigault 2005] D. Bourigault, C. Fabre, C. Frérot, MP. Jacques et S. Ozdowska. *Syntex, analyseur syntaxique de corpus.* In TALN 2005, Dourdan, 6-10 juin, 2005. 203

[Bourrigault 1994] D. Bourrigault. *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes.* PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris, 1994. 56

[Brini 2005] A. Brini, M. Boughanen et D. Dubois. *A Model for Information Retrieval based on Possibilistic Networks.* In 12th Symposium on String Processing and Information Retrieval (SPIRE), 2005. 57

[Brisson 2004] L. Brisson. *Mesures d'intérêt subjectif et représentation des connaissances.* Rapport technique, Laboratoire I3S, Université Sophia Antipolis, 2004. 40, 52, 61

[Brucks 1985] M. Brucks. *The Effects of Product Class Knowledge on Information Search Behaviour.* Journal of Consumer Research, vol. 12, no. 1, pages 1–16, 1985. 69

[Bruner 1957] J.S. Bruner. *On perceptual readiness.* Psychological Review, vol. 64, pages 123–152, 1957. 69

[Bruner 1990] J. S. Bruner. Acts of meaning. 1990. 70

[Brunner 2003] J.S. Brunner. Clusterisation par exploitation de structurations sémantiques de domaine ; une contribution au moteur wishbone. rapport de stage. Master's thesis, IRIN, 2003. 37

[Buitelaar 2005] Buitelaar. *Ontology Learning from Text : Methods, Evaluation and Applications.* Frontiers in Artificial Intelligence and Applications, vol. 123, 2005. 55

[CAF 2002] *Le Français expliqué. Cours Autodidactique de Français Écrit.* Faculté de Lettres de Montréal web site, 2002. 47

[CAF 2005] *Le Français expliqué. Cours Autodidactique de Français Écrit.* Faculté de Lettres de Montréal web site, 2005. 42

[Cardoso 2007] J. Cardoso. *The Semantic Web Vision : Where are We ?* IEEE Intelligent Systems, pages 22–26, 2007. 61

[Cauzinille-Marmèche 1998] E. Cauzinille-Marmèche, D. Dubois et J. Mathieu. Modèles généraux et locaux du développement cognitif, chapitre Catégories et processus de catégorisation. PUF, 1998. 70

[Chandrasekaran 1998] B. Chandrasekaran, J.R. Josephson et V.R. Benjamins. *The Ontology of Tasks and Methods.* In Proceedings of the 11th workshop on Knowledge Acquisition, Modeling and Management (KAW'98), 1998. 97

[Charlet 2000] J. Charlet, M. Zacklad, G. Kassel et D. Bourigault. Ingénierie des connaissances, évolutions récentes et nouveaux défis. Eyrolles, 2000. ISBN 2-212-09110-9. 2

[Charlet 2005] J. Charlet, B. Bachimont et R. Troncy. *Ontologies pour le Web sémantique*. Mission de recherche STIM, AP-HP er INSERM ERM 0202, 2005. 52

[Chausson 2007] C. Chausson. *Le Web sémantique, un vaste terrain d'applications encore à défricher*. Le Monde Informatique, vol. février, 2007. 31

[Chemlal 2006] S. Chemlal et F. Cordier. *Structures conceptuelles, représentation des objets et des relations entre les objets*. Canadian Journal of Experimental Psychology, vol. 60, pages 7–23, 2006. 67, 68

[Cohen 1987] J. Cohen et K. Basu. *Alternative Models of Categorization : Toward a Contingent Processing Framework*. Journal of Consumer Research, vol. 13, pages 455–472, 1987. 68

[Collins 1969] A.M. Collins et M.R. Quillian. *Retrieval time from semantic memory*. Journal of Verbal Learning and Verbal Behavior, vol. 8, pages 240–247, 1969. 71, 81

[Condamines 2005] A. Condamines. Sémantique et corpus. 2005. 54

[Cordier 1985] F. Cordier. *Formal and locative categories. Are there typical instance ?* Psychologica Belgica, vol. XXV, no. 2, pages 115–125, 1985. 78

[Cordier 1993] F. Cordier. Les représentations cognitives privilégiées, typicalité et niveau de base. Presses Universitaires de Lille, 1993. 76, 77, 78

[Dailland 2005] F. Dailland. *Tout savoir sur le MESH ou presque...* Rapport technique, Université de Médecine, Paris XI, Le Kremlin-Bicêtre, 2005. 36

[Daille 2004] B. Daille. *Recent Trends in Computational Terminology*. Special issue of Terminology, vol. 10, 2004. 55

[d'Amato 2008] C. d'Amato, S. Staab et N. Fanizzi. *On the Influence of Description Logics Ontologies on Conceptual Similarity*. In EKAW 2008, International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns, pages 48–63, October 2008. 86

[de Poche 1983] Livre de Poche, editeur. Larousse de poche. 1983. 35

[Després 2007] S. Després, F. Fürst et S. Szulman. *Construction d'une ontologie du domaine HSE*, 2007. 164, 167, 193

[Dice 1945] L.R. Dice. *Measures of the amount of ecological association between species*. Ecology, vol. 26, pages 297–302, 1945. 86, 120

[Dubois 1991] D. Dubois. Sémantique et cognition. Paris : Ed. du CNRS, 1991. 68

[Dubois 2001] J. Dubois, M. Giacomo-Marcellesi et L. Gespin. Dictionnaire de linguistique et des sciences du langage. Larousse, 2001. 67

[Dubost 1999] K.l Dubost. *Spécification du modèle et la syntaxe du cadre de description des ressources. RDF. Traduction de la recommandation du W3C.*, 1999. 20, 23

[Dubost 2004] K. Dubost. *Ontologie, Thésaurus, Taxonomie et Web sémantique.*, 2004. 37

[Dubuc 2005] B. Dubuc. *Le cerveau à tous les niveaux.* Hôpital Douglas, Verdun / Montréal., 2005. 35

[Eco 1984] U. Eco. Sémiotique et philosophie du langage. PUF, 1984. 45

[Eco 1993] U. Eco. Le signe : histoire et analyse d'un concept. Livre de Poche, 1993. 35, 47

[Eco 1994] U. Eco. La recherche de la langue parfaite. Seuil, 1994. 33, 35, 39

[ENS 2005] *Une introduction au Web sémantique.* ENSI Caen web site, 2005. 19

[Eppstein 2005] R. Eppstein. *Ontologies.* INIST web site, 2005. 38

[Euzenat 2004] J. Euzenat et J.F. Baget. *OWL : un langage d'ontologies pour le web.* INRIA Rhônes-Alpes web site, 2004. 24

[Fensel 1998] D. Fensel. *Knowledge Engineering : Principles and Methods.* Data and Knowledge Engineering, vol. 25, pages 161–197, 1998. 38

[Finkelstein 2002] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman et E. Ruppin. *Placing Search in Context : The Concept Revisited.* ACM Transactions on Information Systems, vol. 20, no. 1, pages 116–131, 2002. 146, 186

[Fox 1998] M. S. Fox, M. Barbuceanu, M. Gruninger et J. Lin. Simulating organizations. computational models of institutions and groups., chapitre An Organisation Ontology for Enterprise Modeling, pages 131–152. 1998. ISBN : 0-262-66108-X. 61

[Fürst 2002] F. Fürst. *L'ingénierie ontologique.* Rapport technique, IRIN, Nantes, 2002. 39, 42, 57

[Geeraerts 1986] D. Geeraerts. *Functional Explanations in Diachronic Semantics.* Belgian Journal in Linguistics, vol. I, pages 67–98, 1986. 72

[Geerearts 1985] D. Geerearts. *Polysemization and Humboldt's Principle.* Cahier de l'Institut de linguistique de Louvain, vol. II(3-4), pages 67–93, 1985. 72

[Gómez-Pérez 1999] A. Gómez-Pérez. *Développements récents en matière de conception, de maintenance et d'utilisation des ontologies.* In Terminologie et intelligence artificielle, actes du colloque de Nantes, 1999. 41

[Gomez-Perez 2003] A. Gomez-Perez, M. Fernandez-Lopez et O. Corcho. Ontological engineering. Springer, Advanced Information and Knowledge Processing, 2003. 97

[Gruber 1993a] T.R. Gruber. *Toward principles for the design of ontologies used for knowledge sharing.* In N. Guarino et R. Poli, editeurs, Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, The Netherlands, 1993. Kluwer Academic Publishers. 2, 7, 10, 38, 60, 98, 185

[Gruber 1993b] T.R. Gruber. *A translation approach to portable ontology specifications.* Knowledge Acquisition, vol. 5 (2), pages 199–220, 1993. 50, 97

[Guarino 1995] N. Guarino et P. Giaretta. Towards very large knowledge bases : knowledge building and knowledge sharing., chapitre Ontologies and knowledge bases, towards a terminological clarification. IOS Press, 1995. 38

[Guarino 1998] N. Guarino. *Some ontological principles for designing upper level lexical resources.* In 1st International Conference on Language Resources and Evaluation, 1998. 60

[Guarino 2001] N. Guarino et C. Welty. *Supporting ontological analysis of taxonomic relationships.* Data and knowledge engineering, vol. 39, pages 51–74, 2001. 43, 48, 49

[Guelfi 2007] N. Guelfi, C. Pruski et C. Reynaud. *Les ontologies pour la recherche ciblée d'information sur le Web : une utilisation et extension d'OWL pour l'expansion de requêtes.* In 18èmes journées francophones d'Ingénierie des Connaissances, IC'2007, Plate Forme de l'AFIA, Grenoble, 2007. 189

[Harris 1968] Z. Harris. *Mathematical Structures of Language.* R.E. Krieger Publishing Company, Inc., 1968. 55

[Hernandez 2006] N. Hernandez. *Ontologies de domaine pour la modélisation du contexte en recherche d'information.* PhD thesis, Institut de recherche en informatique de Toulouse, décembre 2006. 56, 57, 91

[Hirst 1998] G. Hirst et D. St-Onge. *Lexical chains as representation of context for the detection and correction malapropisms.* WordNet : An electronic lexical database and some of its applications, pages 305–332, 1998. 146, 152

[Horridge 2004] M. Horridge, H. Knublauch, A. Rector, R. Stevens et C. Wroe. *A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools.* Université de Manchester, 1st édition, 2004. 26

[Howard 1963] J.A. Howard. Marketing : Executive and buyer behavior. New York : Columbia University Press., 1963. 68

[Jaccard 1901] P. Jaccard. *Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines.* Bulletin de la Société Vaudoise de Sciences Naturelles, vol. 37, pages 241–272, 1901. (in french). 4, 85, 120

[Jiang 1997] J. Jiang et D. Conrath. *Semantic similarity based on corpus statistics and lexical taxinomy.* In International Conference en Research in Computationnal Linguistics, pages 19–33, 1997. 4, 88, 146, 152

[Kleiber 1987] G. Kleiber. Etudes linguistiques générale et de linguistique latine offertes en hommage à guy serbat, chapitre Quelques réflexions sur le vague dans les langes naturelles, pages 157–172. Société pour l'information grammaticale, 1987. 72, 74

[Kleiber 2004] G. Kleiber. La sémantique du prototype. Presses Universitaire de France - coll. Linguistique Nouvelle, mars 2004. ISBN 2-1304-2837-1, 2e édition. 71, 72, 75, 79, 101, 102

[Koffka 1935] K. Koffka. Principles of gestalt psychology. Routledge & Kegan Paul PLC, 1935. ISBN 978-0710031211. 4, 84, 117, 185

[Lacombe 2006] E. Lacombe. *Le Web Sémantique, traduction de l'article de Tim Berners-Lee, James Hendler, Ora Lassile.* URFIST, Toulouse., 2006. 34

[Lacot 2005] X. Lacot. *Introduction à OWL, un langage XML d'ontologies Web.* École Nationale Supérieure des Télécommunications., 2005. 19

[Ladwein 1995] R. Ladwein. *Le jugement de typicalité comme heuristique de choix : approche comparative.* In R.A. Perterson A. Jolibert et A. Strazzieri, editeurs, Prceeding of the International Research Seminar, numéro 22, pages 351–362, 1995. 68

[Lakoff 1986] G. Lakoff et M. Johnson. Les métaphores dans la vie quotidienne. Les éditions de Minuit, 1986. ISBN : 2.7073.1059.X. 74

[Lame 2002] G. Lame. *Construction d'ontologie à partir de textes. Une ontologie du droit dédiée à la recherche d'informations sur le Web.* PhD thesis, Ecole des Mines de Paris, 2002. 40, 58

[Larousse 1932] Larousse, editeur. Larousse du xxe siècle. 1932. 33, 34, 35

[Leacock 1998] C. Leacock et M. Chodorow. Wordnet : an electronic lexical database, chapitre Combining local context and Wordnet similarity for word sense identification, pages 265–283. Cambridge, MA, The MIT Press, 1998. 4, 83

[Lecocq 1987] P. Lecocq. *Normes de typicalité sur huit catégories naturelles chez des enfants de CE2-CM1 âgés de 7 à dix ans.* non publié - Université de Lille III, 1987. 78

[Leger 2004] L. Leger. *La discrimination visuelle et sémantique des mots dans les affordances lexicales.* PhD thesis, Psychologie des Processus Cognitifs, Université Paris VIII, 2004. 85

[Lemaire 1999] P. Lemaire. Psychologie cognitive. 1999. 75

[Lemire 2006] D. Lemire. *Initiation à RDF - RDF par l'exemple.* web site, 2006. 20, 21

[Léger 2005] L. Léger, C. Tijus et T. Baccino. *La Discrimination Visuelle et Sémantique : pour la Conception Ergonomique du Contenu de Sites Web.* Revue d'Interaction Homme-Machine, vol. 6, no. 1, 2005. 85

[Lin 1998] D. Lin. *An information-theoric definition of similarity.* In Proceedings of the 15th international conference on Machine Learning, pages 296–304, 1998. 4, 88, 146, 152

[López 1996] M. Fernàndez López, A. Gómez-Pérez et A. De Vivente. *A towards a method to conceptualize domain ontologies.* In Workshop on ontological engineering. ECAI'96., 1996. 54

[López 2000] M. Fernàndez López. *Overview of methodologies for building ontologies.*, 2000. 52

[Lyon 1969] A.J. Lyon. *Criteria and Meaning.* Studium Generale, vol. 22, pages 401–426, 1969. 72

[Maedche 2001] A. Maedche, S. Staab, N. Stojanovic, R. Studer et Y. Sure. *SEAL - A Framework for Developing Semantic Portals.* In Proceedings of the 18th British National Conference on Databases. Oxford, UK., 2001. 49

[Martin 2006] J. Martin. La contrepétrie. Que sais-je ?, 2006. 21

[McEvoy 1982] M.E. McEvoy et D.L. Nelson. *Category norms and instance norms for 106 categories of various sizes.* American Journal of Psychology, vol. 95, pages 462–472, 1982. 78

[Messai 2006] N. Messai, M.D. Devignes, A. Napoli et M. Smail-Tabbone. *Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques BioRegistry.* Ingénierie des Systèmes d'Information : Systèmes d'information spécialisés, vol. 11, no. 1, pages 39–60, janvier - février 2006. 189

[Mikulinger 1990a] M. Mikulinger, P. Kedem et D. Paz. *Anxiety and categorization-1, the structure and boundaries of mental categories.* Personnality and individual differences, vol. 11, no. 11, pages 805–814, 1990. 114

[Mikulinger 1990b] M. Mikulinger, P. Kedem et D. Paz. *Anxiety and categorization-2, hierarchical level and mental categories.* Personnality and individual differences, vol. 11, no. 8, pages 815–821, 1990. 114

[Miller 1991] G. Miller et W. Charles. *Contextual correlates of semantic similarity.* Language and Cognitive Processes, vol. 6, no. 1, pages 1–28, 1991. 146, 147

[Min 2007] Ministère de l'écologie et du développement durable, Direction de la prévention des pollutions et des risques, Service de l'environnement industriel. *Nomenclature des installations classées pour la protection de l'environnement*, Janvier 2007. 193

[Moigno 2002] S. Le Moigno, J. Charlet, D. Bourigault et M.C. Jaulent. *Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale.* In Ingénierie des Connaissances - 13e journées francophones, pages 229–238. Cépaduès, 2002. 56

[Morfaux 1995] L.-M. Morfaux. Vocabulaire de la philosophie et des sciences humaines. Collection L.-M. Morfaux, 1995. 43

[Morris 1946] Ch. W. Morris. Signs, language, and behavior. Prentice Hall, 1946. 47

[Nedungadi 1985] P. Nedungadi et J.W. Hutchinson. *The Prototypicality of Brands : Relationships with Brand Awareness, Preference and Usage.* Advances in Consumer Research, vol. 12, pages 498–503, 1985. Elizabeth C. Hirschman and Morris Holbrook, Provo, UT : Association for Consumer Research. 69, 77

[Niedenthal 2004] P.M. Niedenthal, C. Auxiette, A. Nugier, N. Dalle, P. Bonin et M. Fayol. *A prototype analysis of the french category "Emotion".* In Cognition and Emotion, volume 18, pages 289–312, 2004. 78

[Norta 2010] A. Norta, L. Carlson et R. Yangarber. Utility survey of ontology tools. Department of Computer Science, Department of Linguistics - University of Helsinki, Finland, 2010. 62, 63, 65

[Noy 2000] N. Noy et D. Mc Guinness. *Développement d'une ontologie : guide pour la création de votre première ontologie.* Université de Stanford, 2000. 39, 51

[Ogden 1989] C. K. Ogden et LA. Richards. The meaning of meaning : A study of the influence of language upon thought and of the science of symbolism. Harcourt, 1989. ISBN-13 : 978-0156584463. 45

[Osgood 1953] C. Egerton Osgood. Method and theory in experimental psychology. Oxford University Press, 1953. ISBN-13 : 978-0195010084. 73

[Patwardhan 2003] S. Patwardhan, S. Banerjee et T. Pedersen. *Using Measures of Semantic Relatedness for Word Sense Disambiguation.* In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pages 241–257, 2003. 146

[Peirce 1978] Ch. S. Peirce. Ecrits sur le signe. L'ordre philosophique. Seuil, 1978. ISBN : 978-2020050135. 44, 45

[Peirce 1998] Ch. S. Peirce. Collected papers. Thoemmes Continuum, 1998. ISBN : 1855065568. 46

[Piaget 1972] J. Piaget. Essai de logique opératoire. Paris : Dunod, 1972. 68

[Poitrenaud 1998] S. Poitrenaud. *La représentation des procédures chez l'opérateur : description et mise en oeuvre des savoir faire.* PhD thesis, Université de Paris VIII, 1998. 84

[Prat 2006] M. Prat. Processus cognitifs et émotions. 2006. 75

[Psyché 2003] V. Psyché, O. Mendes et J. Bourdeau. *Apport de l'ingénierie ontologique aux environnements de formation à distance.* STICEF, Technologies et formation à distance, vol. Hors-série, 2003. 41, 42

[Rada 1989] R. Rada, H. Mili, E. Bicknell et M.Blettner. *Development and application of a metric on semantic nets.* IEEE Transaction on Systems, Man and Cybernetics, vol. 19, no. 1, pages 17–30, 1989. 4, 82

[Rastier 2001] F. Rastier. *Sémiotique et sciences de la culture.* Linx, vol. 44-45, pages 149–168, 2001. 45

[Reed 1972] K.S. Reed. *Patter recognition and categorization.* Cognitive Psychology, vol. 3, pages 207–238, 1972. 75

[Resnik 1993] P. Resnik. *Selection and Information : A Class-Based Approach to Lexical Relationships.* PhD thesis, University of Pennsylvania, Institute for Research in Cognitive Science, 1993. Unpublished. 86

[Resnik 1995] P. Resnik. *Using information content to evaluate semantic similarity in a taxonomy.* In 14th International Joint Conference on Artificial Intelligence (IJCAI 95), volume 1, pages 448–453, Montréal, August 1995. 4, 82, 87

[Resnik 1999] P. Resnik. *Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language.* Journal of Artificial Intelligence Research (JAIR), vol. 11, pages 95–130, 1999. 146, 152

[Rips 1973] L.J. Rips, E.J. Shoben et E.E. Smith. *Semantic distance and the verification of semantic relations.* Journal of verbal learning and verbal behavior, vol. 12, pages 1–20, 1973. 73, 76, 85

[Robert 1984] Le Robert, editeur. Petit robert, dictionnaire de la langue française. 1984. 33, 36, 37, 42

[Robertson 1976] S. Robertson et K. Sparck Jones. *Relevance weighting of search terms.* Journal of the American Society for Information Sciences, vol. 27 (3), pages 129–146, 1976. 57

[Roche 1999] C. Roche, J.-C. Marty et S. Lacroix. *Ontologie et terminologie : le modèle OK.* In Terminologie et intelligence artificielle, actes du colloque de Nantes, 1999. 44

[Roche 2007] C. Roche. *Dire n'est pas concevoir.* In Ingénierie des Connaissances - 18es journées francophones, pages 157–168. Cépaduès, juillet 2007. ISBN 978-2-85428-773-8. 51, 54, 55

[Rosch 1973] E. Rosch. *Natural Categories.* Cognitive Psychology, no. 4, pages 328–350, 1973. 3, 75

[Rosch 1975a] E. Rosch. *Cognitive Representations of Semantic Categories.* Journal of Experimental Psychology, no. 104, pages 192–233, 1975. 76, 78, 99

[Rosch 1975b] E. Rosch. Cross-cultural perspectives on learning, chapitre Universals and Cultural Specifics in Human Categorization, pages 177–206. Cross-cultural Research & Methodology. John Wiley & Sons Inc, 1975. ISBN-13 : 978-0470104712. 99

[Rosch 1975c] E. Rosch et C. Mervis. *Family Ressemblances : Studies in the Internal Structure of Categories.* Cognitive Psychology, vol. 7, pages 573–605, 1975. 3, 72, 76, 79, 185

[Rosch 1978] E. Rosch. *Principles of categorization.* Cognition and categorization, pages 27–48, 1978. 76

[Rosch 1981] E. Rosch et C. Mervis. *Categorization of Natural Objects.* Annual Review of Psychology, vol. 32, pages 89–113, 1981. 68

[Rossi 2006] J.P. Rossi. Psychologie de la mémoire. De Boeck, juin 2006. ISBN 2-8041-5222-7. 67

[Roth 1983] E. M. Roth et E. J. Shoben. *The effect of context on the structure of categories.* Cognitive psychology, vol. 15, no. 3, pages 346–378, 1983. ISSN 0010-0285. 70

[Sanderson 1999] M. Sanderson et W.B. Croft. *Deriving concept hierarchies from text.* In Proceedings of the 22nd International ACM SIGIR Conference, pages 206–213, 1999. 87

[Saussure 1962] F. De Saussure. Cours de linguistique générale. Payot, 1962. 46

[Sebeok 1994] Th. A. Sebeok. An introduction to semiotics. Univ of Toronto Pr, 1994. ISBN : 978-0802077806. 46

[Simon 2007] Y. Simon. *Le Web sémantique, infrastructure du social média*, Décembre 2007. 14

[Staab 2000] S. Staab et A. Maedche. *Axioms are objects too : Ontology engineering beyond the modeling of concepts and relations.* Rapport technique 399, Institute AIFB, Karlsruhe, 2000. 49

[Sujan 1985] M. Sujan. *Consumer Knowledge : Effects on Evaluation Strategies Mediating Consumer Judgments.* Journal of Consumer Research, vol. 12, pages 31–46, 1985. 69

[Tajfel 1979] H. Tajfel et J.C. Turner. The social psychology of intergroup relations, chapitre An integrative theory of intergroup conflict, pages 33–48. Pacific Grove, CA/ Brooks/Cole., 1979. 97

[Tajfel 1986] H. Tajfel et J.C. Turner. Psychology of intergroup relations, chapitre The social identity theory of intergroup behavior, pages 7–24. Burnham Inc Pub ; 2 edition, 1986. ISBN-10 : 0830410759. 97

[Thibaut 1997] J.-P. Thibaut. *Similarité et catégorisation.* L'année psychologique, vol. 97, no. 97-4, pages 701–736, 1997. 83

[Trichet 2010] F. Trichet, X. Aimé et C. Thovex. The handbook of research in culturally-aware information technology : Perspectives and models., chapitre OSIRIS : Ontology-based System for Semantic Information Retrieval and Indexation dedicated to community and open web spaces. IGI Global, 2010. ISBN : 978-1615208838. 189

[Troadec 2007] B. Troadec. Psychologie culturelle. le développement cognitif est-il culturel ? Belin, 2007. ISBN 978-2-7011-3389-8. 99

[Tversky 1977] Amos Tversky. *Features of Similarity.* In Psychological Review, volume 84, pages 327–352, 1977. 4, 84

[Uschold 1996] M. Uschold et M. Gruninger. *Ontologies : principles, methods and application.* Knowledge engineering review, vol. 11 (2), pages 93–155, 1996. 40, 42, 58

[Velardi 2002] P. Velardi, P. Fabriani et M. Missikoff. *Using text processing techniques to automatically enrich a domain ontology.* In ACM Conference on Formal Ontologies and Information Systems, pages 270–284, 2002. 56

[W3C 2004] W3C. *OWL Web Ontology Language Guide*, 2004. 25, 26, 27, 28

[Wittgenstein 1973] L. Wittgenstein. Philosophical investigations. Prentice Hall ; 3 edition, 1973. ISBN-10 : 0024288101. 78

[WoodField 2004] A. WoodField. Introduction aux sciences cognitives, chapitre 9 : Un modèle à deux étapes de la formation des concepts, pages 277–293. Folio Essai, 2004. 76

[Wu 1994] Z. Wu et M. Palmer. *Verb semantics and lexical selection.* In Proceedings of the 32nd annual meeting of the Association for Computational Linguistics, pages 133–138, 1994. 4, 83, 146, 152

[Zammuner 1998] V.L. Zammuner. *Concepts of emotion : Eomitionness, and dimensional rating of italian emotion words.* In Cognition and Emotion, volume 12, pages 151–175, 1998. 78

Résumé : En psychologie cognitive, la notion de prototype apparaît de manière centrale dans les représentations conceptuelles. Dans le cadre de nos travaux, nous proposons d'introduire cette notion au sein des activités relevant de l'Ingénierie des Ontologies et de ses modèles de représentation. L'approche sémiotique que nous avons développée est fondée sur les trois dimensions d'une conceptualisation que sont l'intension (les propriétés), l'expression (les termes), et l'extension (les instances). Elle intègre, en sus de l'ontologie, des connaissances supplémentaires propres à l'utilisateur (pondération des propriétés, corpus, instances). Pratiquement, il s'agit de pondérer les liens "is-a", les termes et les instances d'une hiérarchie de concepts, au moyen de gradients de prototypicalité respectivement conceptuelle, lexicale et extensionnelle. Notre approche a été mise en œuvre dans un système industriel de gestion documentaire et de recherche d'information pour la société Tennaxia - société de veille juridique dans le domaine de l'Environnement. Elle a conduit au développement d'une ontologie du domaine Hygiène-Sécurité-Environnement, et de deux applications logicielles : l'application TooPrag dédiée au calcul des différents gradients de prototypicalité, et le moteur de Recherche d'Information sémantique Theseus qui exploite les gradients de prototypicalité. Nous avons enfin étendu notre approche à la définition de deux nouvelles mesures sémantiques, en nous inspirant des lois de similarité et de proximité de la théorie de la perception : Semiosem, une mesure de similarité, et Proxem, une mesure de proximité.

Mots clés : prototypicalité, ontologie, sémiotique, personnalisation, similarité, proximité

Abstract : In cognitive psychology, the concept of prototype appears in a central way in the conceptual representations. In our works, we propose to introduce this concept within the activities concerned by Ontology Engineering and its models of representation. Our semiotic approach is based on the three dimensions on conceptualization : intension (properties), expression (terms), and extension (instances). It adds, in addition to the ontology, some user's additional knowledge (weighting of the properties, corpus, instances). We balance "is-a" links, terms and instances in a hierarchy of concepts, by respectively conceptual, lexical and extensional prototypicality gradients. Our approach was implemented in an industrial document management and information research system for Tennaxia company - a company which offers service and software for legal intelligence in the environment domain. It led to the development of an ontology of the Environment domain, and two software : TooPrag an application dedicated to the calculation of the prototypicality gradients, and the semantic search engine Theseus which exploits the prototypicality gradients. We finally extended our approach to the definition of two new semantic measures inspired by the laws of similarity and proximity as theorized by gestalt psychologists : Semiosem, a measure of similarity, and Proxem, a measure of proximity.

Keywords : prototypicality, ontology, semiotic, personalization, similarity, proximity