UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CONTRIBUTION À LA MAINTENANCE DES ONTOLOGIES À PARTIR D'ANALYSES TEXTUELLES : EXTRACTION DE TERMES ET DE RELATIONS ENTRE TERMES

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

YASSINE GARGOURI

Avril, 2009

« Vous ne savez pas ce que vous ne savez pas jusqu'à ce que vous le sachiez...
La bonne solution est une recherche continue de la bonne solution ».

Dr. Ichak Adizes

AVANT-PROPOS

Cette thèse de doctorat s'insère dans le programme multidisciplinaire de doctorat en informatique cognitive de l'Université du Québec à Montréal. La problématique de la maintenance des ontologies de domaine est donc principalement abordée d'un point de vue informatique. Cependant, les différentes réflexions exposées font appel à des concepts et des techniques provenant de plusieurs domaines, parmi lesquels figurent non seulement les sciences cognitives et la représentation de connaissances, mais aussi l'Intelligence Artificielle (IA) et le Traitement Automatique du Langage Naturel (TALN). Il s'agit d'une recherche de nature pluridisciplinaire dont l'objectif général est de mettre en relation un certain nombre d'hypothèses théoriques, d'explorer plusieurs concepts et d'appliquer des techniques provenant de différentes disciplines afin de proposer une solution aux problèmes complexes de la maintenance des ontologies et, en particulier, la découverte de relations entre termes pertinentes à l'ontologie.

Ce projet n'aurait pu être réalisé sans la précieuse collaboration et le constant soutien de plusieurs professeurs, collègues, parents et amis. Je remercie donc tous ceux et celles qui, durant les dernières années, ont contribué à la réussite de ce projet.

Je tiens plus particulièrement à remercier chaleureusement mon directeur de recherche, Monsieur Bernard Lefebvre, professeur d'informatique cognitive et d'intelligence artificielle au Département d'informatique de l'Université du Québec à Montréal. Membre du Groupe de Recherche Interuniversitaire en Tutoriel Intelligent (GRITI), il a dirigé le projet de Gestion et Diffusion du Savoir en Télécommunication (GDST) qui a été la principale motivation de mon projet de recherche. Bernard Lefebvre m'a prodigué de précieux conseils qui ont contribué à la qualité de cette thèse de doctorat.

Je tiens aussi à remercier mon co-directeur, Monsieur Jean-Guy Meunier, professeur de philosophie du langage et de sciences cognitives au Département de philosophie de l'Université du Québec à Montréal et co-directeur (fondateur) du Laboratoire d'ANalyse Cognitive de l'Information (LANCI). Plusieurs des idées présentées dans cette thèse s'inspirent de celles développées, depuis des années, par Jean-Guy Meunier et ses

collaborateurs. Je le remercie vivement aussi pour m'avoir permis de bénéficier de l'environnement de recherche au sein du LANCI.

J'adresse aussi mes remerciements à tous les collègues avec lesquels j'ai eu le plaisir de collaborer au laboratoire LANCI et au programme de doctorat en informatique cognitive. Certaines idées que l'on retrouve dans cette thèse puisent effectivement leurs racines dans les échanges et les débats que nous avons eus durant les dernières années.

Ce travail rentre dans le cadre du projet GDST (Gestion et Diffusion du Savoir en Télécommunications) réalisé par des chercheurs de l'UQAM (Université du Québec à Montréal) et l'UDM (Université de Montréal) : Bernard Lefebvre, Jean-Guy Meunier, Gilles Gauthier, Omar Cherkaoui, Olivier Gerbé et d'autres étudiants de maîtrise et doctorat. La réalisation de cette recherche a été rendue possible grâce à la participation financière de Bell Canada, à travers son programme de support aux Recherches et Développements des Laboratoires universitaires Bell, ainsi que CRSNG (Conseil de Recherches en Sciences Naturelles et en Génie du Canada). Nous remercions enfin, Ismail Biskri (du LANCI) pour la plateforme SATIM (développée conjointement avec Jean-Guy Meunier) qu'il a mise à notre disposition.

Cette thèse n'aurait pu être menée à terme sans l'inestimable soutien de ma mère Olfa et de ma conjointe Amira. Je les remercie profondément pour leur patience, leur encouragement et leur soutien inconditionnel.

Une pensée toute particulière à l'âme de mon père Abdelmajid.

TABLE DES MATIÈRES

AVANT-PROPOS ..... **ii**
LISTE DES ACRONYMES ET ABRÉVIATIONS ..... **vii**
LISTE DES FIGURES ..... **viii**
LISTE DES TABLEAUX ..... **ix**
RÉSUMÉ ..... **x**
INTRODUCTION **1**
CHAPITRE I
PROBLÉMATIQUE ET OBJECTIFS DE RECHERCHE **4**
1.1 ONTOLOGIES DE DOMAINE 4
1.1.1 Définition de l'ontologie *4*
1.1.2 Systèmes à base d'ontologies de domaine *6*
1.1.3 Évolution des ontologies *7*
1.1.4 Conséquences des changements de l'ontologie *8*
1.2 CONTEXTE DU PROJET DE THÈSE 10
1.2.1 Le projet GDST *10*
1.2.2 L'ontologie des compétences *14*
1.2.3 Diffusion de documents *16*
1.2.4 Maintenance de l'ontologie du domaine *18*
1.3 PROBLÉMATIQUE DE RECHERCHE 19
1.4 MOTIVATIONS COGNITIVES 22
1.4.1 Représentation des connaissances en I.A *23*
1.4.2 Analyse Sémantique *26*
1.4.3 Psychologie Cognitive *26*
1.4.4 Architecture Cognitive *27*
1.5 OBJECTIFS DE LA RECHERCHE 28
1.5.1 Objectifs spécifiques *28*
1.5.2 Hypothèses de recherche *29*
CHAPITRE II
INGÉNIERIE DES ONTOLOGIES **34**
2.1 QU'EST-CE QU'UN TERME ? 34
2.1.1 L'aspect linguistique du terme *35*

2.1.2 L'aspect sémantique et conceptuel du terme *38*
2.2 CONCEPTION D'ONTOLOGIES DE DOMAINE 40
2.2.1 Processus de conception d'ontologie *40*
2.2.2 Outils de développement *48*
2.3 APPROCHES DE MAINTENANCE DES ONTOLOGIES 49
2.3.1 Apprentissage d'ontologie à partir de texte *50*
2.3.2 Apprentissage d'ontologie à partir de dictionnaire *52*
2.3.3 Apprentissage d'ontologie à partir de bases de connaissances *52*
2.3.4 Apprentissage d'ontologie à partir de schémas semi-structurés *52*
2.3.5 Apprentissage d'ontologie à partir de schémas relationnels *53*
2.3.6 Synthèse *53*
2.4 EXTRACTION DE TERMES À PARTIR DE TEXTES 54
2.4.1 Approches linguistiques pour l'extraction de termes complexes *57*
2.4.2 Approches statistiques pour l'extraction de termes complexes *60*
2.4.3 Méthodes mixtes *61*
2.4.4 Conclusion *62*
2.5 EXTRACTION DE RELATIONS À PARTIR DE TEXTES 64
2.5.1 Analyse par règles pour l'extraction de connaissances *65*
2.5.2 Extraction d'information en utilisant des patrons *65*
2.5.3 Approches statistiques *66*
CHAPITRE III
MÉTHODOLOGIE ET MODÈLE PROPOSÉ 72
3.1 MÉTHODOLOGIE DE RECHERCHE 72
3.2 MODÈLE PROPOSÉ 76
3.2.1 Extraction de n-grammes et filtrage de termes *77*
3.2.2 Classification et repérage de termes reliés *82*
3.2.3 Indexation Sémantique Latente *89*
3.2.4 Décomposition en Valeurs Singulières *91*
3.2.6 Vecteurs Conceptuels et thésaurus *94*
3.2.7 Mise à jour de l'ontologie *103*
3.2.8 Conclusion *107*
CHAPITRE IV
IMPLÉMENTATION ET EXPÉRIMENTATION 108
4.1 LOGICIELS UTILISÉS 108

4.1.2 Plateforme SATIM *108*
4.1.3 Chaîne de traitement ONTOLOGICO *109*
4.2 CONSTITUTION DE CORPUS 114
4.3 EXPÉRIMENTATION ET ÉVALUATION DES RÉSULTATS 114
4.4 VALIDATION DES RÉSULTATS 125
CHAPITRE V
CONCLUSION ET PERSPECTIVES 134
5.1 SYNTHÈSE ET CONTRIBUTIONS ORIGINALES 134
5.2 PROBLÈMES ET DÉFIS RENCONTRÉS 137
5.3 CONCLUSION ET PERSPECTIVES DE RECHERCHE 140
ANNEXES 144
ANNEXE 1 145
ONTOLOGIE DE DOMAINE 145
ANNEXE 2 146
ONTOLOGIE DE DOMAINE 146
ANNEXE 3 166
LISTE DE CLASSES 166
ANNEXE 4 169
UN EXEMPLE GÉNÉRÉ PAR ONTOLOGICO (CLASSE DE TERMES N. 7) 169
ANNEXE 5 176
L'ALGORITHME DVS 176
RÉFÉRENCES 188

LISTE DES ACRONYMES ET ABRÉVIATIONS

AFC	Analyse Formelle de Concepts
ART1	Adaptive Resonance Theory
DOMIF	Domaine d'Information
DVS	Décomposition en Valeurs Singulières
GDST	Gestion et Diffusion du Savoir en Télécommunication
IA	Intelligence Artificielle
ISL	Indexation Sémantique Latente
LANCI	Laboratoire d'Analyse Cognitive de l'Information
RC	Représentation de Connaissances
SATIM	Système d'Analyse et de Traitement de l'information multidimensionnelle
SBC	Systèmes à Base de Connaissances
TALN	Traitement Automatique du Langage Naturel
UNIF	Unité d'Information
VC	Vecteurs Conceptuels
RI	Recherche d'Information

LISTE DES FIGURES

Figure 1.1 : Schéma d'organisation du système d'ontologie 14

Figure 1.2 : Architecture du service de diffusion 18

Figure 2.1: Le triangle sémiotique 38

Figure 2.2 : Processus de construction d'ontologie 40

Figure 2.3 : Taxonomie des ressources nécessaires à la maintenance des ontologies et approches correspondantes 50

Figure 2.4 : Diagramme du contexte « ANIMAUX » 71

Figure 3.1 : Architecture de la chaîne de traitement ONTOLOGICO 73

Figure 3.2 : L'interaction entre les intrants et l'archive dans ART1 87

Figure 3.3 : La matrice terme-document W_c 91

Figure 3.4 : Décomposition en Valeurs Singulières de la matrice terme/document W_c 93

Figure 4.1 : Architecture de la chaîne de traitements ONTOLOGICO 111

Figure 4.2 : Interface graphique d'ONTOLOGICO 112

Figure 4.3 : Proportions des relations trouvées par rapport au nombre des relations possibles 121

Figure 4.4 : Nombre moyen de relations par classe de termes 122

Figure 4.5 : Taux de rappel et de précision par classe de termes 125

Figure 4.6 : Proportions des relations trouvées en fonction de l'approche utilisée 129

LISTE DES TABLEAUX

Tableau 2.1 : Un exemple de contexte formel de « ANIMALS » 69

Tableau 3.1 : Description de la famille des modèles du réseau de neurones ART (Adaptive Resonance Theory) 85

Tableau 3.2 : Composition des vecteurs conceptuels de (Wireless, Network) 99

Tableau 4.1 : Extrait de quelques classes de termes générées par Gramexco 116

Tableau 4.2 : Exemple de relations (fortes et très fortes) générées par ONTOLOGICO 120

Tableau 4.3 : Statistiques pour un extrait de 10 classes de termes sur les relations générées par ONTOLOGICO 121

Tableau 4.4 : Tableau Taux de rappel et de précision par classe de termes 124

RÉSUMÉ

Les ontologies sont des nouvelles formes de contrôle intelligent de l'information. Elles présentent un savoir préalable requis pour un traitement systématique de l'information à des fins de navigation, de rappel, de précision, etc. Toutefois, les ontologies sont confrontées de façon continue à un problème d'évolution. Étant donné la complexité des changements à apporter, un processus de maintenance, du moins semi-automatique, s'impose de plus en plus pour faciliter cette tâche et assurer sa fiabilité.

L'approche proposée trouve son fondement dans un modèle cognitif décrivant un processus d'extraction de connaissances à partir de textes et de thésaurus. Nous mettons ainsi, les textes au centre du processus d'ingénierie des connaissances et présentons une approche se démarquant des techniques formelles classiques en représentation de connaissances par son indépendance de la langue. Les traitements textuels sont fondés principalement sur un processus de classification supporté par un réseau de neurones (ART1) et sur l'Indexation Sémantique Latente appliquée sur des classes de termes.

Partant de l'hypothèse que l'extraction de connaissances à partir de textes ne peut se contenter d'un traitement statistique (ni même linguistique) de données textuelles pour accaparer toute leur richesse sémantique, un processus d'extraction de connaissances à partir d'un thésaurus a été conçu afin d'intégrer, le mieux possible, les connaissances du domaine au sein de l'ontologie. Ce processus est fondé principalement sur un calcul d'associations sémantiques entre des Vecteurs Conceptuels.

Le modèle proposé représente une chaîne de traitement (ONTOLOGICO) au sein de la plate-forme SATIM. Ce modèle vise à assister les experts de domaine dans leur tâche de conceptualisation et de maintenance des ontologies en se basant sur un processus itératif supporté par un ensemble de modules, en particulier, un extracteur de termes, un lemmatiseur, un segmenteur, un classifieur, un module de raffinement sémantique basé sur l'Indexation Sémantique Latente et un identificateur de termes reliés basé sur le calcul de similarité sémantique entre les couples de vecteurs conceptuels.

La découverte de relations entre termes pour les besoins d'une conceptualisation de domaine s'avère être le résultat d'une complémentarité de traitements appliqués tant sur des textes de domaine que sur un thésaurus. D'une part, les analyses textuelles fondées principalement sur l'application de l'Indexation Sémantique Latente sur des classes de termes génèrent des relations sémantiques précises. D'autre part, l'extraction de relations sémantiques à partir d'un thésaurus, en se basant sur une représentation par des Vecteurs conceptuels, constitue un choix théorique judicieux et performant. Ce processus joue en effet, un rôle important dans la complétude des relations.

Ce projet de recherche se place au cœur des échanges entre terminologie et acquisition de connaissances. Il amène une réflexion sur les divers paliers à envisager dans une telle démarche de modélisation de connaissances textuelles pour des objectifs de maintenance

d'une ontologie de domaine. La méthodologie proposée constitue une aide précieuse dans le domaine de la maintenance des ontologies. Elle assiste les terminologues chargés de naviguer à travers de vastes données textuelles pour extraire et normaliser la terminologie et facilite la tâche des ingénieurs en connaissances, chargés de modéliser des domaines.

Mots clés : maintenance d'ontologie, Traitement Automatique du Langage Naturel (TALN), Indexation Sémantique Latente, Vecteurs Conceptuels, classification automatique, Réseaux de Neurones.

INTRODUCTION

La maintenance des ontologies est un champ multidisciplinaire impliquant le traitement du langage naturel, la prospection de données, l'apprentissage machine et la représentation de connaissances. De ce fait, il est irréaliste de s'attendre à ce que l'humain comprenne la totalité de l'ontologie et de ses interdépendances internes. Il lui est difficile, voire même impossible, de repérer de nouvelles relations entre termes à partir de la simple lecture de données textuelles et d'évaluer leur pertinence par rapport à l'ontologie actuelle. Cette tâche cognitive est d'autant plus ardue que les nouveaux textes à analyser et l'ontologie existante sont de large taille. Ce problème se pose également lorsque la conceptualisation du domaine est ambiguë ou encore si l'utilisateur ne possède pas suffisamment d'expérience.

Au cours de la dernière décennie, de nombreuses recherches ont été réalisées dans le domaine de l'ingénierie des ontologies. La majorité de ces recherches se sont concentrées sur les problèmes de construction. Toutefois, il n'existe pas, jusqu'à date, de méthodes consensuelles et de lignes de conduite répondant à la problématique de la maintenance. L'absence de telles méthodes consensuelles entrave l'extension d'une ontologie donnée à partir d'autres et sa réutilisation.

Notre objectif de recherche consiste à mettre en place une « passerelle » entre les documents (lexiques, réseaux sémantiques...) et l'ontologie courante. Ceci revient à proposer un modèle permettant, à partir de l'analyse de nouveaux textes, d'identifier les nouveaux concepts spécifiques à un domaine ainsi que leurs relations et d'automatiser certaines tâches relatives à leur intégration au niveau de l'ontologie du domaine.

L'exploitation des sciences cognitives pour s'attaquer à certains problèmes dans le domaine de l'informatique constitue en effet, une piste intéressante pour les chercheurs en informatique. Pour une meilleure plausibilité de la modélisation proposée et dans la mesure où certains problèmes recensés sont de nature cognitive ou associée (extraction de relations sémantiques entre termes à partir d'analyses textuelles, classification de textes, représentation

de termes par des Vecteurs Conceptuels), une telle modélisation pourra tirer profit à notre avis, des travaux de recherche en sciences de la cognition, à savoir, la représentation des connaissances, les modèles connexionnistes de classification, l'Analyse Sémantique Latente, la cooccurrence, etc.

La thèse est articulée autour de quatre principaux chapitres. En guise d'introduction, nous présentons dans un premier chapitre, le contexte de notre projet de recherche ainsi que notre problématique de recherche et les objectifs généraux du projet. Ensuite, nous identifions les aspects cognitif et informatique de la problématique et formulons les principales hypothèses de notre modèle.

Le reste de la thèse comporte deux volets principaux : un volet exploratoire et un volet expérimental. Le volet exploratoire se penche sur les questions de modélisation d'un processus de maintenance d'ontologies de domaine, à la lumière des possibilités offertes par les sciences de la cognition. Il s'organise en deux chapitres. Le premier (chapitre 2) est consacré à une revue de littératures sur le processus de la conception d'ontologies et offre un panorama des approches de maintenance. Ce chapitre lève un pan du voile sur les difficultés auxquelles les ingénieurs d'ontologies sont confrontés lors de l'extraction de termes et de relations entre termes à partir d'analyses textuelles, lesquelles difficultés justifient ce travail de thèse.

Le chapitre fait ressortir les principales questions théoriques relevant des sciences de la cognition, et jugées pertinentes. Il en découle que l'usage des techniques statistiques de traitements automatiques de textes, l'Indexation sémantique Latente, associée avec la classification et l'Analyse Formelle de Concepts sont des pistes prometteuses pour surmonter une bonne partie de ces difficultés. Ces pistes sont examinées dans le troisième chapitre, sous le prisme des sciences de la cognition.

Le troisième chapitre est consacré à la présentation de notre méthodologie que nous envisageons de suivre et à la proposition d'une solution. Pour appuyer l'approche présentée et la rendre plus concrète, une architecture générale pour un système semi-automatique de maintenance d'ontologies est spécifiée et décrite. Cette architecture est en grande partie mise en œuvre dans le volet expérimental de la thèse.

Ce volet expérimental de la thèse (quatrième chapitre) montre entre autres, à travers le cas d'expérimentation qui y est présenté, comment le modèle proposé dans le volet exploratoire pourrait être exploité dans le cadre du développement d'outils d'assistance à la maintenance d'ontologies. L'évaluation du module ainsi que l'analyse des résultats de l'évaluation occupent une place non négligeable dans le chapitre. Il en ressort que le modèle proposé génère des résultats manifestement remarquables.

Nous concluons enfin, par l'énumération des différentes contributions originales de notre projet sur le plan scientifique, ainsi que les défis et les obstacles que nous avions à affronter. Ce travail se veut une ouverture sur d'autres projets de recherche. Nous suggérerons ainsi, quelques voies de recherche susceptibles de déboucher sur des environnements d'assistance à la maintenance des ontologies.

CHAPITRE I

PROBLÉMATIQUE ET OBJECTIFS DE RECHERCHE

Ce chapitre introduit les ontologies et la problématique de la maintenance. Le contexte de recherche de notre projet de thèse est également abordé en vue de présenter un exemple concret sur l'évolution des ontologies et le besoin de maintenance. Le chapitre avance une note d'espoir sur l'importante contribution des sciences de la cognition pour s'attaquer aux difficultés recensées et s'achève par l'identification des objectifs et les hypothèses de la recherche afin de guider notre démarche méthodologique.

1.1 Ontologies de domaine

1.1.1 Définition de l'ontologie

Le terme *ontologie*[1] est utilisé étymologiquement pour désigner l'étude philosophique de ce qui existe. Aujourd'hui, il n'est plus uniquement utilisé dans les débats philosophiques ; certains auteurs lui ont donné, au cours de l'histoire, des connotations se rapprochant de leurs domaines d'étude respectifs. Les chercheurs en IA se sont intéressés aux ontologies dans au moins deux domaines : la Représentation des Connaissances et l'Ingénierie des Connaissances. L'ontologie a été définie chez la communauté de l'Ingénierie des Connaissances comme étant une compréhension commune et partagée d'un domaine qui peut être communiquée entre des personnes et des systèmes (Guarino, 1995). Chez la communauté de représentation des connaissances, la définition d'ontologie la plus utilisée et fortement citée est celle de Gruber (Gruber, 1993) ; « *une ontologie est une spécification formelle et explicite d'une conceptualisation partagée* ». Cette spécification représente un modèle abstrait d'un phénomène du monde réel qui est défini par des concepts et des relations. Le principe général de l'ontologie est en fait analogue à celui des bases de données dans la

[1] La convention veut que la notation *Ontologie* (avec un *O majuscule*) soit attribuée au domaine issu de la philosophie et *ontologie* aux autres.

mesure où elle regroupe le vocabulaire d'un domaine en différentes classes (termes) et relie ces classes par le biais de relations.

Partant de ces définitions, les connaissances intégrées dans les ontologies peuvent être formalisées en mettant en jeu cinq types de composants : les classes, les relations, les fonctions, les axiomes et les instances (Gruber, 1993).

- **Les classes** sont habituellement organisées en taxonomies. Elles réfèrent à des concepts, utilisés dans le sens large. Ces concepts peuvent être abstraits ou concrets, élémentaires (électron) ou composés (atome), réels ou fictifs.
- **Les relations** (R) représentent un type d'interaction entre les notions d'un domaine (C_i). Elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles, c'est-à-dire $R \subset C_1 \times C_2 \times \ldots \times C_n$. Par exemple, les relations binaires sont du type «*sous-classe-de*», «*connecté-à*», etc.
- **Les fonctions** sont des cas particuliers de relations dans lesquelles le nième élément de la relation est défini à partir des n-1 premiers. Formellement, les fonctions (F) sont définies ainsi : $F : C_1 \times C_2 \times \ldots \times C_{n-1} \rightarrow C_n$.
- **Les axiomes** sont utilisés pour affirmer des phrases qui sont toujours vraies.
- **Les instances** sont utilisées pour représenter des éléments selon un principe similaire à la relation classe/objet en UML. Exemple : *Paris* est une instance de *Ville*.

Les *ontologies générales* (ou *ontologies de niveau supérieur*) spécifient les différentes catégories d'entités existant dans le monde. Ce type d'ontologies synthétise des notions très générales, indépendantes de tout domaine ou problème particulier. Par exemple, on spécifie dans une ontologie des régions, des concepts tels que *pays*, *ville*, *région,* etc. Contrairement aux ontologies générales, les *ontologies de domaine* sont plus spécifiques. Elles synthétisent les connaissances spécifiques à un domaine particulier tel que : *ontologie des biotechnologies, ontologie des télécommunications sans fils, ontologie du domaine de l'aviation civile*, etc. Les concepts du domaine considéré, ainsi que les relations entre concepts et les théories gouvernant le domaine y sont spécifiés.

Il convient de souligner que la conceptualisation d'une ontologie de domaine ne peut se faire de manière non ambigüe que dans un champ d'application précis. Cette restriction est

nécessaire pour garantir un certain consensus sur la sémantique associée aux termes du domaine. Par exemple, un même terme peut désigner deux concepts différents, c'est à dire deux objets, qui peuvent être physiquement les mêmes, avec des sémantiques différentes dans deux cadres applicatifs différents. Ainsi, la même sémantique ne sera pas associée à l'objet *Table* dans le cas où la *Table* est un meuble à disposer dans une pièce (contexte applicatif de l'aménagement intérieur) et dans le cas où la *Table* est une marchandise (contexte applicatif de l'import/export). Toutefois, délimiter rigoureusement un domaine de connaissance peut se révéler ardu, du fait de la nature holistique de la connaissance.

1.1.2 Systèmes à base d'ontologies de domaine

Si les Systèmes Experts n'avaient pour objet que la résolution automatique de problèmes, les Systèmes à Base de Connaissances (SBC) qui leur ont succédé ne sont pas censés faire manipuler en aveugle des connaissances à la machine, qui restitue à la fin la solution du problème. Les SBC permettent généralement un dialogue, une coopération entre le système et l'utilisateur humain. Ces systèmes permettent le stockage et la consultation de connaissances, le raisonnement automatique sur les connaissances stockées, leur modification et leur partage entre systèmes informatiques.

Les représentations symboliques utilisées dans les machines doivent avoir du sens aussi bien pour la machine que pour les utilisateurs. Plus précisément, ces représentations doivent non seulement utiliser les termes employés par l'être humain, mais également accaparer la sémantique que ce dernier associe aux différents termes, faute de quoi aucune communication efficace n'est possible.

Les ontologies sont de plus en plus considérées comme partie prenante dans plusieurs applications industrielles et académiques dans des domaines tels que la recherche basée sur la sémantique, l'interopérabilité entre applications, la spécification et la validation de contraintes, les applications de Web sémantique, etc. (Noy et al, 2001). Grâce à une modélisation du domaine sous forme d'ontologie, divers services peuvent être automatisés pour aider des utilisateurs à réaliser des objectifs particuliers en accédant à des informations stockées sous un format compréhensible par la machine.

Les systèmes de gestion de connaissances basées sur des ontologies utilisent ces dernières pour fournir et accéder à des sources de connaissances. Cette technique offre une terminologie appuyant en particulier les processus d'indexation et de recherche de connaissances. L'avantage principal, par rapport à une technique de recherche ou d'indexation basée sur les mots clés est attribué à la description formelle, commune et partagée du domaine, que l'ontologie permet de représenter. En effet, l'extraction de connaissances est améliorée grâce à un mécanisme d'extension de lexique en vue d'enrichir ce processus à l'aide de la terminologie étendue de l'ontologie. De plus, cette terminologie permet de fournir un ensemble d'hypothèses sur les significations réelles des termes utilisés. Par exemple, une recherche visant à repérer des sources de connaissances sur *souris* (matériel informatique) évite les ressources relatives à la « *souris* » animale, grâce à une extension implicite de la recherche du terme *souris* par *matériel informatique*, *ordinateur*, *périphérique* etc.

Par ailleurs, les ontologies sont de plus en plus importantes dans les usages industriels à l'instar des projets KAON (Bozsak et al, 2002) et KnowWork (Hans et al, 2001). Ces systèmes sont conçus pour gérer les connaissances relatives à la gestion des processus d'affaires, des documents, des produits, etc. Dans ce contexte, les ontologies sont continuellement mises à jour en raison de nouvelles situations de marché, de la restructuration des compagnies ou de l'accroissement des connaissances.

Nous discutons dans les sections suivantes des causes d'évolution des ontologies et des conséquences que cette évolution pourrait avoir sur les applications à base d'ontologies.

1.1.3 Évolution des ontologies

L'ontologie, en tant que *conceptualisation partagée* du domaine n'est pas une spécification statique. Elle est plutôt considérée comme un réseau dynamique de significations, dans lequel un consensus est atteint à l'intérieur d'un processus continu de changements d'informations et de significations (Fensel, 2001). Une conceptualisation peut également changer suivant la perspective d'usage. En effet, différentes tâches peuvent impliquer une variété de points de vues dans le domaine et par conséquent, une conceptualisation différente.

La conceptualisation d'un domaine et son mode d'expression évoluent, dans le temps, d'une façon continue selon un rythme qui dépend de la nature et des spécificités du domaine. De nouveaux termes apparaissent, d'autres changent de signification, un même concept prend des modalités d'expression (c'est-à-dire des termes) similaires, de nouvelles règles sont définies, des relations entre termes s'avèrent importantes, etc. Dans le monde réel, les instances des concepts évoluent d'une façon particulièrement fréquente. Par exemple, une fusion de deux compagnies en une seule constitue un changement dans le monde réel, relatif à des instances du concept « *compagnie* ». L'ontologie devrait par conséquent refléter cette modification.

Par ailleurs, des changements de spécifications peuvent être nécessaires lorsque l'ontologie est traduite d'un langage de représentation de connaissances à un autre. Les langages diffèrent, non seulement au niveau de leur syntaxe mais aussi et surtout au niveau sémantique. La préservation de la sémantique d'une ontologie durant la traduction n'est pas une tâche triviale. Des ajustements importants sont potentiellement nécessaires[2].

L'évolution des ontologies entraîne des problèmes d'efficacité opérationnelle affectant leur réutilisation. Il serait donc important d'explorer les conséquences des changements apportés à des ontologies et la stratégie à employer pour garantir leur consistance et leur cohérence et ainsi préserver l'efficience et l'aspect opérationnel des applications basées sur des ontologies.

1.1.4 Conséquences des changements de l'ontologie

L'ontologie doit préserver sa consistance suite aux changements complexes caractérisant son évolution (Stojanovic et al, 2002). Un changement élémentaire peut induire des inconsistances dans d'autres parties de l'ontologie. On distingue les inconsistances syntaxiques de celles sémantiques. Les premières surgissent quand des entités non définies dans l'ontologie ou au niveau des instances sont utilisées, ou encore lorsque les contraintes du modèle de l'ontologie ne sont pas valides. Des inconsistantes sémantiques apparaissent quand la signification d'une entité change suite à une modification dans l'ontologie. Par exemple, considérons une relation entre *livre* et *librairie* spécifiée dans une ontologie. La

[2] Les spécifications formelles des ontologies ne seront pas considérées dans le cadre de cette thèse. C'est plutôt l'aspect *conceptualisation* qui nous intéresse.

signification du concept « LIVRE » est définie à travers une propriété *vend* qui les relie. Si le concept « LIBRAIRIE » est supprimé, la sémantique de "LIVRE" n'est plus définie ; s'agit-il d'un ouvrage, du verbe *livrer*, d'une monnaie ou d'un poids ?

Les changements apportés à une ontologie font l'objet d'un processus complexe d'évolution, en raison de la dépendance de l'ontologie par rapport à différentes composantes :

- D'abord, l'évolution d'une ontologie affecte les données qui y sont reliées. Dans le cadre du Web sémantique, on parle de pages Web qui sont annotées en fonction de termes de l'ontologie. Quand cette ontologie change, ces données peuvent avoir une interprétation différente ou utiliser des termes inconnus.

- D'autres ontologies peuvent être reliées à l'ontologie modifiée. Elles sont construites à partir de l'ontologie source[3] ou encore, elles l'importent. Des changements apportés à la source peuvent affecter les ontologies reliées.

- Quand une ontologie est modifiée, les instances doivent être maintenues de façon à ce que l'ontologie et les instances demeurent réciproquement cohérentes. En d'autres termes, l'information (le texte) annotée doit s'adapter d'une façon continuelle à la nouvelle terminologie sémantique et ses relations.

- Enfin, les applications utilisant une ontologie peuvent aussi être remises en cause par des changements de cette dernière. Idéalement, les connaissances conceptuelles nécessaires à la conception d'une application sont principalement spécifiées dans l'ontologie. Or, dans la réalité, les applications utilisent également un vocabulaire qui lui est propre. Ce vocabulaire peut être incompatible avec l'ontologie modifiée. Par ailleurs, l'efficacité des systèmes de gestion de connaissances peut se voir affectée lorsque certaines des connaissances sont annotées avec une ancienne ontologie alors qu'une version plus récente est utilisée pour la recherche. Ainsi, pour une requête donnée, le système pourrait, non seulement rater des sources de connaissances pertinentes, mais aussi fournir des réponses erronées (Stojanovic et

[3] L'ontologie source : l'ontologie de base (racine) qui est importée (ou utilisée) dans une autre ontologie.

al, 2002). Par conséquent, il est primordial que les connaissances soient annotées d'une façon synchrone avec la version actuelle de l'ontologie.

1.2 Contexte du projet de thèse

Notre travail de thèse rentre dans le cadre du projet GDST (Gestion et Diffusion du Savoir en Télécommunications) réalisé par des chercheurs de l'Université de Québec à Montréal et l'Université de Montréal. Le choix du LANCI comme lieu de recherche se justifie spécialement par la dynamique de recherche du laboratoire par rapport aux analyses textuelles, la classification, la catégorisation, la recherche d'informations, le développement et la maintenance d'ontologies... Les chercheurs du LANCI ont développé, au fil des années, une plateforme modulaire SATIM (Biskri et al, 2002) constituée d'un ensemble de chaînes de traitements d'information. Notre logique « ONTOLOGICO » s'inscrit dans cette même vision modulaire en réutilisant certains modules de la plateforme et en en créant d'autres, en vue de proposer une nouvelle chaîne de traitements visant à découvrir des relations sémantiques.

Après un bref aperçu sur la structure du projet GDST nous décrivons dans cette section l'ontologie du domaine et celle des compétences et expliquons l'importance de ces deux ontologies pour construire un service de diffusion de documents vers les usagers concernés. Nous introduisons ensuite le processus de maintenance que nous proposons pour mettre à jour l'ontologie du domaine. Ce processus est en effet essentiel à la survie à long terme de tout système de diffusion.

1.2.1 Le projet GDST

Bien que les moteurs de recherche jouent un rôle important dans l'appariement de documents à des requêtes spécifiques, ils sont considérés comme des outils restreints et limités. En raison de l'ampleur et de la complexité de la documentation dans les domaines spécialisés, la pertinence de ces outils a été remise en cause. En effet, la recherche dans les documents s'effectue habituellement par le biais de mots clés, mais le seul critère des moteurs de recherche demeure la présence des mots dans le texte. En d'autres termes, l'ordre de présentation des résultats dépend de la proximité des mots recherchés par rapport au contenu

informationnel du texte trouvé. Moins l'écart entre les mots est grand, plus l'ordre de présentation sera élevé (présenté en premier). Les documents n'étant pas spécifiquement annotés pour identifier clairement leur contenu, les résultats de recherche retournent dans la plupart des cas une foule de documents n'ayant aucun rapport avec les besoins et les recherches sont systématiquement imprécises.

Dans un domaine en perpétuel changement comme les télécommunications, ce problème devient sérieux. Par conséquent, les annotations de documents se présentent comme une solution incontournable, améliorant ainsi les résultats de recherche et permettant à des applications de manipuler, extraire et réutiliser cette information. L'annotation des documents revient à utiliser un langage (tel que DAML+OIL, OWL, RDF, etc.) pour rattacher des métadonnées (des explications, des commentaires, etc.) à un document Web ou encore pour représenter son contenu sémantique en se basant sur l'ontologie du domaine.

Dans un environnement organisationnel, l'information ne peut devenir connaissance que si elle est cataloguée, structurée et disponible d'accès pour les bonnes personnes, et ce, au bon moment. L'implantation de solutions informatiques pour la gestion de la connaissance qui répondent à ces objectifs est un phénomène plutôt récent. Cela implique en effet l'intégration assez difficile de concepts et techniques issus de différents domaines comme l'intelligence artificielle, l'ingénierie des systèmes d'informations, la réingénierie des processus ou le comportement des organisations et de leurs ressources humaines (Liebowitz, 1999).

Le projet GDST (Gestion et Diffusion de Savoir en Télécommunication) (Lefebvre et al, 2003) entre dans ce cadre. Il a pour principal objectif de faciliter le développement des compétences des ressources humaines pour les besoins d'une entreprise. Il s'agit d'un environnement informatique qui vise à fournir une aide précise et concrète aux utilisateurs dans la réalisation de leurs activités, et par conséquent, à renforcer leur productivité. Cela doit aussi les aider à enrichir leurs connaissances professionnelles, et donc contribuer de façon permanente à leur formation.

Dans un environnement où les connaissances relatives aux activités professionnelles s'accroissent très rapidement et où elles ne sont pas correctement organisées, les utilisateurs (les experts, les techniciens, ...) ne sont pas toujours conscients de l'existence des

informations utiles à leurs activités. Quand ils en sont conscients, ils ne savent pas nécessairement comment y accéder. L'intérêt d'avoir un service de diffusion actif qui vise principalement à acheminer les informations utiles et pertinentes aux personnes concernées est de ce fait évident.

Un des grands défis de la diffusion active et intelligente de documents vers des utilisateurs est de ne pas les importuner par des informations qui soient éloignées de leurs compétences ou encore de leurs champs d'intérêt. Au delà de la gestion de l'information, on parle de gestion des connaissances, de la sémantique que l'individu attache à une information et du contexte dans lequel elle est utilisée. Le projet GDST représente une réponse à ce défi, basée fondamentalement sur un service de diffusion guidé principalement par des ontologies de domaine et de compétences, mais aussi de documents, de processus, d'usagers et de l'entreprise.

Parmi les travaux qui ont des objectifs comparables à ceux du projet GDST, on peut distinguer celui de (Abecker et al, 1998) dont la plateforme « *KnowMore* » a été développée sur la base d'une architecture de mémoire organisationnelle pour répondre aux besoins d'interrogation dans le contexte d'une tâche. Si ce projet se base principalement sur les connaissances du domaine pour décrire les documents, le projet GDST intègre une dimension additionnelle, celle des compétences, dans l'objectif de mieux caractériser les utilisateurs.

À l'instar du projet « *Ontologging* » (Razmerita et al, 2003), le projet GDST est également fondé sur des agents contribuant à une meilleure personnalisation de la diffusion de connaissances.

En intelligence artificielle, les ontologies sont apparues comme une réponse aux problématiques de représentation et de manipulation des connaissances au sein des systèmes informatiques. L'importance que revêt aujourd'hui l'usage des ontologies pour le développement de systèmes à base de connaissances n'est plus à démontrer. Les chercheurs du Web ont adopté ce terme ontologie pour référer à un document (ou fichier) définissant d'une façon formelle les relations entre termes (Berners-Lee et al, 2001). Dans le cadre du Web sémantique, les ontologies sont utilisées comme noyau du système pour accéder à des informations structurées ainsi qu'à des règles d'inférence supportant le raisonnement

automatique. Ces ontologies offrent également la possibilité, pour un programme, de retrouver les différents termes désignant un même concept. Il s'agit pour ce cas spécifique, d'ontologies de type domaine.

Comme déjà évoqué plus haut, la structure du modèle de connaissances du projet GDST (**figure 1.1**) est principalement fondée sur des ontologies (Lefebvre et al, 2005). Il s'agit des ontologies du domaine, de la compétence, du document, du processus d'affaire, de l'employé et de l'entreprise.

- ***L'ontologie du domaine*** décrit les éléments généraux et les propriétés de l'organisation des connaissances d'un domaine. L'**annexe 1** (*domainOntology.daml*) contient un extrait de cette ontologie de domaine. L'**annexe 2**[4] montre une représentation graphique d'un extrait de cette ontologie.
- ***L'ontologie du document***, dite aussi ontologie de l'information, décrit les différentes sources d'information documentaire liées au domaine (télécommunications sans fil). Ses liens vers les autres ontologies (compétences, entreprise et processus d'affaires) peuvent également servir de filtre pour le service de diffusion en éliminant les documents qui ne correspondent pas aux critères de sélection des usagers.
- ***L'ontologie de la compétence*** représente le noyau du modèle des connaissances et est décrite plus loin. C'est un élément clé dans le processus de filtrage.
- ***L'ontologie du processus d'affaire*** décrit les composantes de l'activité professionnelle pour une entreprise.
- ***L'ontologie de l'employé*** permet de décrire les profils de tous les usagers du système qui sont les employés de l'entreprise.
- ***L'ontologie de l'entreprise*** présente cette dernière principalement en termes de rôles, qui sont joués par les employés.

[4] Pour des raisons de lisibilité, seule la relation de hiérarchie « *est-un* » a été représentée.

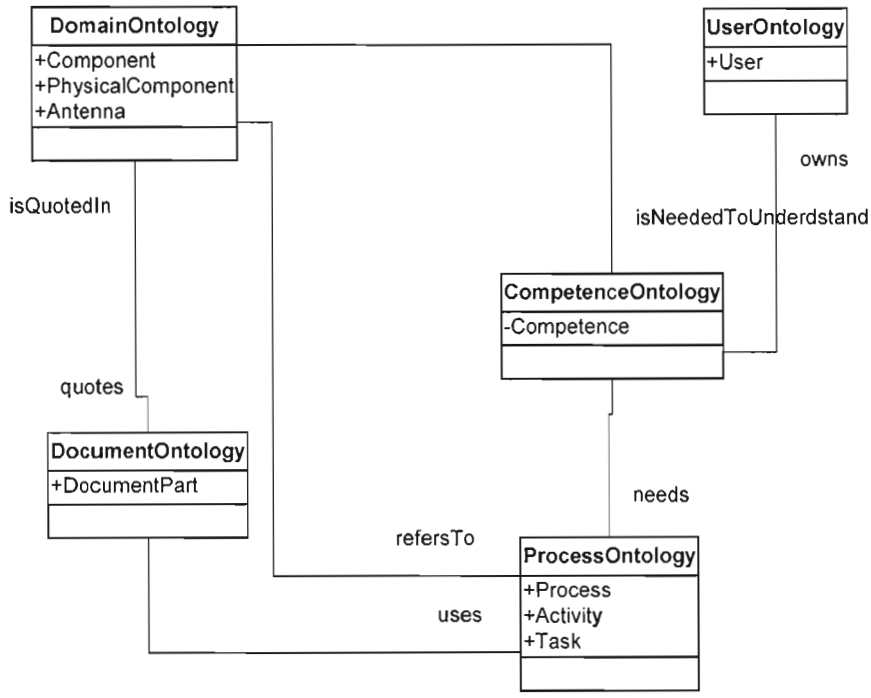


Figure 1.1 : Schéma d'organisation du système d'ontologie

La création, le contrôle et la maintenance des connaissances de l'architecture du projet GDST est une entreprise complexe. Dans l'objectif de faciliter ces tâches, différents services ont été utilisés ou développés dans le cadre de ce projet. Il s'agit d'un éditeur d'ontologies « *Protege-2000* » (Noy et al, 2001), d'un système d'analyse de documents « *SATIM* » (Biskri et al, 2002), d'une chaîne de traitements pour la maintenance de l'ontologie du domaine « *ONTOLOGICO*» (Gargouri et al, 2003, 2004), d'un système d'aide à l'annotation de documents « *AnnoCitaTool* » (Hung, 2003) et de services d'exploration des ontologies « *NORD* » (Hogue, 2003), de diffusion, et d'interrogation documentaire (Duc, 2003).

1.2.2 L'ontologie des compétences

Il est important, dans le contexte du projet GDST, de caractériser le niveau des compétences des employés. Toutefois, la compétence est difficilement mesurable en soi, mais s'exprime par des champs d'actions ou des éléments que nous pouvons hiérarchiser.

La compétence se définit par rapport à d'autres concepts décrivant la capacité, l'habileté et l'expertise d'un employé lors de l'exercice d'une activité professionnelle. Ces éléments et leurs relations constituent le modèle de compétences, tel que détaillé dans (Lefebvre et al, 2005). En particulier, la relation entre les employés et les compétences est non seulement spécifiée d'une façon directe, mais également d'une façon indirecte à travers les rôles exercés par ces employés. Cette relation représente les exigences auxquelles ces derniers doivent se conformer lors de l'exercice d'une activité pour l'entreprise. D'autres relations existent également entre les compétences elles-mêmes. Il s'agit des relations d'analogie, de généralisation et d'agrégation (Nkambou, 1996).

- Relation d'analogie : les compétences peuvent être similaires du point de vue de leur fonctionnalité, leur résultat ou leur définition.

- Relation de généralisation : décrit la relation entre deux compétences ; l'une étant plus générale que l'autre, au sens classique adopté dans le contexte du paradigme de la programmation orienté objet.

- Relation d'agrégation : cette relation établit le fait qu'une compétence est un composant d'une autre.

Dans le contexte du projet GDST, les compétences sont classées en deux groupes : les compétences spécifiques et les compétences transversales. Les premières sont liées directement à la réalisation des processus de travail tandis que les secondes intéressent potentiellement plusieurs processus de travail et sont généralement reliées aux connaissances du domaine. Une compétence est caractérisée par un niveau d'expertise. Le niveau 1 réfère au débutant et le niveau 5 à l'expert. De plus, les compétences peuvent être précisées par des verbes d'habileté qui peuvent présenter divers niveaux de complexité (Paquette, 2002).

Dans la mesure où l'acquisition de compétences est un processus continu, l'ontologie des compétences doit être mise à jour. Au cours de leur vie professionnelle, les employés entretiennent et améliorent leurs acquis par des formations ou de l'expérience pratique. Ainsi, de nouvelles compétences peuvent apparaître, se rapportant entre autres à de nouveaux concepts intégrés dans l'ontologie du domaine. Pour un usager à qui un document

introduisant de nouveaux concepts a été diffusé, on peut inférer l'existence de nouvelles compétences que l'on caractérisera au départ par un niveau d'expertise très bas et par un verbe d'habilité de type *connaître*.

1.2.3 Diffusion de documents

Une diffusion active et efficace consiste à filtrer sémantiquement les employés potentiels pour qui un document peut être utile. Après analyse du contenu sémantique d'un nouveau document, il est question de tenir compte des compétences, des habiletés, des niveaux d'expertise, des rôles et des tâches des employés, en vue de le diffuser vers les personnes aptes à le comprendre et intéressées à le consulter (Lefebvre et al, 2005; Achaba, 2003).

La diffusion de documents aux employés concernés pose de grands défis. En effet, elle vise à offrir une aide précise et concrète dans la réalisation de leurs activités sans les importuner par des informations qui ne rentrent pas dans leurs domaines d'intérêt ou qui ne s'accordent pas avec leurs compétences. La diffusion vise également à enrichir leurs connaissances professionnelles, améliorer leur productivité, procurer un gain de temps et améliorer leur efficacité en accédant aux informations pertinentes, au moment de la réalisation de tâches spécifiques et selon un format convivial.

La réalisation de tels objectifs doit alors se conformer à un ensemble de règles, définies dans une base de connaissances. À titre d'exemple, les documents doivent être filtrés selon la pertinence des concepts clés de ces documents par rapport à la tâche que l'utilisateur doit accomplir. Les documents doivent également tenir compte du niveau d'expertise de l'utilisateur ; les documents jugés de niveau trop élevé ou trop bas par rapport à son niveau sont éliminés. Ce jugement peut être fondé sur la notion de *zone proximale de développement* (Vygotski, 1934) qui est « *la distance entre le niveau de développement actuel tel qu'on peut le déterminer à travers la façon dont l'enfant résout des problèmes seul et le niveau de développement potentiel tel qu'on peut le déterminer à travers la façon dont l'enfant résout des problèmes lorsqu'il est assisté par l'adulte ou collabore avec d'autres enfants plus avancés* ». Ce concept a des conséquences pratiques en apprentissage. Il aide à caractériser le sens du développement et à fixer les objectifs de l'apprenant en se basant sur l'intervention du médiateur. Cette interaction se situe dans la zone proximale de développement de

l'apprenant (à un niveau supérieur à ce qu'il serait capable de faire seul) afin de lui permettre de dépasser ses compétences actuelles.

L'identification de cette zone nécessite la définition de certaines relations entre les compétences, et leur classement selon des niveaux hiérarchiques pour permettre d'en déterminer le niveau des unes par rapport aux autres.

L'architecture du service de diffusion, telle que décrite dans (**figure 1.2**), est principalement composée de deux parties, le *moteur de filtrage* et le *service de diffusion*. Le moteur de filtrage est chargé de sélectionner les usagers selon que le document en question appartient ou non à leur zone proximale de développement. Le moteur de filtrage fournit la liste des usagers filtrés au service de transmission des nouveaux documents.

Le raisonnement sémantique associé au filtrage est réalisé grâce aux relations implicites ou explicites entre les entités ou les classes représentées dans une ontologie. Il est fondé sur des processus d'inférence ainsi que des heuristiques en vue de rattacher à ce raisonnement sémantique, des comportements intelligents et naturels.

Par exemple : l'ensemble des personnes possédant l'expertise "*InstallNetwork_3*" est non seulement composé des personnes ayant exactement cette caractéristique, mais également les personnes possédant les expertises "*InstallNetwork_4*" et "*InstallNetwork_5*" qui sont d'un niveau plus élevé.

D'autres raisonnements heuristiques sont implémentés dans le système concernant la relation d'agrégation ("*isDecomposeIn*") et l'analogie ("*sameAs*") entre les compétences et les qualifications. Les caractéristiques des compétences relatives aux verbes d'habileté sont également prises en compte dans le raisonnement en fonction de la nature plus ou moins générale de ces verbes (Paquette, 2002).

Un service de recherche documentaire utilise également les fonctionnalités du moteur de filtrage pour fournir aux usagers les documents qu'ils recherchent, non seulement en fonction de spécifications précisées dans l'interface, comme la tâche ou le concept du domaine

impliqués, mais aussi en tenant compte des expertises de l'usager et de celles requises par un document candidat.

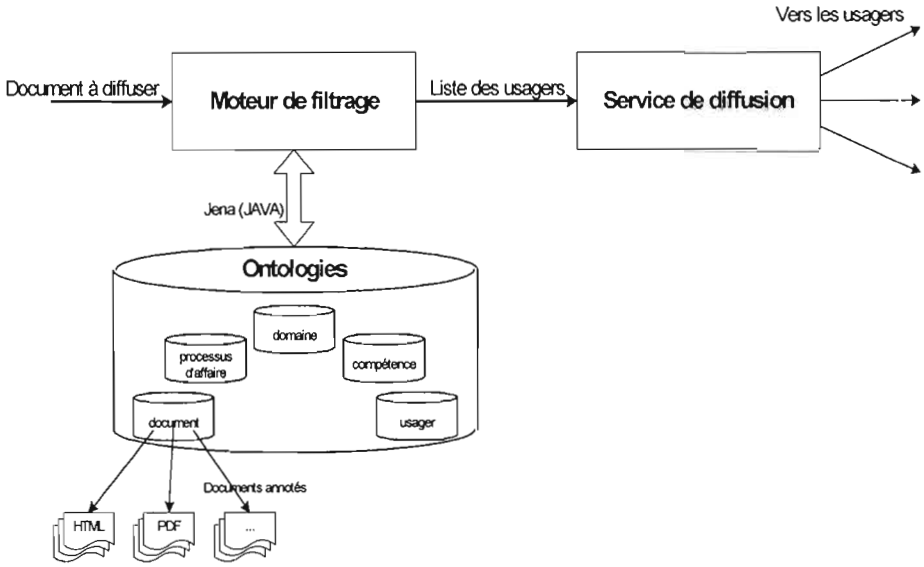


Figure 1.2 : Architecture du service de diffusion

1.2.4 Maintenance de l'ontologie du domaine

Les termes techniques relatifs à un domaine particulier changent et évoluent de façon perpétuelle. Par conséquent, les ontologies de domaine doivent être maintenues pour faire face aux incomplétudes et aux erreurs, ou encore s'adapter aux innovations dans le domaine. La complétude et l'exactitude des termes appartenant à un domaine spécifique au niveau de l'ontologie sont considérées comme un prétraitement d'une grande importance pour garantir l'utilité et la fiabilité de l'ontologie, et par conséquent, l'efficacité du service de diffusion (Gargouri et al, 2005).

Dans un premier temps, la conception de l'ontologie du domaine a été réalisée en se basant sur un ensemble d'outils intégrés dans une plateforme nommée SATIM et développée au laboratoire LANCI de l'UQAM. Le rôle de ces outils et principalement du module GRAMEXCO est d'assister la découverte de termes et concepts dans des documents. Ces outils sont aussi utilisés pour faciliter l'annotation de nouveaux documents et leur rattachement à l'ontologie du domaine.

Sur un autre plan, la maintenance de l'ontologie du domaine est assurée par un processus structuré sous forme d'une chaîne de traitements de la plateforme SATIM, nommée ONTOLOGICO et que nous détaillons dans le chapitre 3 (méthodologie et modèle proposé). Dans la mesure où il est difficile et fastidieux de repérer de nouvelles relations entre termes à partir de la simple lecture de données textuelles et d'évaluer leur pertinence par rapport à l'ontologie actuelle, l'utilisation d'un outil, du moins semi-automatique est incontournable. ONTOLOGICO vise justement à assister les experts de domaine dans leur tâche de maintenance de l'ontologie correspondante.

La maintenance de l'ontologie du domaine assure une meilleure représentation de la base documentaire à l'aide de connaissances à jour. En effet, le contenu sémantique des nouveaux documents (ainsi que des anciens) est mieux structuré grâce à l'intégration quasi-totale des termes clés du domaine dans l'ontologie.

1.3 Problématique de recherche

Au regard de l'évolution de la recherche sur l'ontologie dans le domaine informatique, il ressort qu'au fur et à mesure que la discipline se développe et devient mature, l'intérêt de la recherche et les questions que les auteurs adressent évoluent en conséquence. La progression se manifeste par un changement d'orientation, des problèmes théoriques relatifs aux ontologies vers des problématiques associées à une utilisation pratique d'applications basées sur des ontologies.

Bien que l'utilisation des ontologies pour supporter l'extraction de connaissances ait été déjà mise en évidence (Vargas-Vera et al, 2002), le potentiel de cette approche n'est pas totalement exploré. Nous croyons que le problème d'évolution des ontologies est la principale entrave justifiant la « réticence » de certains vis-à-vis de l'utilisation d'ontologies dans leurs systèmes.

Les documents relatifs à un domaine particulier changent et évoluent de façon continue. Par conséquent, les ontologies sont souvent sujettes à changement, parce que des incomplétudes ou des erreurs se sont révélées dans les versions précédentes, une nouvelle façon de modélisation du domaine a été préférée ou encore le domaine a changé. Les ontologies

doivent supporter ces révisions et faire en sorte que les nouveaux documents relatifs à une nouvelle version d'ontologie soient compatibles avec ceux qui précèdent.

En pratique, les méthodologies les plus utilisées pour la construction et la maintenance des ontologies sont généralement celles basées sur une approche descendante, c'est-à-dire dirigées par un modèle. Les méthodes ascendantes (du texte vers le modèle) sont beaucoup plus rares. Nous estimons que les textes, en tant que source principale des connaissances sur le domaine, sont actuellement sous-exploités. Un de nos principaux défis serait d'affronter la complexité du traitement des données textuelles en vue d'exploiter la richesse implicite de cette source de connaissances.

Il est important de penser à une façon de faire pour automatiser ce processus ou du moins, minimiser le traitement manuel relatif à la maintenance. En effet, il est irréaliste de s'attendre à ce que l'humain comprenne la totalité de l'ontologie et de ses interdépendances internes (Tallis et al, 1999). Il lui est difficile, voire même impossible, de repérer de nouvelles relations entre termes à partir de la simple lecture de données textuelles et d'évaluer leur pertinence par rapport à l'ontologie actuelle. Cette tâche cognitive est d'autant plus ardue que les nouveaux textes à analyser et l'ontologie existante sont de large taille. Ce problème se pose également lorsque la conceptualisation du domaine est ambiguë ou encore si l'utilisateur ne possède pas suffisamment d'expérience.

Étant donné la difficulté de maintenir les ontologies disponibles, celles-ci sont difficilement réutilisables et partageables, même lorsqu'elles sont exprimées selon le même formalisme et couvrent le même domaine. Cette contrainte justifie d'ailleurs le recours quasi systématique des ingénieurs de connaissances à la construction d'une ontologie de domaine à partir de zéro. En effet, les efforts manuels nécessaires pour réutiliser une ontologie existante et l'entretenir, seraient beaucoup plus coûteux.

Au cours de la dernière décennie, beaucoup de recherches ont été réalisées dans le domaine de l'ingénierie des ontologies. La majorité de ces recherches se sont concentrées sur les problèmes de construction. Toutefois, la gestion des changements et les mécanismes de maintenance doivent être abordés différemment. Jusqu'à date, il n'existe pas de méthodes consensuelles et de lignes de conduite répondant à cette problématique.

La complexité de cette problématique se justifie principalement par le fait qu'en pratique, l'élaboration d'ontologies relève plus du savoir-faire que de l'ingénierie. C'est ainsi que, lors du processus de mise au point d'une ontologie, chaque équipe de développement suit habituellement ses propres principes, ses critères de conception et ses étapes d'élaboration. L'absence de méthodes consensuelles entrave, d'une part, le développement d'ontologies communes et acceptées par les équipes et entre elles. D'autre part, elle nuit à l'extension d'une ontologie donnée à partir d'autres et à sa réutilisation.

Pour atteindre ces objectifs, nous sommes aujourd'hui confrontés à la problématique de l'acquisition de connaissances à partir de textes. L'ingénierie des connaissances se trouve face à des difficultés liées à la complexité de la langue et des textes ainsi que la nécessité de mettre en œuvre des méthodologies rigoureuses pour rendre la pratique de l'ingénierie des connaissances plus efficace et plus adaptée aux problèmes réels.

Plus spécifiquement, la problématique de la maintenance des ontologies de domaine implique deux sous problèmes fondamentaux. Le premier est relatif à l'extraction de termes spécifiques au domaine et pertinents à l'ontologie existante. Le second consiste à identifier des relations entre termes. Une variété de techniques de traitement du langage naturel, d'extraction d'information, d'apprentissage machine et d'analyse textuelle sont utilisées pour extraire des termes à partir d'un corpus. Il s'agit de techniques suffisamment matures et présentant des résultats prometteurs pour le domaine de la construction d'ontologie. Cependant, l'extraction de relations entre termes est un problème plus complexe et difficile à résoudre. Ceci constitue d'ailleurs la problématique de base qui affecte directement la performance des techniques de maintenance des ontologies.

Deux méthodologies classiques sont proposées pour l'analyse semi-automatique de larges données textuelles pour extraire des connaissances pertinentes, à savoir « *les méthodes numériques* » et « *les méthodes linguistiques* ». Ces deux techniques sont plutôt complémentaires. En raison de ses caractères sémiotique et linguistique, le traitement classique d'information est habituellement linguistique. En effet, un texte est considéré comme étant une succession de phrases qui doivent faire l'objet d'analyseurs linguistiques.

Cette approche semble être complètement naturelle dans la mesure où elle correspond, en théorie, au processus normal de lecture chez l'humain (Meunier, 1996).

Il est évident que les approches numériques permettent d'extraire de plus amples régularités dans le texte, comparativement aux approches strictement linguistiques (basées sur la grammaire). Les techniques numériques, en particulier celles basées sur des stratégies de classification, permettent un gain de temps considérable lors de l'exploration du corpus, par conséquent, elles sont essentielles lorsqu'on est confronté à un large corpus textuel. Par ailleurs, elles sont extrêmement utiles pour une détection rapide d'associations textuelles sémantiques. D'ailleurs, lorsque associées avec d'autres ressources, telles que les thesaurus, elles offrent une assistance précieuse pour des analyses globales.

Trois questions servent de fil conducteur dans ce travail de recherche : comment doit-on maintenir une ontologie face à un monde dynamique caractérisé par des données textuelles continuellement instables? Comment peut-on extraire, à partir de ces textes, des termes et des relations entre termes, qui soient pertinents pour une ontologie et assister à sa maintenance? Les ressources terminologiques sont-elles nécessaires pour compléter la richesse textuelle? Ces questions constituent les piliers de la stratégie que nous avons adoptée pour mener à bien cette thèse. Les objectifs préalablement fixés sont précisés plus loin.

1.4 Motivations cognitives

Le domaine de la maintenance des ontologies à partir d'analyses textuelles s'intéresse principalement aux processus permettant de découvrir des relations d'association entre termes, pertinentes au domaine. L'automatisation partielle des procédures de traitements de textes et d'usages de thésaurus constitue à notre avis une piste prometteuse pour s'attaquer à une bonne partie des problèmes auxquels fait face une maintenance d'ontologies.

Sur le plan cognitif, l'élaboration de systèmes d'assistance à la maintenance des ontologies nécessite l'intégration de processus de raisonnement et d'apprentissage, principalement dédiés à la découverte de relations sémantiques ou d'association entre termes à partir de données textuelles. Les motivations cognitives se justifient entre autres, par le fait que certaines des difficultés du Traitement Automatique du Langage Naturel (TALN) ne sont pas

dues à des causes contingentes comme la taille de la mémoire, la puissance des microprocesseurs ou la performance des algorithmes, mais bien à des conceptions théoriques sur le traitement de données textuelles.

Les sciences cognitives représentent *un ensemble de disciplines s'appliquant à analyser les comportements intelligents*[5] *(celui de l'homme, des animaux ou des machines) et à analyser les supports matériels qui paraissent conditionner ces comportements (le cerveau ou l'ordinateur, par exemple)* (Vignaux, 1992). Les sciences cognitives ont un impact indéniable sur les théories informatiques, notamment les théories de modélisation. Dans les travaux de recherche en informatique (informatique cognitive notamment), les modèles dits cognitifs sont aujourd'hui monnaie courante (modèles connexionnistes, modèles symboliques et modèles dynamiques). À la suite de ces modèles, on a assisté à l'émergence d'une catégorie de logiciels dits « intelligents » (systèmes experts, systèmes tutoriels intelligents, systèmes multi-agents...) qui manipulent les modèles suscités et simulent l'activité mentale pour résoudre certains problèmes du monde réel.

Cette thèse s'appuie sur l'avancement des connaissances dans plusieurs disciplines inter-reliées ; principalement le TALN, l'analyse sémantique, la représentation des connaissances, l'intelligence artificielle, la psychologie cognitive, etc. Ces disciplines s'inscrivent à des degrés divers dans les sciences cognitives et informatiques.

1.4.1 Représentation des connaissances en I.A

«*Les systèmes informatiques dits intelligents possèdent entre autres une base de connaissances à laquelle ils se réfèrent, pour intégrer les intrants nouveaux, prendre des décisions et enfin effectuer des raisonnements* » (Meunier, 1992). La base de connaissances constitue un maillon essentiel dans un système informatique dit intelligent. Qu'est-ce qu'une activité intelligente? Newell et Simon (Newell et al, 1976) avancent que toute activité intelligente de la part d'un humain ou même d'une machine nécessite trois éléments importants :

---

[5] L'intelligence est l'ensemble des facultés mentales permettant de comprendre les choses et les faits et de découvrir les relations entre eux.

- la représentation des aspects significatifs du domaine d'un problème, ou en d'autres termes, des connaissances pertinentes associées au domaine d'un problème,
- la définition (ou la description) des opérations applicables sur les représentations obtenues afin de générer les potentielles solutions au problème,
- des stratégies de sélection d'une solution parmi celles potentielles générées.

Confrontés à la complexité des problèmes qu'ils sont appelés à résoudre, les systèmes d'IA ont besoin non seulement de connaissances mais également de mécanismes et stratégies pour manipuler efficacement ces connaissances afin de fournir des solutions aux problèmes. Une représentation adéquate des connaissances à manipuler est une étape importante pour favoriser la performance et l'efficacité des systèmes. Luger (Luger, 2002) place d'ailleurs la représentation et la recherche des connaissances au cœur de la recherche moderne en IA.

Les approches de représentation des connaissances en IA, peuvent être regroupées sous deux grands paradigmes : le paradigme symbolique et le paradigme subsymbolique (ou connexionniste). En général, le choix du paradigme et de la forme de représentation pour un système donné, dépend du domaine d'application associé au système.

1.4.1.1 Le paradigme symbolique

L'approche symbolique consiste à se placer à un niveau supérieur de description permettant d'avoir une vision symbolique globale du fonctionnement. Ce paradigme, d'inspiration philosophique, a pour base *l'hypothèse du système symbolique physique* énoncée par (Newell et al, 1976). Il est caractérisé par l'utilisation de symboles pour représenter les connaissances, et par la formalisation de la manipulation des symboles. Selon ce paradigme, les approches de représentation de connaissances peuvent être « *orientées syntaxe* » ou « *orientées sémantique* ».

Les approches « *orientées syntaxe* » sont celles basées sur la logique propositionnelle, la logique des prédicats, une logique pour raisonnement non monotone ou la logique floue. Il s'agit de la manipulation de symboles de forme arbitraire, selon des règles portant uniquement sur cette forme et ne portant ni sur la forme matérielle de leurs référents ni sur leur sens (Harnad, 1990).

Pour ce qui est des approches « *orientées sémantique* », on sacrifie la rigueur (mais également la rigidité) du raisonnement logique au profit d'une simulation psychologique et linguistique inspirée du modèle humain. Il s'agit des approches basées sur les réseaux sémantiques, sur les graphes conceptuels, ou encore sur les réseaux bayésiens. Toutefois, Harnad (Harnad, 1990) soutient que ce n'est que par l'ancrage des symboles que nous pouvons déterminer la sémantique d'une représentation propositionnelle. Un modèle de l'IA classique manipule des structures de symboles qui sont dénués de sens pour le système. Ce dernier ne reconnaît que leur forme. Tout sens est attribué par une entité extérieure via une fonction d'interprétation.

1.4.1.2 Le paradigme subsymbolique

Le paradigme subsymbolique se base sur des modèles connexionnistes en proposant de modéliser aussi bien les processus perceptifs de bas niveau que ceux de haut niveau tels que la classification, la reconnaissance d'objets, la résolution de problèmes, la planification et la compréhension du langage (Smolensky, 1988).

Pour les partisans de l'approche subsymbolique, la représentation des connaissances est structurée en deux couches : une première, de description du système formel au niveau supérieur et une deuxième, d'interprétation sémantique au niveau inférieur. Les entités susceptibles de recevoir une interprétation sémantique sont des configurations d'activités sur un grand nombre d'unités du système, et les entités manipulées par des règles formelles sont l'ensemble des activations individuelles des cellules du réseau.

En IA, les modèles connexionnistes sont de plus en plus populaires grâce à leur puissance computationnelle et leur efficacité dans la résolution de certains problèmes relativement complexes. Pour le cas de la classification textuelle (qui nous intéresse le plus dans notre recherche), de nombreuses études empiriques fournissent des modèles plausibles, inspirés des modèles biologiques, mais qui opèrent en tant que « boite noire » ; en effet, on ne sait pas très bien comment les processus cognitifs de l'être humain opèrent la classification, en particulier

dans le domaine textuel qui relève à la fois de la perception visuelle, du langage et de la structuration des connaissances de l'agent cognitif.

Le modèle que nous proposons se base tant sur l'approche symbolique que l'approche connexionniste.

1.4.2 Analyse Sémantique

L'extraction de connaissances à partir de textes ne peut généralement se passer d'une discipline telle que l'analyse sémantique qui a pour objet la description des significations propres aux langues et leur organisation théorique. Même si nous affirmons que l'analyse sémantique joue un rôle primordial en ingénierie de connaissances, notre objectif n'est pas la compréhension automatique et complète de la documentation. Il s'agit plutôt d'analyser la documentation dans le cadre d'une tâche bien déterminée, celle de l'identification de relations conceptuelles entre termes. Ainsi les outils de TAL sont vus comme des « *outils d'aide à l'analyse des textes* ». Ceci nous amène à privilégier les outils et techniques effectuant des analyses partielles, mais robustes et fonctionnant sur des données textuelles réelles.

Sur le plan sémiotique, aucune, parmi les techniques proposées dans la littérature, ne semble capturer toute la richesse sémantique encapsulée dans les données textuelles, du moins, pas aussi efficacement que font les processus cognitifs de l'être humain. En effet, les architectures supportant ces techniques, ont rarement été élaborées sur la base de comportements humains, affectant ainsi, d'une certaine manière, les hypothèses relatives à la représentation, l'organisation, l'utilisation et l'acquisition de connaissances à partir de textes. Ce constat nous a amené à explorer les fondements de la « *Psychologie Cognitive* » en vue de mettre en évidence son efficacité à résoudre le problème de la sémantique.

1.4.3 Psychologie Cognitive

La cognition est un ensemble de processus intellectuels, à travers lesquels l'information est obtenue, transformée, stockée, retrouvée et utilisée. L'approche à adopter pour la maintenance des ontologies devrait s'inspirer de façon étroite des mécanismes intellectuels de compréhension, de production et d'apprentissage chez l'être humain (psycholinguistique).

Les recherches en psychologie cognitive montrent que la plupart des mots sont assimilés par la lecture[6] (Landauer et al, 1997). Étant exposé à des textes, un apprenant tente, tout le long de son processus de lecture, de raffiner graduellement la signification du mot en utilisant les cooccurrences conjointes de ce mot avec d'autres. Par exemple, en absence d'une définition explicite du mot « *micro-processeur* », l'apprenant est en mesure, à travers la lecture de textes, d'acquérir la signification du mot parce que celle-ci est confirmée dans le contexte dans lequel ce mot apparaît avec d'autres, tels que « *carte* », « *ordinateur* », « *électronique* », « *matériel* », « *unité centrale de traitement* », etc. Toutefois, une simple cooccurrence répétée d'un mot avec d'autres semble être insuffisante pour l'assimilation de sa signification. En effet, toutes les cooccurrences de tous les mots à travers le texte sont plutôt nécessaires.

En se basant sur cette hypothèse d'apprentissage du sens, nous supportons l'idée de l'application de la technique de classification de documents pour identifier des groupes de termes qui apparaissent ensemble et qui ont des relations sémantiques, ou du moins, des similarités sémantiques lorsque utilisés dans des contextes comparables. Nous détaillons davantage les apports de la classification dans la section « modèle proposé ».

1.4.4 Architecture Cognitive

Notre orientation vers une architecture cognitive est motivée par un objectif de mise en place d'un système intelligent, supportant les potentialités de l'humain. Cet objectif se rapporte, d'une certaine manière, au paradigme fondateur du test de Turing[7]. Les architectures cognitives supportent également l'objectif central de l'IA et des sciences cognitives ; à savoir la création et la compréhension de comportements intelligents mettant en application les potentialités de l'humain.

Par ailleurs, un problème central, auquel les concepteurs d'architectures cognitives se trouvent confrontés, est de faire en sorte que les agents puissent accéder à différentes sources de connaissances. Par exemple, les connaissances relatives à l'environnement sont saisies à travers la perception, les connaissances relatives aux implications de la situation courante

---

[6] Bien que la conversation orale joue un rôle important dans l'apprentissage de mots, cette dimension ne sera pas traitée car notre modèle traite plutôt les textes écrits.

[7] Ce test consiste à faire dialoguer un humain et le système et déterminer si l'humain peut déceler si le système n'est pas humain.

sont saisies à travers la planification, le raisonnement et la prédiction, les connaissances produites par les agents sont échangées via la communication et les connaissances induites à partir des expériences passées sont acquises à travers la souvenance et l'apprentissage. Plus ces caractéristiques sont supportées par l'architecture, plus cette architecture peut tirer profit de ces sources de connaissances en vue d'influencer son comportement.

Partant de ce principe, nous avons opté pour une architecture qui se base sur différentes sources de connaissances, à savoir les textes relatifs au domaine ainsi qu'une base terminologique spécifique à la langue (tel que Wordnet).

1.5 Objectifs de la recherche

L'objectif général de ce projet est de proposer un modèle permettant d'assister les ingénieurs d'ontologie dans leur tâche de maintenance d'ontologie de domaine en se basant principalement sur une analyse textuelle. Ceci revient à proposer un modèle permettant, à partir de l'analyse de nouveaux textes, d'identifier de nouveaux termes spécifiques à un domaine ainsi que leurs relations. La maintenance est ici vue comme la mise à jour incrémentale de l'ontologie au fur et à mesure que de nouveaux termes sont extraits de textes du domaine.

Au niveau cognitif, notre objectif consiste à explorer et à mettre à contribution la pertinence de certains travaux en sémantique cognitive, en analyses textuelles et en psycholinguistique dans leur application à l'informatisation des processus d'assistance à la maintenance. Notre démarche visera donc à mettre en valeur la complémentarité des différentes approches abordées et à démontrer comment ces travaux permettent d'opérationnaliser informatiquement les processus de repérage de termes et de relations entre termes.

1.5.1 Objectifs spécifiques

Plus spécifiquement, nos objectifs de recherches s'orienteront vers les points suivants :

- La problématique de la maintenance des ontologies de domaine implique deux sous-problèmes fondamentaux ; le premier est relatif à l'extraction de termes spécifiques au domaine et pertinents à l'ontologie existante. Le second consiste à identifier des relations entre termes.

- Le repérage de termes qui sont particulièrement spécifiques au domaine en question est un élément clé pour maintenir une ontologie. Cet objectif revient à éviter d'alourdir l'ontologie par des termes non pertinents par rapport au domaine. Le choix d'un corpus représentatif du domaine découle en partie, de cet objectif.
- Bien que l'ontologie soit spécifique à un domaine particulier, notre objectif est de proposer un modèle de maintenance qui soit applicable pour d'autres domaines, par le biais de simples procédures d'adaptation. Par conséquent, la solution proposée ne doit pas être excessivement dépendante d'un domaine particulier.
- L'analyse textuelle visée n'est pas spécifique à une langue particulière et par conséquent, ne doit prendre avantage des connaissances linguistiques, à l'exception de certaines procédures simples (telles que la lemmatisation, le filtrage de termes, etc.).
- Nous avons opté pour un modèle de maintenance d'ontologie de type monolingue. Nous n'aborderons pas la complexité impliquée par l'utilisation d'une ontologie multilingue. En effet, celle-ci suscite des réflexions cruciales, principalement attribuées au domaine de la traduction automatique, laquelle problématique s'éloigne en fait des objectifs de notre projet.
- Les méthodes linguistiques de génération d'hypothèses de relations entre termes génèrent souvent moins d'hypothèses de relations que les méthodes statistiques de regroupement. Elles proposent cependant des relations étiquetées, directement vérifiables en contexte. Selon notre point de vue, pour un cogniticien en phase de modélisation de connaissances, la complétude des hypothèses de relations qui seront proposées au cogniticien pour validation est un objectif prioritaire par rapport à l'étiquetage. En effet, repérer une relation entre deux termes à partir d'un corpus nécessite un effort cognitif beaucoup plus important que pour étiqueter une relation préalablement identifiée. Cet objectif nous conduit à assigner plus d'importance au taux de rappel de relations candidates, présentées au cogniticien pour validation, qu'à l'étiquetage automatique de relations.

1.5.2 Hypothèses de recherche

L'hypothèse générale du projet est formulée ainsi : le repérage de termes et de relations entre termes nécessaires à la maintenance d'une ontologie de domaine peut être assisté

efficacement en employant certaines techniques issues de l'intelligence artificielle pour des fins d'analyses textuelles et aussi d'analyses de ressources terminologiques.

À la lumière des objectifs spécifiques que nous nous sommes fixés, nous avons formulé certaines hypothèses afin de guider notre démarche méthodologique, à savoir :

H1 *Les textes propres à un domaine sont une source principale pour enrichir le modèle de l'ontologie*

Partons des définitions retenues pour le concept d'ontologie ; il est important de souligner la présence, quasiment systématique, de l'aspect «*conceptualisation partagée*» dans ces définitions. Cet aspect nous a amenés à faire des textes, propres à un domaine particulier (ou à une entreprise, etc.), une source principale pour enrichir le modèle de l'ontologie.

Les textes jouent un rôle central dans l'appropriation et la transmission des connaissances ; nous croyons qu'ils sont indispensables dans tout processus de maintenance d'ontologie de domaine. L'exploitation des textes est aujourd'hui facilitée par la numérisation de quantités de plus en plus importantes de données documentaires.

Nous estimons que les textes, en tant que source principale de connaissances sur le domaine, sont actuellement sous-exploités. Un de nos principaux défis serait de mettre les textes et plus particulièrement la documentation spécifique à un domaine, au cœur de l'ingénierie des connaissances et d'affronter la complexité du traitement des données textuelles en vue d'exploiter la richesse implicite de cette source de connaissances.

H2 *D'autres sources de connaissances sont nécessaires pour découvrir des relations entre termes*

Notre modèle est fondé sur l'hypothèse que l'extraction de connaissances à partir de textes ne peut se contenter d'un traitement statistique (ni même linguistique) de données textuelles pour accaparer toute leur richesse sémantique. En effet, certaines connaissances implicites, spécifiques au domaine, ne peuvent être extraites à partir du corpus. Cette hypothèse est en fait, fondée sur la base de fondements cognitifs cohérents ; lors d'un processus de lecture de textes par un expert du domaine, ce dernier fait souvent usage de certaines de ses propres

connaissances du domaine (qu'on ne retrouve nécessairement pas dans les textes), pour repérer des relations d'associations conceptuelles entre termes. Il fait également recours à un dictionnaire ou un thésaurus pour compléter ses connaissances par certaines informations, telles que les définitions, les synonymies, les hyperonymies, etc.

L'expert fait enfin appel à d'autres connaissances qui ne sont pas intégrées dans les thésaurus et dont la spécification est impensable, vu leur nature purement humaine, qui ne peut être reproduite par la machine.

Le modèle que nous proposons ne vise pas à remplacer complètement l'humain pour la découverte de relations entre termes, mais plutôt d'assister les experts du domaine dans cette tâche. Nous avons ainsi opté pour un modèle qui soit un compromis acceptable mais qui soit suffisamment fonctionnel. Le modèle proposé se base ainsi sur différentes sources de connaissances, à savoir les textes relatifs au domaine, mais aussi des données terminologiques (thésaurus, dictionnaire électronique, base terminologique, etc.) en vue de faire intégrer, le mieux possible, les connaissances du domaine mais aussi certaines connaissances reliées plutôt à la langue, au sein de l'ontologie.

H3 *La cooccurrence est un critère de choix pour le repérage de relations entre termes*

La proximité sémantique entre les termes repose principalement sur la cooccurrence d'un ensemble de termes à travers différents segments d'un corpus. La base théorique de cette hypothèse repose sur celle formulée par Harris (Harris, 1968) ; on peut classer les divers sens d'un terme en fonction des constructions auxquelles ce dernier participe. Des termes qui ont des distributions comparables ont souvent un élément de sens commun. Partant de cette hypothèse, nous avons privilégié l'emploi de la classification textuelle pour repérer, dans un premier temps, des regroupements de termes sémantiquement reliés. Nous argumentons davantage ce choix théorique dans la section « modèle proposé ».

H4 *Les significations de termes contribuent activement au repérage de relations entre termes*

Les définitions des termes constituent une source d'information importante, permettant de contribuer activement au repérage de relations entre termes. En effet, en présence d'un regroupement de termes qui sont potentiellement reliés (identifiés à l'aide de la classification

par exemple), les définitions de chacun de ces termes peuvent enrichir le contenu informationnel de cette classe et concourir, par conséquent, à la découverte de relations conceptuelles plus fortes entre des couples de termes. Cette contribution consiste en fait, en un rapprochement entre les groupes de termes clés, faisant partie de chacune des définitions des termes en question. Cette hypothèse met en évidence l'importance de l'utilisation d'une ressource terminologique (dictionnaire électronique, thésaurus, etc.).

Dans le cadre de la représentation de connaissances et de la signification de lexique, « *les vecteurs conceptuels* » (VC) (Lafourcade et al, 2001) ont prouvé leur efficacité pour construire des taxonomies hiérarchiques et mettre en évidence des relations entre termes. Les vecteurs conceptuels sont généralement utilisés avec certaines mesures pour prendre des décisions par rapport à la qualité d'association entre les termes. Ils sont d'ailleurs, largement utilisés en recherche d'information (Salton et MacGill, 1983) ainsi qu'en représentation de signification par le modèle ISL (Deerwester et al, 1990), relatif aux études d'Analyse Sémantique Latente (ASL) en psycholinguistique.

La question qui se pose à cet effet est la suivante : doit-on composer les VC à partir des items lexicaux de la signification, ou plutôt des relations sémantiques et des relations d'association extraites d'un thésaurus ? Cette question sera traitée dans la **section 3.2.6** (Vecteurs Conceptuels et thésaurus).

H5 : *L'intervention d'un expert est une opération incontournable*

Notre objectif consiste à construire une méthodologie supportant l'utilisateur lors de sa découverte de relations entre termes qui sont potentiellement utiles pour la maintenance de l'ontologie. Il est relativement facile de remettre en cause des systèmes automatiques prétendant accomplir cette tâche sans biais ou imperfections. Il semble plus raisonnable, et plus réaliste en terme de faisabilité, de suivre un processus plutôt semi-automatique, impliquant une simple intervention d'un expert du domaine, à travers certaines étapes, et spécialement pour la validation des résultats.

H6 : *L'Indexation Sémantique Latente présente de meilleurs résultats quand elle s'applique sur des classes de termes plutôt que sur tous les termes d'un document:*

L'Indexation Sémantique Latente est un des principaux choix théoriques dans notre méthodologie. Cette hypothèse repose sur l'idée que l'application de la technique d'Indexation Sémantique Latente sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du corpus, possède l'avantage de réduire la matrice de cooccurrence de termes dans les documents à une dimension raisonnable. Cette réduction sera détaillée dans la ***section 3.2.4***. Nous présumons que ce choix méthodologique constitue un remède à la difficulté d'identifier dans la théorie, une dimension adéquate et précise de cette matrice. En effet, il est facile de prouver qu'une dimension immense pourrait empêcher l'émergence de suffisamment de relations sémantiques entre les termes, et aussi, une dimension trop réduite pourrait entraîner une grande perte d'information.

CHAPITRE II

INGÉNIERIE DES ONTOLOGIES

Dans plusieurs disciplines aujourd'hui, on assiste au développement d'ontologies standardisées pour faciliter les échanges entre experts et expliciter les connaissances du domaine. L'ingénierie des ontologies suscite un engouement sans précédent, non plus seulement dans les laboratoires de recherche en Intelligence Artificielle (IA) mais de plus en plus parmi les experts de divers domaines.

Dans ce chapitre, nous présentons le processus de conception d'ontologie et les approches de maintenance. Dans la mesure où la problématique de recherche s'intéresse au processus de conceptualisation, un état des lieux est fait sur les techniques d'extraction de termes (simples et complexes) et des relations entre termes à partir de textes. Les étapes d'ontologisation, d'opérationnalisation et d'évaluation ne seront pas traitées dans cette thèse.

Avant de passer en revue ces techniques, nous rappelons la définition de « terme » et distinguons « terme » de « concept ».

2.1 Qu'est-ce qu'un terme ?

Cette question vise à distinguer, *concept*, *terme* et *instance*, spécifiés à l'intérieur d'une ontologie. Un concept, tel que défini par le Petit Robert, est une représentation mentale, générale et abstraite d'un objet. Il peut être exprimé par un terme, un symbole ou autre. Ainsi, les termes représentent la manifestation linguistique d'objets réels ou immatériels qui partagent des propriétés communes. Par exemple, les termes *voiture*, *auto*, *automobile*, *bagnole*, ou encore *car* (en anglais) réfèrent au même concept (ou catégorie) "VOITURE ". Quant à la relation concept/instance, celle-ci est analogue à celle de classe/objet dans le

formalisme UML[8]. L'instance représente ainsi, un objet particulier dans le monde ou encore un membre de l'extension d'une catégorie. Par exemple, la *Chevrolet* de mon voisin est une instance du concept " VOITURE ".

L'emploi même de *terme* est ambigu dans la littérature spécialisée. De nombreux spécialistes de domaine envisagent le terme sans considérations théoriques véritables : pour eux, il s'agit d'une entité formelle associée à un contenu informationnel (« étiquette » associée au nœud d'une ontologie ; « mot-clé » pour la recherche d'information ; « entrée » d'un index, etc.) (Ibekwe-San Juan et al., 2005).

Le concept est défini indépendamment de la langue. Il indique une notion universelle qui ne tient pas compte de la diversité des langues. Le concept peut être vu comme un signifié normé. La norme en question est induite par l'ensemble des conventions et usages en vigueur dans un champ disciplinaire donné. (Rastier, 1991) définit le concept comme :

> *« ... un sémème construit, dont la définition est stabilisée par les normes d'une discipline, de telle façon que ses occurrences soient identiques à son type. La validité conventionnelle de ses normes disciplinaires permet la traduction des concepts, qui échappent de ce fait à la variété des langues comme à la diversité des contextes ».*

Dans cette section, l'accent sera mis sur les aspects linguistique, sémantique et conceptuel du terme. Ces aspects influent considérablement les choix de conception d'une ontologie.

2.1.1 L'aspect linguistique du terme

On distingue dans le vocabulaire d'une langue les termes simples et les termes construits à partir de termes existants. En français par exemple, on distingue différents types de créations de termes ; les plus fréquents sont les créations morphologiques, syntagmatiques, sémantiques et par réduction.

---

[8] UML : « langage de modélisation unifié ») (« *Unified Modeling Language* ») est un langage graphique de modélisation des données et des traitements. UML est la référence en modélisation objet.

- **Les créations morphologiques**

Les créations dites morphologiques, comprennent les dérivations affixales, la dérivation zéro et les dérivations par composition.

- *Les dérivations affixales*

Les langues de spécialité se distinguent de la langue générale par un certain nombre de traits syntaxiques qui sont remarquables, par rapport à l'usage général, notamment par une fréquence d'apparition plus élevée. Ces traits résultent des opérations dites de dérivations. Il s'agit, à partir d'un système lexical déjà existant, de fabriquer de nouvelles séries de mots destinées à désigner de nouveaux objets et procédés techniques.

Ainsi, le procédé de suffixation consiste à associer à un mot de base un suffixe dont la fonction est de changer le mot de classe et d'orienter le mot nouveau dans un champ lexical différent du premier. Le procédé de suffixation en « -age » est le plus utilisé. (Par exemple, le verbe *monter* qui se transforme en *montage*). Par contre, la préfixation se différencie de la suffixation par le fait qu'elle ne détermine pas un changement de la classe du mot. On peut citer par exemple, le préfixe « dé » à valeur négative (*faire*, *défaire*) et le préfixe « re » équivalent à « de nouveau » (*faire*, *refaire*).

- *La dérivation zéro*

Ce procédé, connu aussi sous le nom de conversion, consiste à passer d'un mot d'une catégorie syntaxique à une autre, sans changer sa structure morphologique. Il s'agit donc de former un mot nouveau sans utiliser des préfixes, des suffixes, etc., mais en changeant simplement sa catégorie syntaxique. Par exemple, le mot « *pour* » peut être employé dans différents contextes :

- « Je suis pour la liberté d'expression » (« *pour* » ici est une préposition).
- « Peser le pour et le contre » (« *pour* » ici est un nom).

- *La composition*

Dans la langue de spécialité, un grand nombre d'unités lexicales comportent deux ou plusieurs éléments. La composition en français est productive, mais moins productive que celle dans les langues anglaise et germanique où la plupart des unités lexicales des langues de spécialité comportent deux ou plusieurs éléments (Binet et al, 1987). Il y a deux sortes de mots composés en français : ceux formés sur des bases françaises, tels que *porte-document* qui est du type complément d'objet du verbe, ceux formés sur des bases latines ou grecques, telles que *misanthrope*.

- **Les créations syntagmatiques**

Ce sont des créations linguistiques qui consistent en des collocations[9] de plusieurs lexèmes, en un seul lexème complexe. Les créations peuvent être de type : substantif (nom) + adjectif, ou encore, de type synapsie, c'est-à-dire une unité de signification composée de plusieurs morphèmes lexicaux et reliés par diverses expansions. La synapsie *machine à coudre* se distingue du mot composé *timbre-poste*.

- **Les créations par réduction :**

Il s'agit d'un procédé de réduction syllabique qui répond d'une part, à un besoin d'économie linguistique, d'autre part, à un intérêt au concept devenu plus essentiel que le mot. On trouve par exemple les acronymes et les sigles : l'acronyme est un mot formé d'initiales, de lettres ou de syllabes appartenant à plusieurs mots, exemple : OVNI (Objet Volant Non Identifié). Le signe O.N.U. désigne lui, plus un concept qu'une simple abréviation du terme *Organisation des Nations Unies*. Dans ce type de création on peut classer aussi les mots dits mots valises (Binet et al, 1987), qui résultent d'une sorte de télescopage linguistique entre deux autres mots ; c'est le cas par exemple du mot *bionique* où les composants sont *biologie* et *électronique*.

---

[9] Collocation : présence d'au moins deux unités linguistiques distinctes dans un énoncé, qui sont liées par un rapport de proximité syntaxique et de relative dépendance.

- **Les créations sémantiques :**

Elles consistent souvent en une migration de mots de la langue générale vers les vocabulaires spécialisés, pour leur donner une teinte particulière : par exemple dans la terminologie de l'atome, le mot *âge* désigne l'aire de ralentissement d'un neutron, calculée par la méthode de l'âge de Fermi. Parmi les créations sémantiques, on trouve aussi les métaphores et les métonymies (appellations par analogie). Par exemple, dans la terminologie de l'atome, *pile atomique* désigne un empilement.

2.1.2 L'aspect sémantique et conceptuel du terme

Les mots ont été analysés par Platon et Aristote selon un double point de vue ; d'abord, la différence entre *signifiant* et *signifié*, ensuite, la différence entre la signification et la référence. Pour illustrer ces notions, nous évoquons «le triangle sémiotique», le problème de sens et les relations sémantiques entre termes.

- **Le triangle sémiotique :**

Pour mettre en évidence la relation entre termes et concepts, plusieurs linguistes comme (Rastier, 1990 ; Bessé, 1990 ; Pierce, 1978) se réfèrent couramment à un modèle triadique, décrit par le triangle sémiotique d'Ogden et ses collaborateurs (Ogden et al, 1969) :

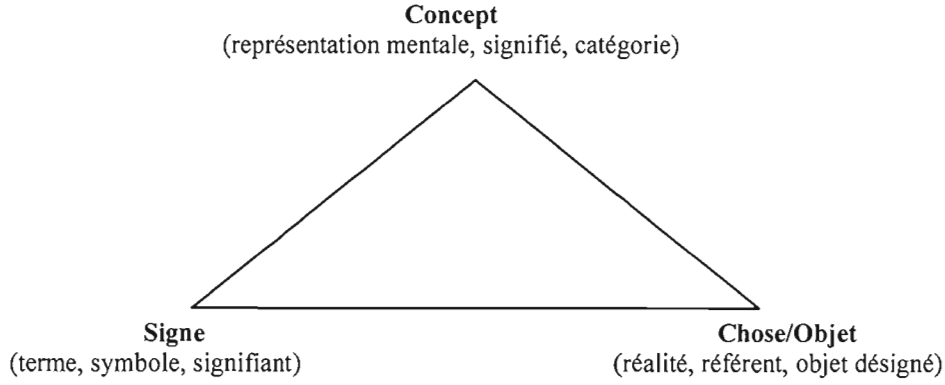


Figure 2.1: Le triangle sémiotique

Dans ce triangle, le signe désigne l'objet et symbolise le concept. Ce dernier se rapporte à l'objet, c'est-à-dire le référent. Les mots entre parenthèses sont des mots équivalents.

On distingue le référent d'un signe et son signifié (son sens). Le référent est identifié par rapport à une catégorie appartenant au domaine d'étude. Pour référer à des objets, la langue dispose de tout un matériel linguistique permettant par exemple de faire une description définie (article défini + substantif) ou de désigner un objet (démonstratif + substantif, déterminant + substantif). Quant au sens, il désigne le processus sous-jacent (se déroulant à l'intérieur du système qui parle et qui agit sur les référents désignés par ses mots) qui permet d'identifier son référent.

- **Le problème du sens :**

D'un point de vue grammatical, toute unité linguistique du vocabulaire isolée possède un sens de base appelé « sens grammatical ». C'est un sens linguistique abstrait donné par exemple par le sens dérivationnel, la flexion (singulier, pluriel), la mise dans le temps (présent, passé, ..), etc. Par exemple le mot *blanc* est un adjectif mais *blancheur* est un substantif. Ces 2 mots ont des sens différents. De même, pour le verbe *aimer*, on peut utiliser les flexions suivantes : *j'aimerai*, pour exprimer le contexte du temps (le futur), ou encore *nous aimons*, pour exprimer la pluralité.

Par ailleurs, le sens des unités linguistiques est sensible au contexte. En d'autres termes, il est inconcevable de faire abstraction du milieu où on parle. De plus, dans un texte écrit, les mots ne sont pas isolés, ils entretiennent des rapports avec les autres mots qui les entourent dans l'énoncé : le contexte. C'est le cas par exemple des mots polysémiques et les homographes (mots ayant des signifiants et des formes graphiques identiques).

Par rapport à un domaine de spécialité, le lexique de termes désigne un objet ou une opération d'une façon relativement moins ambiguë. Par exemple, dans le domaine des télécommunications le terme réseau n'est pas équivoque. Bien que son type ne soit unique, cela n'entraîne pas de confusions. On peut donc définir un terme d'un domaine de spécialité comme étant un signe linguistique qui désigne un concept du domaine de la façon la moins ambiguë possible (plus un terme est spécialisé moins sa polysémie est étendue).

2.2 Conception d'ontologies de domaine

Depuis le début des années 90, on a assisté à l'émergence d'un certain nombre de méthodologies pour le développement et la maintenance des ontologies. Dans bien des cas, elles sont le reflet des expériences personnelles de construction d'ontologie. Parmi les méthodologies les plus connues, nous citons : *TOVE* (*TOronto Virtual Enterprise*) (Gruninger et al, 95), *Enterprise Model Approach (Uschold et al, 1995)*, *METHONTOLOGY* (Gomez-Perez et al, 1996) et *IDEF5* (KBSI, 1994). D'autres approches qui, à défaut d'être des méthodologies à part entière de développement d'ontologies apportent un éclairage et proposent de grandes lignes directrices pour le développement d'ontologies, telles que : *Ontolingua* (Farquhar et al, 1997), *ONIONS* (*ONtological Integration Of Naive Sources*) (Steve et al, 96), *Mikrokosmos* (Mahesh, 1996), *SENSUS* (Swartout et al, 1997) etc.

2.2.1 Processus de conception d'ontologie

Les ontologies étant destinées à être utilisées comme des composants logiciels dans des systèmes opérationnels, leur développement doit s'appuyer sur les mêmes principes que ceux appliqués en génie logiciel. En particulier, les ontologies doivent être considérées comme des objets techniques évolutifs et possédant un cycle de vie qui nécessite d'être spécifié. Bien qu'aucune méthodologie générale n'ait pour l'instant réussi à s'imposer, de nombreux principes et critères de construction d'ontologies ont été proposés. Plusieurs de ces méthodologies s'entendent sur un ensemble de processus (**figure 2.2**) guidant l'ingénieur d'ontologie dans toutes les étapes de la construction.

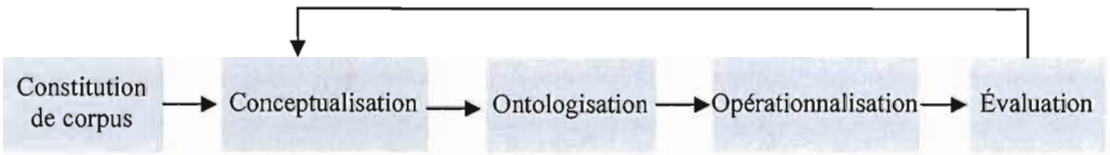


Figure 2.2 : Processus de construction d'ontologie

Nous détaillons dans ce qui suit chacune de ces étapes :

2.2.1.1 Constitution d'un corpus

Le repérage de termes qui sont particulièrement spécifiques au domaine en question est un élément clé pour créer ou maintenir une ontologie. Le choix d'un corpus représentatif du domaine restreint le champ de recherche de termes spécialisés. La tâche de construction du corpus est à la fois primordiale et délicate. Puisque, d'une part, le corpus est l'une des sources d'information essentielles et que, d'autre part, il demeurera, une fois le processus achevé, l'élément de documentation de l'ontologie construite.

La constitution d'un corpus représentatif du domaine suppose que ce domaine soit préalablement délimité. Un domaine n'est pas seulement défini par le champ de connaissances qu'il couvre, mais aussi par le point de vue sous lequel les utilisateurs de l'ontologie considèrent ce champ de connaissances (Uschold et al, 1995). En effet, le corpus constitué pour construire une ontologie des biotechnologies ne sera pas le même si le public visé est constitué de biologistes ou si le public visé ne possède que des connaissances générales dans ce domaine, en particulier en ce qui concerne la granularité du corpus et des connaissances modélisées. Par ailleurs, des choix liés aux contextes d'usage de l'ontologie doivent préalablement être effectués.

La collecte doit se faire avec l'aide des spécialistes et en fonction de l'application cible visée. Il convient en effet de s'assurer auprès des spécialistes que les textes choisis ont un statut suffisamment consensuel pour éviter toute remise en cause ultérieure.

Le critère de la taille est évidemment important, même s'il est impossible de donner un chiffre idéal. Le corpus doit être suffisamment « gros » pour s'assurer que des outils de traitement de la langue basés sur des statistiques puissent s'en servir de façon efficace. Mais il doit être suffisamment petit et redondant pour pouvoir être appréhendé de façon globale par l'analyste, même à l'aide d'outils de TAL. Une fourchette entre 50000 et 200000 mots semble raisonnable (Bourigault et al, 2003).

2.2.1.2 Conceptualisation

La conceptualisation consiste à identifier les connaissances du domaine et à expliciter la nature conceptuelle des connaissances en termes de concepts, termes exprimant ces concepts, relations entre termes, propriétés des concepts et des relations et axiomes. C'est aussi une représentation mentale d'un objet ou d'une idée ou un travail d'abstraction pour exprimer un principe général. L'identification de ces connaissances se base sur l'analyse de documents ainsi que sur des interviews d'experts de domaine.

Lors de ces analyses, il est important de distinguer les connaissances spécifiques au domaine de celles qui, bien que présentes dans le corpus, ne participent qu'à l'expression des connaissances du domaine. Les connaissances non spécifiques au domaine sont plutôt détaillées dans d'autres ontologies qui peuvent formellement être intégrées à l'aide de la technique d'importation. Par exemple, la conceptualisation du domaine du transport aérien fait référence à des connaissances non spécifiques au domaine telles que celles relatives à la géographie.

L'analyse textuelle est généralement assistée par des outils de TALN en vue de détecter les termes et les structures sémantiques présentes dans le corpus. L'analyse de corpus ne peut suffire à elle seule à spécifier la sémantique du domaine. Les interviews, les brainstormings menés avec des experts de domaine doivent en effet générer la sémantique différentielle des concepts. L'échange entre experts est le meilleur moyen de faire émerger une sémantique claire et non ambiguë (Fernandez al, 1997). Toutefois, les connaissances humaines sont essentiellement subjectives et ne seront pas exprimées de la même façon par tous les experts. Il convient alors de réaliser une normalisation sémantique consensuelle afin d'objectiver les connaissances.

Le processus de conceptualisation mène ainsi à la construction d'un modèle conceptuel qui décrit en langage naturel ou semi formel, au travers des éléments terminologiques et sémantiques les connaissances du domaine. Pour être utilisable par la machine, il convient de formaliser le modèle obtenu. C'est l'objet du processus d'ontologisation. Le cœur de nos travaux porte plutôt sur le processus de conceptualisation. Nous y reviendrons en détail dans les chapitres 3 et 4.

2.2.1.3 Ontologisation

L'ontologisation est une traduction dans un certain formalisme de connaissances exprimées a priori en langage naturel. Ce processus consiste à structurer et à formaliser autant que possible la conceptualisation pour construire une ontologie spécifiant la terminologie et la sémantique du domaine à travers un modèle doté d'une sémantique formelle (mais non opérationnelle). Selon (Gruber, 1993), cinq critères doivent être respectés afin d'atteindre les objectifs généraux des ontologies :

- **la clarté et l'objectivité des définitions,** indépendamment de tout choix d'implémentation ;
- **la cohérence des axiomes**, c'est-à-dire leur consistance logique ;
- **l'extensibilité d'une ontologie**, c'est-à-dire la possibilité de l'étendre sans remise en cause ;
- **la minimalité des postulats de formalisation** afin d'assurer une bonne portabilité ;
- **la minimalité du vocabulaire**, c'est-à-dire l'expressivité maximale de chaque terme.

La construction de la hiérarchie de concepts débute par les concepts spécifiques. Ces derniers sont ensuite regroupés selon des concepts plus généraux (Uschold, 1996). Cette hiérarchisation doit s'accorder avec les propriétés des concepts et des relations et être cohérente avec les intensions et extensions des concepts. Le respect de la sémantique du domaine doit être assuré par un engagement ontologique, notion initialement proposée par (Gruber, 1993) comme un critère pour réaliser une spécification partagée d'un vocabulaire. L'engagement ontologique consiste à associer à chaque concept son extension et à manipuler ce concept conformément au sens prescrit par cette extension (Guarino, 1994).

Bachimont (Bachimont, 2000) distingue l'engagement sémantique de l'engagement ontologique. Le premier permet, à travers des principes différentiels, de préciser le sens des concepts (concepts sémantiques) de manière non ambiguë. Ainsi, deux concepts sémantiques ne sont identiques que si leurs interprétations sont les mêmes, conformément aux principes différentiels utilisés. L'engagement ontologique associe plutôt des extensions à des concepts (concepts formels). Deux concepts formels sont identiques s'ils ont une même extension et

une même intension. Les quatre principes différentiels proposés par Bachimont sont les suivants :

- **Principe de similarité avec le père** : un concept partage l'intension de son concept père[10];
- **Principe de différence avec le père** : l'intension d'un concept est différente de celle de son concept père ;
- **Principe de sémantique unique** : une propriété est commune aux concepts frères issus du même concept père mais s'exprime différemment pour chaque frère. Par exemple, les concepts HOMME et FEMME portent la propriété « sexe » héritée de leur concept père HUMAIN, mais cette propriété vaut « masculin » chez HOMME et « féminin » chez FEMME ;
- **Principe d'opposition** : les frères doivent tous être différents, sinon il n'y aurait pas besoin de tous les définir.

D'autres caractéristiques des concepts ont été proposées par (Guarino al, 2000 ; Welty et al, 2001) pour structurer des ontologies en imposant certaines contraintes sur l'utilisation des liens de subsomptions.

- **L'identité** : un concept porte une propriété d'identité si celle-ci permet de conclure quant à l'identité de deux instances de ce concept. Cette propriété peut porter sur des attributs du concept ou sur d'autres concepts. Par exemple, le concept de CLIENT porte une propriété d'identité liée au numéro du client. Deux clients sont identiques s'ils ont le même numéro.

- **La rigidité** : une propriété est rigide, si et seulement si, elle est essentielle pour toutes les instances d'un concept. Une propriété est par contre anti-rigide si, et seulement si, elle n'est nécessaire pour aucune de ses instances. Par exemple, la propriété *Humain* est rigide, alors que la propriété *Client* est anti-rigide (on peut toujours trouver des situations ou un individu qui a été client peut ne plus l'être).

[10] Un concept père est un concept plus général (exemple : ANIMAL est le concept père de CHEVAL)

- **L'unité** : un concept composé de plusieurs concepts est un concept unité si, pour chacune de ses instances, les différentes parties de l'instance sont liées par une relation qui ne lie pas d'autres instances de concepts. Par exemple, les deux parties d'une chaise roulante, chaise et roues sont liées par une relation « *roulante* » qui ne lie que cette chaise et ces roues.

- **La dépendance :** un concept C1 est dépendant d'un concept C2 si pour toute instance de C1 il existe une instance de C2 qui ne soit ni partie ni constituant de l'instance de C1. Par exemple, PARENT est un concept dépendant de ENFANT (et inversement), car l'existence d'un parent suppose celle d'un enfant, mais COUTEAU et MANCHE ne sont pas dépendants, car le manche fait partie du couteau.

Le processus d'ontologisation conduit ainsi à la construction d'une ontologie mais celle-ci ne peut, telle quelle, être utilisée dans un SBC, du fait de l'absence de sémantique opérationnelle des représentations qu'elle contient. Le modèle conceptuel structuré doit en effet être exprimé dans un langage formel de représentation d'ontologies.

2.2.1.4 Opérationnalisation

Les connaissances représentées dans une ontologie, n'ont habituellement pas de sémantique opérationnelle, car leur représentation doit rester neutre vis-à-vis des usages qui seront faits de l'ontologie. Décrire les connaissances en termes de concepts, de relations et de propriétés sur ces concepts et relations ne suffit généralement pas pour atteindre l'objectif opérationnel d'un SBC. L'utilisation opérationnelle d'une ontologie nécessite en fait, le passage vers un processus d'opérationnalisation qui permettra, au delà de la spécification des connaissances au niveau conceptuel, de préciser la sémantique opérationnelle de la forme de l'ontologie qui sera utilisée dans un SBC.

La sémantique opérationnelle décrit la façon dont les représentations dotées de cette sémantique opèrent sur d'autres représentations, leur mise en œuvre, leur concrétisation dans une action pour atteindre un but. Elle dépend donc de l'objectif opérationnel visé par l'utilisation de l'ontologie et ne peut généralement pas être incluse dans l'ontologie. Comparée à la sémantique formelle qui ne fait que restreindre les interprétations possibles

d'une représentation, la sémantique opérationnelle, étant plus spécifique, fixe non seulement l'interprétation de la représentation, mais également l'usage qui est fait de cette représentation pour raisonner. Il s'agit à cet effet, de tirer profit de ce qui fait la spécificité du support informatique par rapport au support écrit traditionnel, c'est-à-dire son aspect dynamique. Un système informatique peut en effet, manipuler les connaissances pour en inférer de nouvelles (Bachimont, 1999). Une ontologie est donc opérationnelle lorsqu'elle est dotée d'une sémantique opérationnelle et de mécanismes de raisonnement.

2.2.1.5 Évaluation

L'évaluation d'une ontologie est un processus primordial qui se passe en deux étapes (Darwiche, 2002). D'abord, la **vérification**, qui consiste à s'assurer que l'ontologie est correctement construite du point de vue du modèle de représentation de connaissances adopté. Cette vérification représente donc un processus formel, dépendant du modèle et non du domaine, portant sur des propriétés formelles qui ne peuvent être violées par l'ontologie afin de préserver son expressivité. La deuxième étape est la **validation**, qui consiste à s'assurer que la sémantique exprimée dans l'ontologie doit être celle du domaine considéré.

2.2.1.5.1 Vérification

La vérification d'une ontologie revient à s'assurer des principes de la *conformité*, la cohérence et la minimalité de l'ontologie.

La **conformité** d'une ontologie à un modèle de représentation est respectée lorsque les représentations de connaissances incluses dans l'ontologie sont bien conformes au modèle utilisé. La conformité est indépendante du domaine. Elle se rattache à la forme (syntaxe) de l'ontologie plutôt qu'aux connaissances du domaine. Il s'agit de préciser entre autres, la signature de chaque relation.

La **cohérence** d'une ontologie implique l'absence de contradictions logiques entre les représentations. Elle fait appel à la sémantique formelle du modèle de représentation plutôt qu'à la syntaxe. Par exemple, deux axiomes de l'ontologie syntaxiquement corrects, peuvent être logiquement contradictoires. Encore une fois, les tests de cohérence appliqués sur

l'ontologie ne dépendent pas du domaine de connaissance et repose uniquement sur les représentations contenues dans l'ontologie.

La **minimalité** d'une ontologie revient à s'assurer qu'elle ne contient pas de connaissances superflues; il s'agit entres autres des connaissances redondantes ou celles qu'on peut facilement déduire du reste de l'ontologie ou encore des axiomes non exécutables du fait de la présence de connaissances contradictoires dans leurs hypothèses.

Les tests pratiques de minimalité reposent sur l'utilisation de critères de construction des hiérarchies qui contraignent les choix de modélisation et imposent d'utiliser des propriétés conceptuelles de vérification. C'est le cas par exemple, des principes différentiels énoncés par (Bouaud et al, 1994) qui peuvent être facilement contrôlés formellement de manière à tester la cohérence générale de la hiérarchie des concepts d'une ontologie. Les méta-propriétés proposées par (Welty et al, 2001) permettent également de vérifier la cohérence sémantique de l'ontologie dans la mesure où ces méta-propriétés imposent des contraintes sur les liens de subsomption.

2.2.1.5.2 Validation

La validation d'une ontologie consiste à tester sa fidélité à la sémantique du domaine de connaissances considéré (Gomez-Perez, 1999). Ceci revient à tester, non seulement la *conformité* de l'ontologie par rapport au domaine, mais aussi sa *complétude*.

La **conformité** de l'ontologie par rapport au domaine vise à s'assurer que les connaissances représentées dans l'ontologie correspondent exactement à la sémantique du domaine. En pratique, des tests sont appliqués sur un système implémentant une version opérationnelle de l'ontologie. Ces tests vérifient si les axiomes de l'ontologie permettent de déduire des réponses correctes à des questions qu'on pose au système.

La **complétude** de l'ontologie par rapport au domaine est validée si toutes les connaissances du domaine sont présentes dans l'ontologie ; les tests de complétude concernent tant l'aspect terminologique que l'aspect sémantique. La complétude du niveau terminologique vise à contrôler si toutes les primitives conceptuelles du domaine sont bien présentes dans l'ontologie. L'impossibilité de représenter une question implique une incomplétude du niveau

terminologique de l'ontologie. La complétude du niveau sémantique vise à contrôler si les axiomes de l'ontologie représentent bien toutes les connaissances du domaine. Ce test valide que les axiomes permettent de répondre aux questions.

2.2.2 Outils de développement

Depuis une décennie, de nombreux outils de construction d'ontologies offrant différentes fonctionnalités ont été développés. En plus de la construction d'ontologie, les autres fonctionnalités se rapportent à l'évaluation d'ontologie, l'alignement et la fusion d'ontologie, l'annotation basée sur des ontologies, le raisonnement et la requête basés sur des ontologies et l'extraction d'ontologie à partir de documents. Les outils les plus aboutis intègrent tous plusieurs de ces fonctionnalités. Nous citons à titre d'exemple les outils suivants :

- *DOE* (Differential Ontology Editor) (Bachimont et al, 2002)
- *LinKFactory*® (Ceusters et al, 2001)
- *OILEd* (Bechhofer et al, 2001)
- *OntoEdit* (Sure et al, 2002)
- *Ontolingua* (Farquhar et al, 1997)
- *Ontosaurus* (Swartout et al, 1997)
- *KnoME* (Rogers et al, 2001)
- *Protégé-2000* (Noy et al, 2001) (Noy et al, 2000)
- *SymOntoX* (Missikoff et al, 2002)
- *WebODE* (Arpírez et al, 2001)
- *WebOnto* (Domingue, 1998)

Les outils orientés conceptualisation sont essentiellement dédiés à l'extraction, à partir de documents, des concepts du domaine et des relations existant entre eux, mais offrent également des fonctionnalités de structuration permettant de construire de véritables ontologies. Il s'agit d'outils comme *TERMINAE*, *Text-To-Onto* et *OntoBuilder* ;

- **TERMINAE:** développé au LIPN de l'Université Paris-Nord[11], permet, à travers l'outil d'ingénierie linguistique *LEXTER*, d'extraire les candidats termes d'un domaine à partir d'un corpus textuel (Biébow et al, 1999). Ces concepts doivent ensuite être triés par un expert et organisés hiérarchiquement, puis la sémantique du domaine peut être précisée à travers des axiomes. *TERMINAE* propose des candidats termes à l'utilisateur en se basant sur des techniques d'analyse syntaxique de texte. L'outil présente également les différents sens des concepts donnés par les usages qui en sont faits dans le corpus.

- **Text-To-Onto**: développé à l'institut AIFB de l'Université de Karlsruhe[12], cet outil offre les mêmes fonctionnalités d'extraction d'ontologie à partir de corpus ou de documents Web, mais en utilisant des ontologies existantes (Maedche et al, 2001). *Text-To-Onto* fait partie de la plateforme logicielle *KAON* (« KArlsruhe ONtology and Semantic Web Infrastructure ») d'édition et de maintenance d'ontologie (Bozsak et al, 2002).

- **OntoBuilder**, développé au Technion de Haifa[13], permet de bâtir une ontologie à partir de ressources Web (Gal et al, 2004). L'extraction de l'ontologie est suivie d'une phase de raffinage guidée par l'utilisateur. *OntoBuilder* autorise également la fusion d'ontologies extraites de différents sites Web.

2.3 Approches de maintenance des ontologies

La maintenance des ontologies implique principalement un processus d'« *apprentissage d'ontologie* ». Ce processus met en commun plusieurs activités de recherche traitant certes différents types d'intrants, mais visant un même objectif, celui de la conceptualisation du domaine. Il s'agit en effet, d'un champ multidisciplinaire impliquant le traitement du langage naturel, la prospection de données, l'apprentissage machine et la représentation de connaissances.

La plupart des recherches relatives à l'apprentissage des ontologies n'ont toujours pas atteint un stade de maturité satisfaisant et impliquent inévitablement une intervention plus ou moins

[11] http://www-lipn.univ-paris13.fr/
[12] http://www.aifb.uni-karlsruhe.de/english
[13] http://iew3.technion.ac.il/OntoBuilder/

importante d'un expert du domaine. Nous proposons dans cette section de passer en revue les approches les plus importantes présentées dans la littérature. En se basant sur le type de ressources (intrants) utilisées par l'apprentissage, Alexander Maedche et Steffen Staab (Maedche et al, 2001) proposent la classification suivante **(figure 2.3).** Cette figure illustre pour chaque type de ressource, les approches de maintenance proposées dans la littérature et que nous décrivons dans ce qui suit.

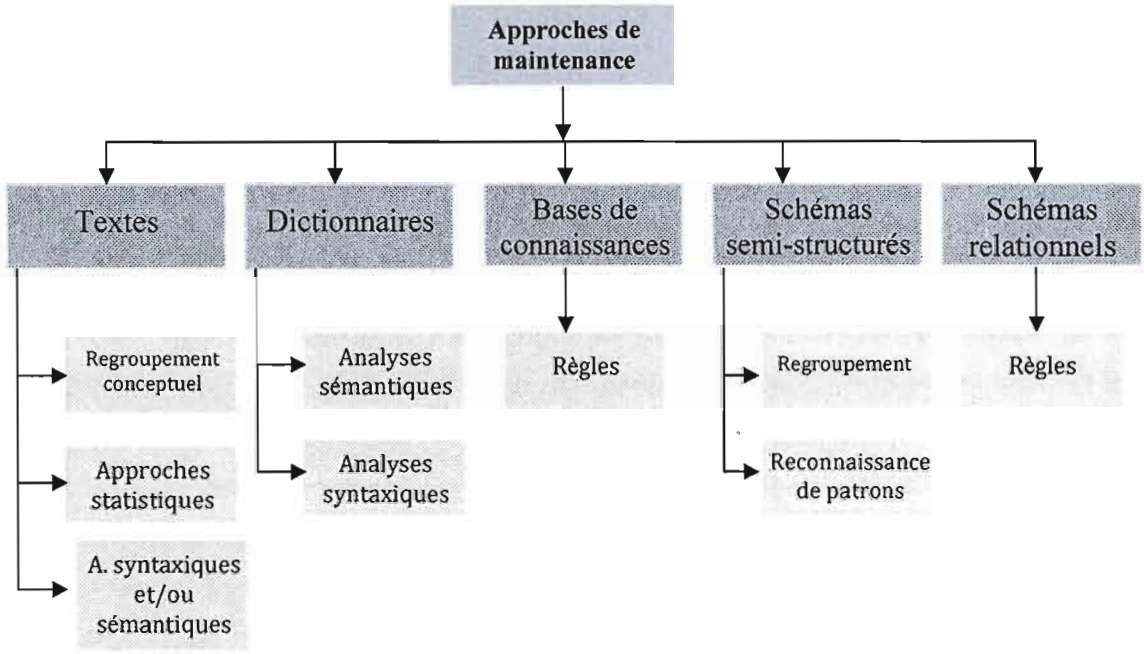


Figure 2.3 : Taxonomie des ressources nécessaires à la maintenance des ontologies et approches correspondantes

2.3.1 Apprentissage d'ontologie à partir de texte

Ce type d'apprentissage a été largement employé chez la communauté de l'ingénierie de connaissances. Il s'agit en particulier des travaux de : (Aguirre et al, 2000 ; Alfonseca E. et al, 2002a, 2002b ; Aussenac-Gilles et al, 2000 ; Bachimont et al, 2002 ; Faatz et al, 2002 ; Gupta et al, 2002 ; Hahn et al, 2001 ; Hearst, 1998 ; Hwang, 1999 ; Khan et al, 2002 ; Kietz

et al, 2000 ; Lonsdale et al, 2002 ; Missikoff et al, 2002 ; Moldovan et al, 2001 ; Nobécourt, 2000 ; Roux et al, 2000 ; Wagner, 2000 ; Xu et al, 2002). Toutefois, aucune méthodologie suffisamment détaillée n'a été présentée pour assister le processus d'apprentissage d'ontologie. En effet, la littérature se limite à la présentation de lignes de conduite plus ou moins générales.

Ces méthodes sont principalement fondées sur des techniques d'analyse du langage naturel. Elles utilisent un corpus à travers différentes étapes du processus. Seuls les travaux de Maedche et ses collaborateurs (Maedche et al, 2001) utilisent tant des corpus généraux que ceux du domaine pour écarter les concepts non spécifiques au domaine de l'ontologie existante. Les autres travaux traitent uniquement des documents relatifs au domaine en vue d'apprendre de nouveaux concepts et de nouvelles relations.

Selon le point de vue technologique, les outils développés dans le cadre de telles approches peuvent être regroupés sous trois catégories principales, dépendamment de la technique d'apprentissage adoptée :

- les outils basés sur le « clustering » conceptuel, tels que ASIUM (Faure et al, 1999), MO'K (Bisson et al, 2000), SVETLAN (Chalandar and al, 2000), TERMINAE (Biébow et al, 1999) ;

- les outils basés sur des approches statistiques, tels que *LTG* (Mikheev et al, 1997), *Text-To-Onto* (Maedche et al, 2001), *TFIDF* (Xu et al, 2002), *WOLFIE* (Thompson et al, 1997), *SubWordNet* (Gupta et al, 2002), *KEA* (Jones et al, 2002) ;

- les outils basés sur des approches linguistiques et/ou sémantiques, tels que *Prométhée* (Morin, 1998, 1999), *Corporum-Ontobuilder* (Engels R, 2001a, 2001b), *TextStorm* (Pereira, 1998), *Welkin* (Alfonseca et al, 2002), *OntoLearn* (Missikoff et al, 2002), *DOE* (Bachimont B., 2000), *SOAT* (Wu et al, 2002).

Aucun de ces outils n'est complètement automatique. Certains sont orientés vers l'assistance à l'acquisition de connaissances lexico-sémantiques, d'autres visent à repérer des concepts ou des relations à partir d'un corpus prétraité, avec l'aide de l'utilisateur, etc. Par ailleurs,

compte tenu de la complexité du processus d'apprentissage, l'évaluation de l'exactitude de ces outils a toujours fait défaut. De plus, aucune comparaison de résultats obtenus en utilisant différentes techniques d'apprentissage n'a été proposée.

2.3.2 Apprentissage d'ontologie à partir de dictionnaire

L'apprentissage d'ontologie a également fait usage, dans certains travaux, de dictionnaires électroniques. La performance de ces méthodes (Hearst, 1992 ; Jannink et al, 1999 ; Rigau, 1998) est basée sur l'utilisation d'analyses sémantiques et linguistiques pour extraire de nouveaux concepts ou des relations à partir de dictionnaires. La plupart de ces méthodes utilisent le dictionnaire WordNet pour enrichir leur ontologie avec de nouveaux concepts ou de nouvelles relations. Les outils, tels que *SEISD* (Rigau, 1998) et *DODDLE* (Yamaguchi, 1999), mettant en application de telles techniques, procèdent principalement à des analyses syntaxiques. Ils nécessitent également l'intervention de l'utilisateur pour valider leurs résultats.

2.3.3 Apprentissage d'ontologie à partir de bases de connaissances

Jusqu'à date, l'apprentissage d'ontologie à partir de bases de connaissances n'est pas suffisamment exploré par la communauté de construction d'ontologie. D'ailleurs, nous n'avons pas retrouvé des outils fondés sur une telle approche. Seuls Suryanto et Compton ont proposé une approche (Suryanto et al, 2001, 2002) visant à générer une ontologie à partir de règles d'une base de connaissances.

2.3.4 Apprentissage d'ontologie à partir de schémas semi-structurés

Des connaissances ontologiques peuvent également être extraites à partir de ressources semi-structurées telles que (XML Schemas, RDF, DAML+OIL, OWL, etc.) en se basant sur des approches de « *regroupement* » ou de « *reconnaissance de formes* ».

(Deitel et al, 2001) ont proposé une approche permettant l'apprentissage d'ontologies à partir des annotations sémantiques RDF d'une base de documents du Web. La méthode consiste à apprendre une ontologie à partir de descriptions de ressources extraites du graphe RDF que constitue l'ensemble des annotations de ces ressources. Une hiérarchie de concepts est

construite en générant systématiquement les généralisations les plus spécifiques de tous les regroupements possibles de ressources. Le résultat de l'apprentissage est évalué par un expert du domaine pour un objectif de validation.

Des outils, tels qu'*OntoBuilder* (Modica et al, 2001) ont été développés pour partir d'une ontologie existante et l'enrichir avec de nouveaux concepts repérés à partir de ressources semi-structurées.

2.3.5 Apprentissage d'ontologie à partir de schémas relationnels

Les schémas relationnels constituent également une source plausible pour extraire des connaissances ontologiques et les inclure d'une façon manuelle dans une ontologie existante. En effet, certains travaux à l'instar de (Stojanovic et al, 2002 ; Kashyap, 1999) et (Rubin et al, 2002) se basent sur l'hypothèse que les connaissances spécifiques à un domaine sont intégrées dans les données et les schémas de bases de données sélectionnées. Ils proposent de construire une ontologie à partir de schémas de bases de données relationnelles en suivant un processus d'appariement. Ce processus s'appuie sur un ensemble de règles pour migrer les éléments du modèle de la base de données (les relations, les attributs, les types d'attributs, les clés primaires, etc.) vers l'ontologie.

Toutefois, nous n'avons pas retrouvé, dans la littérature, des outils appropriés qui permettent de procéder à un tel apprentissage d'ontologies.

2.3.6 Synthèse

L'apprentissage d'ontologies est un processus fondamental au sein des activités de maintenance. Il facilite en effet, l'acquisition de connaissances pour enrichir une ontologie et réduit, par voie de conséquence, le temps nécessaire à cette tâche.

L'utilisation de corpus de textes non structurés offre différents avantages. D'abord, les textes sont généralement facilement accessibles. La prolifération de documents électroniques sur le Web facilite en effet le traitement automatique. Ensuite, les textes non structurés contiennent des connaissances de domaine qui peuvent être pertinentes pour les ontologies de domaine.

Enfin, de plus en plus de méthodes et d'outils permettent la construction semi-automatique d'ontologies à partir de corpus de textes non structurés (Biemann, 2005).

Cependant, il n'existe pas de solution intégrée qui permet de combiner différentes techniques d'apprentissage et des sources de connaissances hétérogènes.

2.4 Extraction de termes à partir de textes

L'extraction de termes à partir de textes pour des fins de construction d'ontologie a été inspirée des recherches en terminologie et spécialement en terminologie spécifique à un domaine. Pour passer d'une terminologie à une ontologie, deux étapes s'imposent. La première consiste à filtrer les candidats termes en éliminant le bruit et en se basant sur un critère préférentiel (tel que la fréquence etc.). La deuxième consiste à transformer l'ensemble de ces termes en une ontologie en structurant cet ensemble selon les liens qui existent entre les termes, mais aussi selon les besoins de l'application utilisant l'ontologie.

Différentes techniques sont aujourd'hui utilisées pour repérer des syntagmes susceptibles d'être des termes. Elles sont opérationnelles dans des logiciels comme *SATIM* (Biskri et al, 2002), *NOMINO*[14] (David et al, 1990), *LEXTER* (Bourigault, 1996), etc. Ces approches peuvent être regroupées en trois grandes catégories possibles : les approches structurelles basées sur l'utilisation de grammaires formelles ; les approches non structurelles, telles que les approches statistiques et quantitatives, plutôt utilisées pour extraire des termes complexes et de plus en plus utilisées grâce à la disponibilité de gros corpus en format électronique et les approches mixtes associant analyses statistiques et méthodes structurelles.

Les approches structurelles d'extraction de termes sont essentiellement des « *méthodes utilisant une grammaire* » ou encore des « *méthodes de surface* ».

- **Les méthodes utilisant une grammaire** : les approches structurelles, utilisées pour les systèmes de traitement automatique de la langue, requièrent souvent des grammaires (par exemple des grammaires probabilistes, par règle, etc.) (Brill, 1993) et parfois des lexiques ou des dictionnaires électroniques de la langue utilisée. Ce type d'approches

[14] Nomino est appelé à l'origine Termino

visent à dissocier les traitements à chaque niveau d'analyse et servent principalement de connaissances linguistiques pour chaque étape.

L'approche classique utilisée pour l'analyse de textes consiste en une analyse morpho-lexicale suivie d'une analyse syntaxique (Sabah 1989). La première produit à partir de ressources lexicales exhaustives, une liste d'unités lexicales (des mots arbitraires). La seconde produit, pour chaque phrase, un ou plusieurs arbres syntaxiques.

En raison de ses caractères sémiotique et linguistique, le traitement classique de l'information est habituellement linguistique. En effet, un texte est considéré comme étant une succession de phrases qui doivent faire l'objet d'analyseurs linguistiques. Cette approche semble être complètement naturelle dans la mesure où elle correspond, en théorie, au processus normal de lecture chez l'humain (Meunier, 1996). Cependant, un problème délicat concerne la théorie des textes. Les textes, sont-ils des phénomènes linguistiques ? La réponse dépend de la définition et de la compréhension du concept « linguistique ». Si ce concept est strictement considéré en tant que « grammaire », alors un texte n'est pas un phénomène grammatical. Bien que certains auteurs le pensent (Pavel, 1975 ; Dijk, 1977), d'autres, tels que (Rastier, 1994 ; Meunier, 1996) refusent une telle vision.

De point de vue technique, la difficulté majeure de ces méthodes est leur aspect combinatoire. En effet, chaque unité lexicale se voit affecter une étiquette et l'ensemble de ces étiquettes permettent d'attribuer à la phrase une structure syntaxique attendue, souvent codée sous forme d'une grammaire. De plus, les systèmes utilisant ces méthodes ne garantissent pas une information exacte pour des mots ou des séquences de mots inconnus (à moins de prévoir un traitement des exceptions). Par ailleurs, dans les approches basées sur des grammaires, on ne reconnaît que les occurrences se trouvant sous une des formes explicitement attendues.

- **Les méthodes de surface** : d'autres approches structurelles reposent sur des méthodes dites « *de surface* ». Ces approches se caractérisent notamment par l'utilisation de patrons syntaxiques de reconnaissance. À titre d'exemple, (Bourigault, 1994) utilise des

marqueurs[15] de frontières complétés par un étiquetage grammatical des mots du corpus afin d'acquérir des syntagmes susceptibles d'être des termes. Son outil d'extraction terminologique *LEXTER* (Bourigault, 1996) utilise des bornes de syntagmes nominaux qui sont des mots appartenant, pour la plupart, aux catégories grammaticales suivantes : verbes, pronoms, déterminants et adverbes.

Les termes complexes sont caractérisés par une compositionnalité limitée[16] (Manning et al, 1999). Les termes complexes (par exemple, « *pomme de terre* ») ne sont pas complètement compositionnels dans la mesure où un élément de sens est souvent ajouté à la combinaison. La plupart des termes techniques sont des mots composés, et un outil de traitement de corpus qui détruit des liens organiques entre composants présente des inconvénients majeurs. Par exemple, pour éviter de séparer « pneu d'hiver » en trois unités, il faudrait que le système sache détecter soit les bornes du composé, soit un lien interne entre ses constituants.

L'identification complète et précise des termes à partir d'un corpus spécifique est considérée comme un prétraitement d'une grande importance tant pour la construction d'ontologie que pour la recherche documentaire, l'indexation, la traduction, le résumé automatique et la terminologie. Ainsi, un certain nombre d'outils ont été développés dans l'objectif d'extraire des termes complexes. Ces outils se basent sur des calculs statistiques ou d'analyse linguistique.

Le problème majeur relatif au repérage de collocations d'une façon générale, est principalement attribué à leur apparition sous différentes formes. Les collocations varient considérablement selon le nombre de mots qu'elles renferment, les catégories syntaxiques des mots, les relations syntaxiques entre les mots et la rigidité de leur structure. Par exemple, dans certains cas, les mots d'une collocation doivent être adjacents, alors que dans d'autres, ils peuvent être séparés par un nombre variable d'autres mots. Ainsi, parmi les types de collocations, que nous pouvons trouver, nous citons :

---

15 Un marqueur est une formule, opérationnelle ou non, d'éléments linguistiques qui est rattachée à une relation lexicale.

16 Une expression du langage naturel est dite compositionnelle si le sens de l'expression peut être prédit à partir des sens des parties.

- ***Les termes complexes***, impliquant des séquences ininterrompues de mots telles que « marché de change »,
- ***Les relations prédicatives*** qui consistent en deux ou plusieurs mots, utilisés répétitivement ensemble selon des relations syntaxiques similaires. Ces relations lexicales sont plus difficiles à identifier dans la mesure où elles correspondent à des séquences de mots interrompus dans le corpus.
- ***Les patrons de phrases*** qui consistent en des expressions rigides, longues et souvent représentatives d'un domaine donné.

Les connaissances collocationnelles portent principalement sur les ensembles de mots qui cooccurrent ensemble et la façon syntaxique avec laquelle ces mots sont combinés. Ces affinités ne peuvent pas être formulées sur la base de règles sémantiques ou syntaxiques, mais peuvent être observées sous forme de régularités dans les textes (Cruse, 1986).

Les approches de repérage de collocations basées sur de larges corpus connaissent une évolution de plus en plus importante (Nagao et al, 1994 ; Ikehara et al, 1996 ; Kupiec, 1993 ; Fung, 1995 ; Kitamura et al, 1996 ; Smadja, 1993 ; Smadja et al, 1996 ; Haruno et al, 1996 ; Remaki et al, 2000). Bien que ces approches accomplissent des résultats intéressants, la plupart visent à extraire des collocations « rigides », spécialement des groupes nominaux, et nécessitent de l'information qui est dépendante de la langue telles que les dictionnaires.

Nous décrivons dans ce qui suit les approches linguistiques et statistiques d'extraction de termes complexes.

2.4.1 Approches linguistiques pour l'extraction de termes complexes

Bien que différentes études aient été proposées (Church, 1988 ; Zernik et al, 1992 ; Calzolari et al, 1990 ; Garside et al, 1985 ; Hindle et al, 1991 ; Brown et al, 1992), ces auteurs s'entendent sur le fait que le corpus doit être préalablement analysé et étiqueté. Ces tâches d'analyse et d'étiquetage sont assistées par des humains, ou encore, nécessitent une correspondance entre des corpus de langages différents. En raison de cette assistance manuelle, ces algorithmes sont devenus relativement contraignants dans les applications actuelles.

Nous décrivons dans cette section, quatre types de solutions proposées pour la détection de termes complexes, à savoir : le repérage de relations lexicales par le calcul des cooccurrences, le repérage de structures syntaxiques internes, le repérage des bornes du syntagme et le repérage de séquences lexicales répétées.

2.4.1.1 Repérage de relations lexicales par le calcul des cooccurrences

Il est d'usage en linguistique de classifier les mots, non seulement sur la base de leurs significations, mais aussi sur la base de leur cooccurrence avec d'autres mots (Church, 1989). Il s'agit en quelque sorte de caractériser un mot donné par son contexte.

> « *You shall know a word by the company it keeps* » (Firth, 1957).

Il existe des mots dont l'apparition est plus probable dans tel contexte que dans tel autre, avec lequel ils entretiennent des relations privilégiées. Par exemple si « *réseau* » apparaît fréquemment dans le contexte de « *télécommunication* », alors ces deux mots sont probablement liés de quelque manière.

2.4.1.2 Repérage de structures syntaxiques internes

Le repérage de la structure interne d'un terme complexe consiste à associer à chaque mot du corpus une étiquette représentant sa catégorie grammaticale et à apparier des séquences lexicales dans un corpus avec des schémas syntaxiques préétablis, tels que : NOM NOM (*lutte antipollution*), NOM PRÉPOSITION NOM (*tableau de bord*), NOM ADJECTIF (*champ magnétique*), etc. L'assignation de catégories (ou étiquetage) se déroule en deux phases ; la reconnaissance des mots du texte et l'assignation de la catégorie.

La reconnaissance se fait à l'aide d'un dictionnaire et/ou d'un analyseur morphologique. Le choix de l'assignation se fait selon des règles préétablies (Brill, 1994), ou selon des procédures stochastiques (Schütze et al, 1994 ; Brill, 1994). Ce sont ces dernières qui semblent les plus performantes. Elles se basent sur un corpus d'apprentissage, étiqueté d'une façon manuelle, pour inférer des règles à partir de mesures de cooccurrences des catégories. Par exemple, avant un nom, il y a telle probabilité pour que l'on trouve un adjectif, ou un déterminant, ou un adverbe, etc. Ces règles sont ensuite appliquées à un corpus plus volumineux et le résultat est corrigé à la main. Le système est ensuite en mesure de traiter le

reste du corpus. Le taux d'erreur de ces systèmes est de l'ordre de 3 à 5% selon les méthodes, ce qui semble peu, mais qui n'est pas suffisant pour assurer une analyse syntaxique correcte.

L'analyse syntaxique consiste à analyser les phrases en leurs constituants (par exemple : groupe nominal sujet, groupe nominal objet, verbe) en se basant sur les étiquettes grammaticales et des règles de bonne formation syntaxique. Les performances sont actuellement assez modestes. En effet, le taux de réussite, c'est-à-dire le nombre de phrases correctement analysées, est assez faible. Selon l'étude de (Pereira et al, 1992) élaborée sur des analyseurs probabilistes, si les règles inférées par le système à partir de calculs sur un corpus d'apprentissage ne sont pas vérifiées manuellement, alors le taux de réussite du système sur un corpus de travail ne dépasse pas 35%. Selon les mêmes auteurs, il ne serait que de 78% sous contrôle humain. (Briscoe et al, 1993) obtiennent 75% de réussite à l'aide d'un contrôle semi-automatique.

2.4.1.3 Repérage des bornes du syntagme

Le repérage des bornes d'un syntagme nominal peut se faire par la détection du premier mot à gauche et à droite qui ne peut pas faire partie du syntagme. Par exemple, il peut être borné à gauche par la phrase précédente et à droite par un auxiliaire (est). *LEXTER*, conçu pour l'extraction de terminologie (Bourigault, 1994 ; Bourigault et al, 1994) est un outil fondé sur ce principe.

Cette méthode présente deux avantages qui la rendent beaucoup plus performante que toutes les autres méthodes, qu'elles soient statistiques ou syntaxiques. D'abord, l'assignation de catégories grammaticales n'a besoin d'être performante que pour quelques catégories constituées de listes relativement fermées (déterminants, prépositions, conjonctions, etc.), ainsi que les verbes et les adverbes. Ensuite, elle collecte la quasi-totalité des syntagmes nominaux du corpus, sans avoir à spécifier leurs schémas syntaxiques. Elle présente toutefois, le même inconvénient que l'approche par repérage de structures syntaxiques internes, à savoir la difficulté à distinguer les syntagmes figés des syntagmes discursifs.

2.4.1.4 Repérage de séquences lexicales répétées

Ces méthodes présument que le problème principal du repérage de termes complexes est leur délimitation. D'un côté, les méthodes purement statistiques n'y parviennent pas, étant donné qu'elles se contentent de lister les cooccurrents d'un mot donné, dont la fréquence dépasse un certain seuil dans une fenêtre arbitraire, sans indication de lien syntaxique. De l'autre côté, les méthodes syntaxiques doivent spécifier les relations syntaxiques entre les composants. Lebart et Salem (Lebart et al, 1988) ont proposé une méthode fondée sur une idée relativement simple : si une séquence de mots est spécifique à un domaine, alors il y a de fortes chances qu'elle soit répétée dans un corpus. Le critère de répétition est un critère objectif qui permet le bornage automatique de ces séquences lexicales répétées. L'inconvénient de cette méthode est qu'elle génère beaucoup de données redondantes ou sans intérêt.

2.4.2 Approches statistiques pour l'extraction de termes complexes

Les méthodes statistiques sont souvent réalisées sur de gros corpus (étiquetés ou pas). Elles se caractérisent par l'utilisation de la notion de seuil. Cette notion est utilisée pour filtrer ou repérer les informations contenues dans le corpus, ce qui explique en fait la possibilité de perte d'information. Plusieurs méthodes statistiques peuvent être appliquées, parmi lesquelles on trouve celles utilisant la notion d'«*information mutuelle* » ou la notion de « *segments répétés* ».

Les méthodes utilisant la notion d'**information mutuelle** consistent à détecter des associations récurrentes de mots, par exemple des paires de mots, ayant une forte valeur d'association mutuelle dans une fenêtre de n mots. Parmi ces associations, on trouve entre autres, les termes composés d'un texte dont la fréquence est assez élevée pour qu'ils soient repérables. Des associations sémantiques peuvent également être repérées, en augmentant la distance prise en compte. On parle alors de notion d'information mutuelle. Dans (Brown et al, 1992), l'observation par exemple de paires à information mutuelle forte dans une fenêtre de 2 à 5 mots montre qu'on peut avoir des termes composés, alors que la recherche dans une fenêtre supérieure à 5 mots ne permet d'avoir que des mots à forte affinité sémantique. (Smadja, 1993) a, de son côté, proposé un outil, appelé Xtract, de traitement de la distance

pouvant séparer des paires de mots fortement associés en les recherchant dans une fenêtre de 5 mots. Cet outil permet de repérer des collocations de structures telles que : N2 N1, N1 of N2, ADJ-N, etc.

La spécificité de telles méthodes est qu'elles utilisent des scores pour mesurer le poids d'association de deux mots. Ainsi, dans le cadre de l'extraction de ressources lexicales monolingues, (Church et al,1990) utilisent un score d'association fondé sur la notion d'information mutuelle. Il s'agit d'un score d'association de deux lemmes qui permet de comparer la probabilité d'observer ces deux lemmes ensemble avec la probabilité de les observer séparément.

D'autres méthodes statistiques ont été proposées en utilisant **la notion de segments répétés**. Dans l'objectif de ne pas se limiter à l'extraction des récurrences de mots associés en paires, comme celles obtenues par l'information mutuelle, Lebart et Salem (Lebart et al, 1994) ont proposé d'étendre l'extraction à des suites de plus de deux mots et répétées dans le corpus. On parle alors d'extraction de segments répétés de textes. Cette méthode privilégie le voisinage des chaînes et met en évidence l'importance du contexte dans l'apparition d'une séquence de mots. Les données obtenues par cette méthode sont linguistiquement hétérogènes (elles contiennent par exemple des syntagmes nominaux, des syntagmes verbaux, des formes figées, etc..). Mais on trouve aussi, des morceaux de syntagmes nominaux plus au moins figés, ou simplement des fragments de texte. Ces données doivent donc être filtrées, voire retraitées afin d'obtenir des objets linguistiques homogènes. Ceci justifie donc l'utilisation de méthodes mixtes.

2.4.3 Méthodes mixtes

Les approches mixtes tentent de tirer profil, tant des méthodes statistiques qui sont multilingues et pouvant traiter de grands volumes de données textuelles, que des approches linguistiques qui sont capables de rendre compte de certains néologismes dans des domaines spécifiques.

(Justeson et al, 1995) ont proposé un prototype, appelé *TERMS*. Leur idée consiste à utiliser des contraintes syntaxiques sous forme de schémas syntaxiques pour repérer des syntagmes à

partir de textes étiquetés. Le critère de répétition est utilisé pour ne retenir que des schémas de composés, répétés et de longueur deux et trois.

2.4.4 Conclusion

La difficulté majeure des méthodes basées sur des grammaires est leur aspect combinatoire. En effet, on affecte une étiquette à chaque unité lexicale, et l'ensemble des étiquettes affectées permet d'attribuer à la phrase une structure syntaxique attendue, souvent codée sous forme d'une grammaire. Un tel processus est relativement lourd quand il est question d'un grand corpus, ce qui est souvent le cas en TALN.

L'information obtenue par ces méthodes d'analyse est exacte sauf pour le cas de mots ou de séquences de mots inconnus, où le système ne permet pas de fournir une information exacte (à moins de prévoir un traitement des exceptions). D'ailleurs, le taux d'erreur dans l'assignation grammaticale est estimé à 5%. En d'autres termes, la probabilité pour qu'une phrase de 20 mots contienne un mot incorrectement analysé est très élevée, et moitié moins élevée seulement pour une phrase de 10 mots, qui est une longueur moyenne.

De plus, certains auteurs s'interrogent sur la validité discursive des règles syntaxiques. Des règles établies en fonction de phrases idéalement formées ne décrivent-elles pas que des artefacts dont les occurrences réelles sont finalement assez rares dans les corpus ? Une grammaire d'unification DCG (« Definite Clause Grammar ») fondée sur plusieurs années d'intuition et de méthodologie linguistique, essentiellement d'inspiration génératiViste, se révéla incapable d'analyser correctement de vraies phrases choisies au hasard dans de vrais journaux.

Enfin, les textes utilisés pour ce type d'analyse peuvent provenir de différentes sources électroniques et donc une des difficultés de l'analyse est la prise en compte de la présence de fautes d'orthographe, d'erreurs grammaticales ou encore d'erreurs de reconnaissance de caractères.

Bien que les approches statistiques accomplissent des résultats intéressants, elles présentent également quelques inconvénients. La plupart visent à extraire des collocations « rigides »,

spécialement des groupes nominaux, et nécessitent de l'information qui est dépendante de la langue tels que les dictionnaires.

Il se pose également la question de la taille du corpus et du seuil de fréquence au-delà duquel on ne prend plus en compte les cooccurrents. Ces méthodes sont souvent réalisées sur de gros corpus. (Grefensette, 1994) montre qu'une taille minimale de corpus est à respecter et qu'il faut choisir la méthode statistique en fonction de la taille du corpus. Des corpus comme le « Brown Corpus » (1 million de mots) ne sont pas considérés comme suffisamment gros pour effectuer par exemple des tâches d'étiquetage (Church et al, 1991) et d'autres corpus même de 150 millions de mots ont été jugés inadéquats pour certaines approches statistiques appliquées dans un but de résolution d'ambiguïté (Dagan et al, 1991).

La seconde contrainte est relative aux seuils de fréquence. Si ces derniers sont bas, la liste risque de présenter des cooccurrents sans valeur générale, car liés à quelques spécificités du corpus. Si les seuils sont élevés, on risque alors de ne conserver que les cooccurrents les plus typiques, et donc de perdre de l'information. Il faut par conséquent augmenter ou abaisser les seuils en fonction des objectifs de l'application ; mais est-on sûr que les seuils soient les mêmes quels que soient les mots à étudier ? Il n'y a donc probablement pas de seuil unique qui soit valable pour tous les mots d'un corpus.

C'est pour pallier ce genre de difficultés que beaucoup d'auteurs couplent les résultats statistiques avec des analyses syntaxiques, à l'instar de (Daille, 1994 ; Smadja, 1993). Une technique statistique ne peut de toute façon, se passer de l'utilisation de la linguistique, du moins pour procéder à la lemmatisation et au filtrage du lexique.

Enfin, il n'existe pas de modèle morphosyntaxique ou statistique capable de décider qu'un syntagme est un terme. Pour plusieurs outils, la validation par élagage de la liste des candidats termes extraits doit être réalisée.

2.5 Extraction de relations à partir de textes

Différents travaux ont envisagé la découverte de relations entre des termes dans des corpus. Il s'agit donc de présenter et d'évaluer différentes méthodes dont les résultats permettent de relever des informations sémantiques explicitées dans des textes.

Les relations sémantiques entre termes sont essentiellement du type *généralisation-spécialisation*. Toutefois, d'autres relations peuvent faire associer des termes, telles que la *composition*, la *dépendance* et la *disjonction*. De telles relations véhiculent ainsi une sémantique plus riche pour décrire un domaine.

Pour des systèmes disposant d'un outil d'extraction de termes, l'acquisition de relations sémantiques entre termes se situe en aval de leur acquisition. Cependant, ces systèmes d'extraction de terminologie ne se sont pas vraiment intéressés au problème de l'acquisition de relations sémantiques entre termes. En effet, l'analyse sémantique est encore une discipline complexe et il est difficile de pouvoir modéliser les mécanismes linguistiques et cognitifs auxquels elle fait appel. Il existe tout de même quelques systèmes qui traitent de ce problème qu'on peut classer en deux catégories :

- les systèmes qui effectuent le repérage de relations sémantiques à partir de règles préétablies et qui utilisent une liste de marqueurs morphosyntaxiques (verbes, prépositions, etc...) ;
- les systèmes qui, au contraire, effectuent le repérage de ces marqueurs morphosyntaxiques directement à partir de textes en utilisant des algorithmes de repérage de séquences de mots et en exploitant la liste des termes du domaine.

Un autre principe consiste à fouiller par des méthodes statistiques la distribution de classes de mots en corpus afin de proposer des relations entre ces mots, sans se soucier de la nature sémantique caractérisant ces relations.

Nous présentons brièvement dans ce qui suit ces trois types d'approches.

2.5.1 Analyse par règles pour l'extraction de connaissances

Cette approche consiste à utiliser une analyse du texte basée sur des règles décrivant des schémas de relations à rechercher dans le corpus. L'objectif est de repérer les contextes de ces relations dont les contraintes morphosyntaxiques sont décrites dans des règles déclaratives du type : (SI conditions ALORS conclusion). Cette démarche a été appliquée au système SEEK (Jouis, 1995) en se basant sur la grammaire applicative et cognitive de (Desclés, 1990).

En appliquant des règles morphosyntaxiques préétablies, il est possible de repérer des relations à partir de textes connaissant des indices et des marqueurs linguistiques. Cependant, (Jouis et al, 1997) ont montré que l'analyse d'un nouveau domaine, utilisant ces mêmes règles prédéfinies, peut se heurter à des difficultés, dans la mesure où l'expression des relations peut être décrite par d'autres indices ou d'autres types de marqueurs. En effet, pour décrire les concepts d'un domaine, la langue utilise des moyens d'expression très variés et très riches tels que la synonymie, la métaphore, la paraphrase, l'introduction de néologismes, etc. Ces moyens d'expression peuvent contenir des ambiguïtés et leur évolution incessante ne permet pas de fixer des règles préétablies pour les décrire.

2.5.2 Extraction d'information en utilisant des patrons

Parmi ces systèmes, on peut citer PALKA : «Parallel Automatic Linguistic knowledge Acquisition» (Kim et al, 1995), dont le but est de faciliter la construction d'une base de connaissances de schémas syntactico-sémantiques, relatifs à des structures de relations entre termes. La base de connaissances est ainsi organisée sous forme d'une hiérarchie de concepts et de structures relatives à des patrons. Le cadre (« *frame* ») d'une structure est représenté par un nœud, un ensemble de propriétés et des contraintes sémantiques.

Pour extraire ces structures de schémas, il faut d'abord remplir manuellement des structures prédéfinies appelées *patrons* («*templates*») à partir du texte. Pour construire les structures, le système effectue à partir des patrons, des mises en correspondance entre les phrases et les *cadres* existants.

2.5.3 Approches statistiques

La base théorique de ces approches repose sur l'hypothèse formulée par Zellig Harris (Harris, 1968). On peut classer les divers sens d'un terme en fonction des constructions auxquelles il participe. Des termes qui ont des distributions comparables ont souvent un élément de sens commun. Partant de cette hypothèse, une première série de travaux a étudié la distribution lexicale en corpus afin de proposer des hypothèses de relations entre ces mots. Pour l'anglais, certains travaux comme ceux de (Smadja, 1993) ont étudié les fréquences de cooccurrences de mots pour proposer des relations entre ces mots. Pour le français, certains auteurs, comme (Toussaint et al, 1997), ont étudié ces phénomènes de cooccurrences afin de former des regroupements de termes. Présentés à l'expert, ces regroupements permettent le repérage de relations de synonymie, d'hyponymie[17] ou encore de méronymie[18].

Bien que les approches statistiques soient suivies d'une interprétation humaine systématique, elles sont reconnues comme des méthodes robustes et ne sollicitant pas des connaissances préalables sur le domaine. Nous pensons qu'elles sont très pertinentes pour distinguer des classes d'usage de termes dans l'espoir de les organiser en systèmes structurés reflétant une organisation conceptuelle. Parmi les outils utilisant de telles approches, nous citons : ***TFIDF*** (Xu et al, 2002), ***SVETLAN*** (Chalandar et Grau, 2000), ***Text-To-Onto*** (Maedche et Volz, 2001), ***WOLFIE*** (Thompson et Mooney, 1997), ***TERMINAE*** (Biébow et Szulman, 1999).

2.5.3.1 Cooccurrence

La notion de cooccurrence de termes dans un corpus a été mise à profit dans cette thèse pour appliquer la classification de documents ainsi que l'Indexation Sémantique Latente. Cette notion fait référence au phénomène général par lequel, des termes sont susceptibles d'être utilisés dans un même contexte (Manning, et al, 1999). Autrement dit, on considère qu'il y a cooccurrence lorsque la présence d'un mot dans un texte donne une indication sur la présence d'un autre mot.

---

[17] L'hyponyme est le nom de l'élément d'un tout dont le sens est inclus dans le sens du nom du tout (exemple : *Chien, chat, âne...* sont des hyponymes de *animal)*

[18] Méronymie : Relation hiérarchique existant entre deux concepts ou deux signes linguistiques, dans laquelle le premier est une partie d'un tout que constitue le second.

La proximité sémantique entre les termes repose principalement sur la cooccurrence d'un ensemble de termes à travers différents segments d'un corpus. La base théorique de cette hypothèse repose sur celle formulée par Harris (Harris, 1968) ; on peut classer les divers sens d'un terme en fonction des constructions auxquelles ce dernier participe. Des termes qui ont des distributions comparables ont souvent un élément de sens commun.

Selon un point de vue sémantique, le phénomène de cooccurrence ne dit rien du type de relations entre les cooccurrents. Par exemple, *avion* et *aéroport*, deux termes qui, selon toute vraisemblance, sont utilisés la plupart du temps dans un contexte commun, celui de l'aviation. Bien qu'il soit évident que ces deux termes partagent quelque chose en commun, ils ne sont reliés selon aucune relation sémantique (telle que la synonymie, l'hyperonymie, l'hyponymie, l'holonymie, la meronymie, ou l'antonymie). Ces deux termes sont tout simplement reliés selon une relation d'association.

Par définition, le calcul de cooccurrences suppose la définition de zones dans lesquelles les termes cooccurrent ensemble. Curieusement, chaque réalisation ne fait jamais intervenir qu'un seul type de contexte : la phrase (délimitée par une ponctuation forte), une fenêtre de n mots, le texte... Aussi, quand il n'y a pas d'usage fixé, soit on argumente pour démontrer que, parmi toutes les définitions de contexte que l'on pourrait envisager, l'une est plus pertinente que les autres ; soit on considère que la définition du contexte peut varier suivant les types de textes et les applications visées, et que c'est un paramètre à ajuster, souvent sur des considérations heuristiques (tel choix « marche mieux » que tel autre dans tel cas de figure).

Plusieurs mesures de cooccurrences permettent d'évaluer la relation sémantique entre termes. La mesure la plus simple serait d'utiliser directement les fréquences, en ne retenant que les cooccurrences d'une fréquence minimale donnée. Cependant, une telle mesure peut difficilement s'appliquer à l'ensemble des termes d'une langue, car les fréquences varient énormément d'un terme à l'autre. Plusieurs autres mesures d'association ont toutefois été proposées pour quantifier la force d'une combinaison de termes, dont le *rapport de vraisemblance* (« *log-likelyhood ratio* »), l'*information mutuelle*, le *test t* et le *test du ki-carré*.

L'information mutuelle et le test t ont tendance à surestimer la force des combinaisons de faible fréquence ou dont un des composants est rare (Dunning, 1993). En revanche, le rapport de vraisemblance[19] s'appuie sur des fondements statistiques solides pour comparer directement l'importance d'évènements rares et d'évènements fréquents. Il s'agit par ailleurs de la mesure la plus apte à isoler les collocations d'un ensemble de combinaisons (Orliac, 2004). Pour ces raisons, nous avons choisi le rapport de vraisemblance comme mesure de la force de nos cooccurrences.

Considérant les deux hypothèses présentées ci-après, le rapport de vraisemblance indique laquelle de ces deux hypothèses est la plus probable. L'hypothèse 1 reflète le cas où il y aurait indépendance complète entre les deux termes : les probabilités conditionnelles qu'un terme 2 soit présent connaissant respectivement la présence ou l'absence du terme 1 sont les mêmes. Par contre, l'hypothèse 2 voulant que les deux termes soient dépendants implique que ces probabilités conditionnelles soient différentes.

Hypothèse 1 : P(terme2 | terme1 présent) = p = P(terme2 | terme1 absent) (indépendance)
Hypothèse 2 : P(terme2| terme1 présent)=p1 ≠ p2=P(terme2 | terme1 absent) (dépendance)

S'en tenir exclusivement à des mesures de cooccurrences entre des couples de termes à travers un corpus s'avère une méthode moins performante que les techniques de classification et de l'Indexation Sémantique Latente (que nous décrivons dans le chapitre 3). En effet, deux termes peuvent avoir une relation sémantique, non pas parce qu'ils cooccurrent ensemble à travers le corpus, mais aussi et surtout, parce qu'ils cooccurrent avec d'autres termes à travers le corpus.

2.5.3.2 Analyse Formelle de Concepts

L'Analyse Formelle de Concepts (AFC) représente une approche théorique de regroupement conceptuel des données permettant d'identifier des concepts et de dégager des liens et patrons sous forme de règles d'association. Cette approche a été initiée par Wille au début des années

---

[19] Le rapport de vraisemblance $\boxed{\Lambda(\mathbf{r}) = \frac{p(\mathbf{r}|H_1)}{p(\mathbf{r}|H_0)}}$ compare directement la vraisemblance des observations sous chacune des hypothèses. Il peut être interprété comme une valeur reflétant dans quelle proportion une hypothèse est plus probable qu'une autre

80 (Wille, 1982) et développée par la suite par Ganter et ses collaborateurs (Ganter et al, 1999). L'AFC est de plus en plus utilisée pour des applications dans plusieurs disciplines telles que la linguistique, le développement de logiciels, la psychologie, l'intelligence artificielle et la recherche d'information. Les publications sont abondantes spécialement dans le domaine du développement de logiciels (Godin et al, 1993 ; Mili et al, 1997).

L'AFC est basée sur une interprétation philosophique considérant un concept comme constitué de deux composants : son *extension* formée par tous les objets appartenant à ce concept et son *intension* qui inclut des attributs communs à tous ces objets.

Les objets et les attributs sont importants lorsqu'on parle de relation hiérarchique (sous-concept/super-concept) entre concepts ou la relation d'implication entre attributs ou encore la relation de conséquence (un objet possède un attribut). La définition d'un jeu d'objets et d'attributs et de relations binaires forme ce qu'on appelle un *contexte formel*. L'AFC se sert également de *concepts formels*, que l'on peut définir à partir de l'exemple suivant (Wolff, 1994) :

ANIMAUX	prédateur	volant	oiseau	mammifère
LION	X			X
PINSON		X	X	
AIGLE	X	X	X	
LIÈVRE				X
AUTRUCHE			X	

Tableau 2.1 : Un exemple de contexte formel de « ANIMALS »

Le tableau (**Tableau 2.1**) décrit pour ces animaux, quels attributs ils possèdent parmi ceux mentionnés. La présence d'un attribut pour un objet est spécifiée dans le tableau par un « X ». La structure mathématique utilisée pour décrire formellement des tableaux de X est appelée *contexte formel*.

Pour chercher un exemple de concepts formels à partir de ce contexte, nous repérons les attributs de « *PINSON* », à savoir « *volant* » et « *oiseau* » et nous identifions les animaux (de ce contexte) ayant les mêmes attributs que ceux de « *PINSON* ». Ainsi, l'ensemble A constitué des objets « *PINSON* » et « *AIGLE* » est étroitement lié à l'ensemble B formé de tous les attributs valides pour tous les objets de A. Chaque paire (A, B) est appelée *concept formel* de ce contexte. L'ensemble A est l'extension et B est l'intention du concept (A, B).

Si on considère l'ensemble de tous les concepts d'un contexte, un ordre hiérarchique (sous-concept à super-concept) peut être défini, tel que le concept « *oiseaux volants prédateurs* » qui décrit un sous-concept de « *oiseaux volants* ». L'extension de ce sous-concept est seulement « *AIGLE* » et l'intention est constituée des attributs « *prédateur* », « *volant* » et « *oiseau* ». L'extension de ce super-concept est formée de « *PINSON* » et « *AIGLE* » et l'intention est composée de « *volant* » et « *oiseaux* ». En général, un concept C est considéré comme un sous-concept d'un concept D, si l'extension de C est un sous-ensemble de l'extension de D, ou encore si l'intention de C inclut l'intention de D.

À l'instar de l'exemple de « *PINSON* », il est possible de construire pour chaque objet g, son *concept objet* (A, B), tel que B est l'ensemble de tous les attributs de g et A est l'ensemble de tous les objets possédant les mêmes attributs que ceux de B. De la même façon, chaque attribut m détermine son *concept attribut* (C, D), tel que C est l'ensemble de tous les objets de m et D est l'ensemble de tous les attributs valides pour tous les objets de C. Le diagramme suivant (**Figure 2.4**) illustre la hiérarchie conceptuelle de tous les concepts du contexte « *ANIMAUX* ».

Ce diagramme est constitué de nœuds, lignes et noms de tous les objets et les attributs du présent contexte. Les nœuds représentent les concepts et l'information décrivant ce contexte peut être exprimée à partir des lignes du diagramme en suivant la règle suivante :

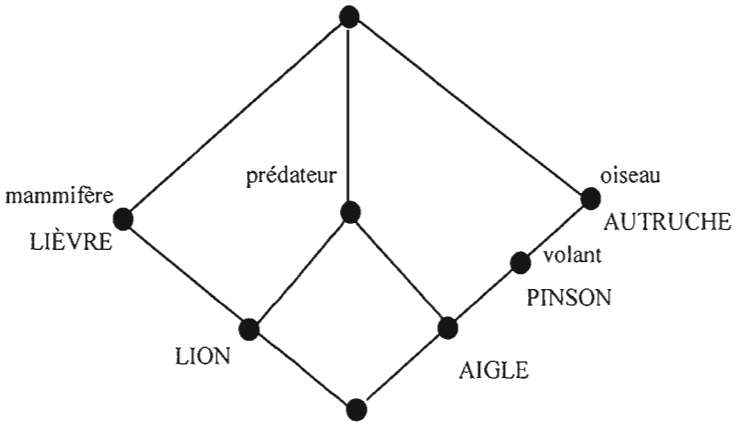


Figure 2.4 : Diagramme du contexte « ANIMAUX »

Un objet g possède un attribut m si et seulement s'il existe un chemin ascendant allant du nœud g au nœud m. Par exemple, « *PINSON* » possède les attributs « *volant* » et « *oiseau* ». L'extension et l'intention peuvent être déduites en regroupant, respectivement, tous les objets au dessous et tous les attributs au dessus du nœud du concept en question. Ainsi, le concept objet « *PINSON* » possède « *PINSON* » et « *AIGLE* » comme l'extension et « *volant* » et « *oiseau* » comme intention. L'extension du nœud supérieur est souvent constituée de l'ensemble de tous les objets. Alors que son intention ne possède aucun attribut dans ce contexte.

L'AFC a principalement pour objectif de trouver des ensembles intéressants (appelés concepts) dans les jeux de données (Ferré et al, 2004). Cette technique vise à générer des structures conceptuelles à partir de données. Ces structures peuvent être représentées graphiquement sous forme d'hiérarchies conceptuelles permettant l'analyse de structures complexes et la découverte de dépendances dans les données.

Notre approche de Vecteurs Conceptuels (que nous détaillons dans la **section 3.2.6**) s'inspire fortement de l'AFC pour découvrir des relations d'association sémantiques entre termes.

CHAPITRE III

MÉTHODOLOGIE ET MODÈLE PROPOSÉ

Dans ce chapitre, nous exposons notre modèle proposé et la méthodologie suivie pour assister les ingénieurs d'ontologies à la maintenance d'une ontologie de domaine.

3.1 Méthodologie de recherche

La méthodologie de recherche que nous proposons suit un cheminement relativement classique chez la communauté scientifique. Comme toute problématique, il est primordial, dans un premier temps d'explorer les approches et les outils qui se rattachent d'une façon directe ou indirecte au problème posé. L'analyse critique des approches existantes, de leurs points forts et points faibles nous a permis de formuler un ensemble d'hypothèses (évoquées à la section 1.5.2) sur le modèle à proposer. Notre méthodologie est ainsi fondée sur la validation de ces hypothèses.

Dans cette section, nous décrivons brièvement la méthodologie que nous adoptons, pour proposer un modèle de maintenance des ontologies. Ce dernier sera détaillé dans la section suivante.

Le modèle se base ainsi sur différentes sources de connaissances, à savoir les textes relatifs au domaine, mais aussi des données terminologiques (thésaurus, dictionnaire électronique, base terminologique, etc.) en vue d'intégrer le mieux possible, les connaissances du domaine ainsi que certaines connaissances reliées plutôt à la langue, au sein de l'ontologie.

Maintenir une ontologie de domaine consiste principalement à extraire, à partir de documents, des termes et des relations entre termes qui sont pertinents par rapport à l'ontologie courante. L'objectif principal est de fournir de l'assistance à l'utilisateur pour assurer un raffinement

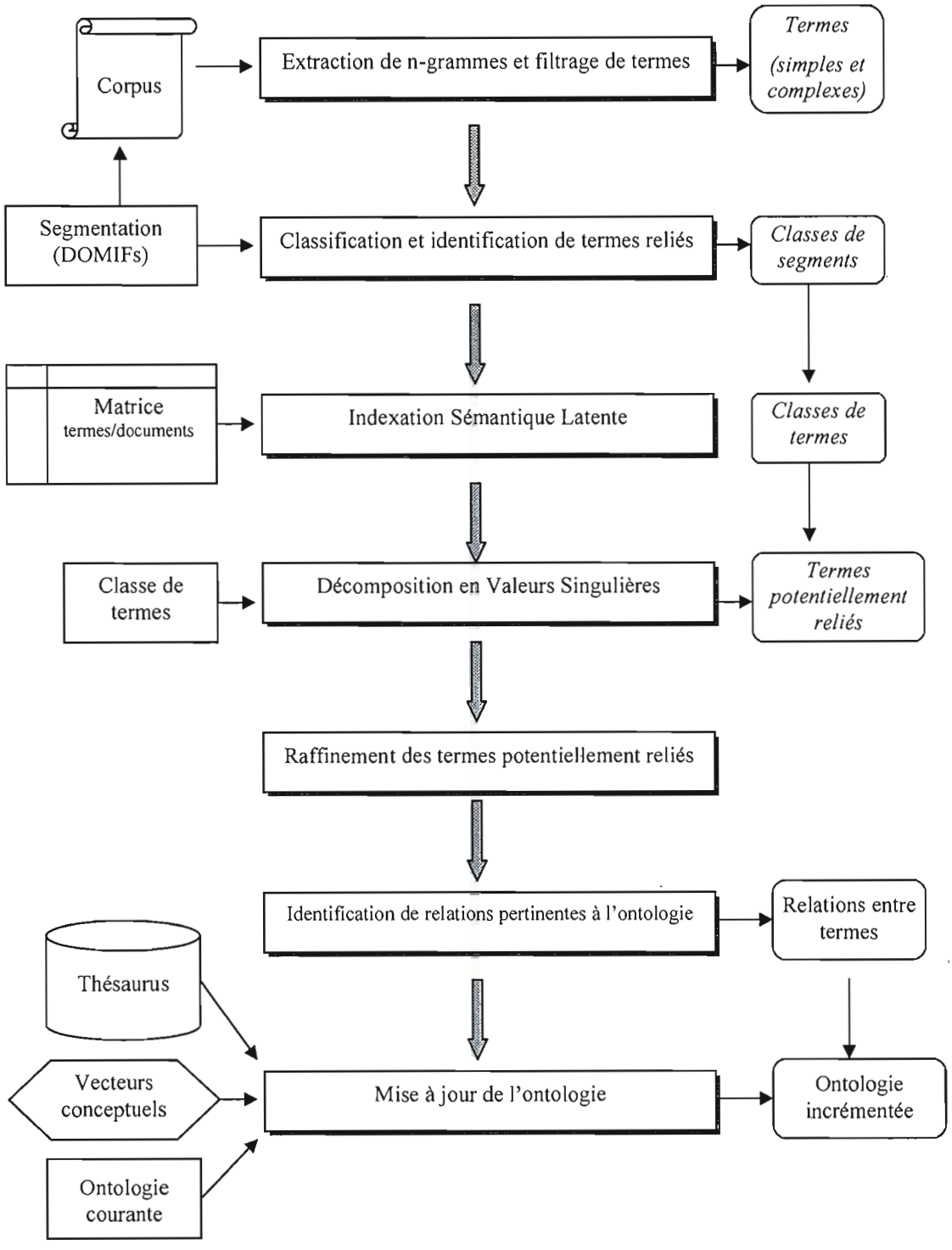


Figure 3.1 : Architecture de la chaîne de traitement ONTOLOGICO

continuel de l'ontologie. Le modèle que nous proposons ne vise pas à remplacer complètement l'humain pour la découverte de relations entre termes, mais plutôt d'assister les experts du domaine dans cette tâche. Il est relativement facile de remettre en cause des systèmes automatiques prétendant accomplir cette tâche sans biais ou imperfections. Il semble plus raisonnable de suivre un processus plutôt semi-automatique impliquant une simple intervention d'un expert du domaine, à travers certaines étapes et spécialement pour la validation des résultats.

Les textes constituent un support considérablement utile, rassemblant des connaissances stables et qui sont utilisées en tant que référence ainsi qu'une ressource précieuse d'analyse et de forage. Par ailleurs, l'accès aux termes et aux textes, justifiant les définitions et les utilisations des termes, assure une meilleure lisibilité du modèle et facilite, par conséquent, la maintenance des ontologies. Partant de cette hypothèse (H1), nous proposons une chaîne de traitements textuels «**ONTOLOGICO**» (Gargouri et al, 2003) constituée des modules suivants : un extracteur de termes, un classifieur, un lemmatiseur, un segmenteur, un module de raffinement sémantique (basé sur l'Indexation Sémantique Latente) et un identificateur de termes reliés (basé sur le calcul de similarité sémantique entre les couples de vecteurs conceptuels).

Le processus débute par l'application de la technique de classification sur un corpus pour identifier dans un premier temps, des groupes de termes qui apparaissent conjointement dans une classe de documents et qui ont potentiellement des relations sémantiques. La capacité des techniques statistiques, à traiter de larges données textuelles, a été le critère de base qui a nous a orientés vers ces approches de classification plutôt que celles fondées sur la linguistique. En effet, bien que plus précises, ces dernières se trouvent confrontées à une difficulté majeure, celle de l'aspect combinatoire. Pour procéder à la classification numérique, nous utilisons **GRAMEXCO** ; une instance de séquence de modules construite à partir de la plate forme **SATIM**[20] (Biskri et Meunier, 2002). La classification, basée sur le réseau de neurones ART1 (Adaptive Resonance Theory) (Grossberg, 1988) permet l'exploration et l'expérimentation de différents types d'analyses grâce à sa modularité, ses diverses fonctions d'analyse et sa capacité d'adaptation par rapport à la croissance des

[20] SATIM peut accepter d'autres types d'information que les textes, tels que : les images, le son, etc.

données textuelles. En particulier, **GRAMEXCO** permet l'exécution d'une séquence de traitements sur des textes pour classifier les segments en se basant sur l'approche des n-grammes (Damashek, 1989).

De ces regroupements de termes on cherche ensuite à extraire des couples de concepts fortement reliés. Cette tâche est accomplie en utilisant la technique d'*Indexation Sémantique Latente* (ISL) (Deerwester et al, 1990) (Srivastava et al, 2002) associée avec la *Décomposition en Valeurs Singulières* (DVS). À l'intérieur des méthodes statistiques de repérage de relations conceptuelles entre termes, la technique de l'ISL a spécialement montré sa fiabilité. Cette technique est entre autres, privilégiée pour sa simplicité et sa justification par des fondements mathématiques précis et solides. Nous pensons toutefois, que cette technique a tout l'intérêt d'être associée avec la classification textuelle pour plusieurs raisons (que nous détaillons plus loin).

La complémentarité entre, d'une part, la technique de classification de documents qui est, dans notre cas, essentiellement basée sur les réseaux de neurones (ART1) d'un côté et l'approche de l'ISL de l'autre côté, constitue un processus de peaufinage très puissant (Gargouri et al, 2003). Ce processus permet de faire émerger à partir des regroupements de termes issus de la classification, les termes qui sont les plus représentatifs de l'information contenue dans les documents d'une même classe. Les termes de plus fort poids générés par l'ISL sont des indicateurs intéressants de la nature des concepts qu'ils représentent.

Partant de l'hypothèse que les données textuelles ne peuvent, toutes seules, supporter la modélisation d'un domaine, du moins à cause des problèmes reliés à l'ambiguïté sémantique, nous avons opté pour l'utilisation d'un thésaurus, afin d'enrichir les connaissances extraites à partir de données textuelles et faire intégrer, le mieux possible, les connaissances du domaine au sein de l'ontologie.

Dans l'objectif de raffiner davantage le processus de repérage de relations entre termes, nous proposons une méthode fondée sur la représentation de termes par des vecteurs conceptuels. Ces vecteurs sont construits à partir des items lexicaux associés à un terme, selon un lien de synonymie, d'antonymie, d'hyperonymie. Ces liens peuvent être extraits à partir d'un thésaurus ou d'une base terminologique telle que Wordnet. En utilisant une mesure de

similarité sémantique entre ces vecteurs conceptuels, qui est le cosinus, nous identifions les couples présentant une forte probabilité d'être sémantiquement associés. Cette méthode repose sur une hypothèse sémantique de Firth selon laquelle, les mots ayant un même environnement lexical sont postulés avoir une sémantique reliée et leurs vecteurs conceptuels sont par conséquent similaires.

Ces relations entre termes proposées d'une part, par la classification et l'ISL et d'autre part, par les vecteurs conceptuels doivent finalement être vérifiées et analysées par un expert pour confirmer leur pertinence par rapport à l'ontologie courante et au domaine. L'expert se charge également de l'étiquetage de ces relations. Enfin, les termes et les relations entre termes (retenues par le processus précédent) sont intégrés dans l'ontologie courante.

Nous proposons dans ce qui suit de détailler notre processus itératif d'ingénierie, organisé sous forme de sept principales étapes.

3.2 Modèle proposé

Le modèle proposé consiste dans ses premières étapes à appliquer une chaîne de traitements textuels. Ainsi, un corpus doit être constitué en sélectionnant des textes à partir d'une documentation technique relative au domaine. Le choix d'un corpus représentatif du domaine ainsi que des contextes d'usage de l'ontologie restreint, en effet, le champ de recherche de termes spécialisés et pertinents pour l'ontologie à maintenir. Ce choix est primordial et délicat, puisque d'une part, le corpus est l'une des sources d'information essentielles et d'autre part, il demeurera, une fois le processus achevé, l'élément de documentation de l'ontologie construite.

La collecte doit se faire avec l'aide des spécialistes et en fonction de l'application cible visée. Il convient en effet de s'assurer auprès des spécialistes que les textes choisis ont un statut suffisamment consensuel pour éviter toute remise en cause ultérieure.

3.2.1 Extraction de n-grammes et filtrage de termes

3.2.1.1 Identification des unités et des domaines d'information

La première étape de notre modèle consiste à identifier, à partir des documents à analyser, les unités d'information (UNIFs) (i.e. les traits descriptifs qui serviront d'ancrage à l'analyse des segments de documents) et les domaines d'information (DOMIFs) (i.e. les segments de documents). Les UNIFs et les DOMIFs issus du corpus serviront d'éléments constitutifs de la matrice qui sera soumise au module de classification. L'extraction des UNIFs et des DOMIFs repose sur plusieurs enjeux théoriques issus principalement de l'analyse statistique et linguistique des données textuelles.

L'identification de ces UNIFs consiste à extraire les éléments sur la base desquels les différents segments du corpus à analyser seront comparés. Ces unités d'information peuvent prendre différentes formes. Elles peuvent être des mots, des mots composés, des phrases, des n-grammes, etc. Cette étape implique des choix théoriques importants. En effet, une analyse des données textuelles fondée sur des mots d'un corpus nous impose d'importants questionnements sur la nature même d'un mot et, de manière plus générale, des unités d'information présentes au sein d'un texte (Meunier, 1995). S'agit-il uniquement d'une suite de caractères ? Un mot se définit-t-il par la séparation spatiale, l'identité morphologique, etc.? Ces questions fondamentales doivent nécessairement être abordées avant même d'envisager tout projet d'analyse textuelle.

Ces choix théoriques doivent prendre en compte les buts envisagés, les résultats que l'on espère découvrir lors de nos recherches, car c'est à partir de ces éléments que la classification sera effectuée. Dans le cadre de notre projet, les unités d'information retenues sont composées de l'ensemble des n-grammes du corpus. Ces n-grammes[21] peuvent être définis comme une séquence de N caractères (par exemple, les séquences de trois caractères sont appelées des tri-grammes). Ces deux objets forment la matrice qui sera utilisée par le classifieur. En d'autres termes, les segments seront comparés et classifiés sur la base de la cooccurrence de n-grammes.

---

[21] Par exemple, dans la phrase «La vie est belle» se retrouve l'ensemble de tri-grammes suivants : {la_, a_v, _vi, vie,...}.

Cette approche constitue un remède à la problématique d'identification des termes complexes dans un corpus. En effet, les n-grammes prennent en compte, d'une façon indirecte, les formes collocationnelles des termes. Une classification basée sur des mots aurait considérée, d'une façon séparée, les mots « pomme », « de » et « terre » comme critère de cooccurrence à travers un corpus. Par contre, une représentation basée sur des n-grammes prend en considération, non pas simplement la cooccurrence des mots, mais également celle de leurs juxtapositions, à l'instar de « E_D » ou « E_T » dans « Pomme de terre ».

Les espaces entre les mots sont considérés dans la composition des n-grammes. Ainsi, le terme « Pomme de terre » est transformé en trigrammes comme suit[22] :

{POM, OMM, MME, ME_, E_D, _DE, DE_, E_T, _TE, TER, ERR, RRE, RE_}

Comparativement à d'autres techniques, les n-grammes capturent automatiquement les racines des mots les plus fréquents. Ainsi, pour des mots comme (déplace, déplacez, déplacement, déplacés, ...), il n'est pas nécessaire de chercher la racine.

Contrairement aux modèles basés sur les mots pour lesquels il faut utiliser des dictionnaires spécifiques (féminin-masculin ; singulier-pluriel ; conjugaisons ; etc.) pour chaque langue, les n-grammes peuvent éventuellement être utilisés sans lemmatisation, bien que cette dernière soit avantageuse pour certains corpus. Les n-grammes opèrent ainsi, indépendamment des langues. L'analyse de texte en termes de n-grammes demeure valide pour un texte écrit en toute langue basée sur un alphabet et la concaténation d'opérateurs de construction de textes.

Il s'agit également d'un avantage considérable répondant à la problématique mentionnée plus haut : qu'est-ce qu'un mot ? Avec les n-grammes, il n'est pas nécessaire de segmenter préalablement le texte en mots ; ceci est intéressant pour le traitement de langues dans lesquelles les frontières entre mots ne sont pas fortement marquées, comme le chinois. C'est moins intuitif, mais les résultats sont théoriquement justifiés et tout à fait acceptables en pratique.

[22] Le caractère « _ » est utilisé pour représenter un espace.

Par ailleurs, les n-grammes sont tolérantes aux fautes d'orthographes. Par exemple, si le mot « piscine » est écrit «pissine », un modèle basé sur les mots aura du mal à reconnaître ce mot. Par contre, un modèle basé sur les n-grammes est capable de prendre en compte les autres n-grammes comme « _PI », « PIS », « INE », « NE_ », etc.

Enfin, l'utilisation des n-grammes de caractères à la place des mots, offre un autre avantage important : elle permet de contrôler la taille du lexique utilisé par le processeur, tel qu'illustré dans (Lelu et al, 1998). La taille du lexique à traiter constitue effectivement un facteur important à considérer quand il est question d'un traitement complexe de corpus.

Il importe, par la suite, d'identifier les différents segments de textes qui seront comparés entre eux. Plusieurs méthodes de segmentation sont explorées en littérature. Parmi les plus fréquemment citées, on retrouve celles fondées sur des marqueurs de discours (Callan, 1994) et sur les marqueurs sémantiques (Kaszkiel et al, 2001). Dans le cadre de notre projet, nous privilégions une segmentation fondée non pas sur certaines propriétés sémantiques des documents (comme c'est le cas dans ces deux méthodes), mais plutôt sur des séquences ou suites de mots.

Encore une fois, des enjeux théoriques s'imposent. De quelle nature doivent être les segments à comparer ? Doivent-ils être identifiés sur la base de critères linguistiques ? Est-il préférable de segmenter le corpus à analyser en textes, en paragraphes ou en phrases ? Cette décision relève du chercheur et dépend des données à analyser. Examinons tour à tour les arguments avancés pour la segmentation.

Pour ce qui est de la phrase, par les constructions syntaxiques qui la structurent, par le voisinage étroit propice aux influences et interactions sémantiques, la phrase a une pertinence cognitive, en ce que sa taille correspond aux capacités de la mémoire à court terme. Cet atout linguistique se double de vertus statistiques : la phrase courante a une taille suffisamment petite pour être sélective. D'une part, cela évite le foisonnement de cooccurrences examinées lors du processus de classification et d'autre part, la significativité statistique des écarts de répartition ainsi mesurés est accrue.

Le paragraphe correspond conventionnellement à une unité au plan sémantique. Il n'est pas assujetti aux limites de la phrase qui peuvent être trop restrictives en ce qui concerne le développement d'une thématique. Le paragraphe possède également une dimension cognitive, au plan de la mémorisation comme de la perception.

Une segmentation par texte revient à considérer une unité sémantiquement autonome. C'est aussi tirer profit de la cohérence sémantique qui traverse le texte, et qui fait que des notions se font écho du début à la fin du texte, d'une partie à une autre. Pour la classification des genres « brefs » et « focalisés » (tels que des annonces sur des forums électroniques), le texte est le segment à considérer. Pour les documents plus longs, un découpage par paragraphe donne des contextes plus développés que les phrases, tout en ayant une certaine autonomie, une longueur à priori plus régulière et une cohérence interne forte.

En général, il semble préférable d'opter pour une segmentation qui n'est ni trop volumineuse (spécialement lorsqu'il s'agit de documents très homogènes), ce qui donnerait lieu à une classification très grossière et à une perte d'information importante, ni trop fine car à l'inverse, les unités à comparer seraient alors beaucoup trop différentes les unes des autres pour les soumettre à toute forme de classification. Dans notre projet, nous utilisons une segmentation par paragraphe. Cette méthode est en effet, plus avantageuse compte tenu des avantages cités, mais aussi parce qu'elle est «computationnellement» peu coûteuse, tout en ne nécessitant aucune propriété structurelle explicite des documents (Kim et al, 2004). Le grand volume de données textuelles que nous traitons justifie en effet ce choix.

3.2.1.2 Filtrage du lexique

Si les n-grammes servent uniquement à la classification, l'extraction du lexique joue un rôle plus actif dans les étapes suivantes. L' « *extracteur de termes* » (un module de **GRAMEXCO**) est utilisé pour identifier le lexique (ensemble de lexèmes) à partir d'un corpus. Avant de traiter ce lexique et d'extraire les n-grammes, des opérations de filtrage doivent être réalisées pour garantir des résultats plus fiables.

Cette étape consiste à appliquer différents filtres statistiques et linguistiques. La nature et l'importance de cette étape sont le centre de plusieurs débats théoriques. Plusieurs recherches

montrent qu'un filtrage adéquat du lexique de départ permet non seulement de diminuer substantiellement le temps nécessaire au traitement du corpus, mais aussi d'éliminer plusieurs éléments susceptibles d'affecter l'analyse et l'interprétation des résultats (Frakes et al, 1992).

Le lexique (représenté par les n-grammes) constitue l'une des deux composantes principales de la matrice traitée par le classifieur. La dimension de l'espace vectoriel dans lequel les domaines d'information seront comparés affecte directement le temps nécessaire à la classification de chacune des entrées de la matrice.

L'opération de filtrage du lexique est composée traditionnellement de plusieurs sous opérations. Il s'agit d'abord, de supprimer certains termes non pertinents à l'analyse. Le fait de conserver ces termes, en plus d'affecter directement la qualité de la classification obtenue, ajoute aussi du bruit dans les résultats.

Le filtrage du lexique peut être effectué à l'aide de plusieurs techniques, certaines étant d'inspiration linguistique, d'autres de nature statistique. Une première opération a pour but de supprimer l'ensemble des mots fonctionnels présents dans le texte. Les mots fonctionnels ou les « mots vides » tels que {le, la, dans, à, etc.} peuvent être définis comme étant l'ensemble des mots non pertinents à l'égard des buts poursuivis.

Par la suite, une deuxième opération de nature statistique peut être appliquée au lexique. Il s'agit d'éliminer les termes qui, tout en ne figurant pas dans la liste des mots fonctionnels, ne sont pas pertinents à l'analyse. Dans une perspective de classification, la pertinence des termes est évaluée en fonction du rôle discriminatoire de ces termes. C'est en effet sur la base de ce rôle que la classification de segments de documents sera réalisée. Ainsi, afin d'optimiser les résultats obtenus, il importe de supprimer les mots dont la fréquence est supérieure ou inférieure à un certain seuil (Schultz, 1969 ; Van Rijsbergen, 1979). Toutefois, compte tenu du fait qu'après le processus de classification, le lexique subira d'autres traitements, la suppression des mots très fréquents risque d'éliminer des termes pertinents pour l'ontologie à maintenir. Il importe par conséquent, de ne pas filtrer de tels mots, bien que le résultat de la classification soit légèrement de moindre qualité.

Une troisième étape de filtrage du lexique consiste à supprimer manuellement tous les termes non pertinents qui ont néanmoins résisté aux deux premières opérations.

Finalement, une dernière étape de traitement du lexique est nécessaire afin d'optimiser le processus de classification. Dans la mesure où le classifieur utilisé est purement statistique, les différentes variantes sémantiques et syntaxiques présentes dans un corpus ne sont pas considérées par le classifieur. Il importe alors d'appliquer au lexique du corpus une opération de lemmatisation en vue de remplacer les termes par leurs lemmes correspondants. En effet, des termes tels que {*informe, information, informant, etc.*} réfèrent au même concept, et doivent par conséquent, être analysés en tant qu'un terme unique dans les étapes suivantes. Cette opération, non seulement réduit substantiellement le nombre de lexèmes présents dans un corpus, mais aussi elle ajuste la fréquence des termes et par conséquent, leurs rôles discriminatoires.

L'opération de lemmatisation est généralement réalisée en effectuant d'abord un marquage morphosyntaxique des différents lexèmes à analyser, puis en les comparant à un dictionnaire. Ce processus permet de générer une liste de lemmes propres à la langue en question. La lemmatisation est toutefois, une opération très délicate, car elle implique un processus complexe de désambiguïsation sémantique. Les logiciels automatiques de désambiguïsation sémantique ne sont pas totalement précis et nécessitent une intervention humaine (Brunet, 2002).

3.2.2 Classification et repérage de termes reliés

L'objectif de la classification est d'extraire certains types de régularités sémantiques entre les segments du texte (Manning et al, 1999 ; Sebastiani, 2002 ; Gelbukh et al, 1999). Ces segments contiennent un type d'information similaire et servent par conséquent, à détecter des indices précieux aux associations entre termes. Il s'agit d'une opération de classification des segments de documents qui consiste à regrouper les différentes données dans des classes les plus homogènes possible en employant uniquement les traits caractéristiques ayant servi à décrire chaque segment.

Compte tenu des récents travaux dans le domaine de la classification des données, plusieurs possibilités sont ouvertes. Le choix d'une approche au détriment d'une autre fait intervenir des considérations concernant tant la nature des données à regrouper que les caractéristiques de la classification souhaitée (exclusivité[23], hiérarchie[24], dynamicité[25], etc.).

Dans le domaine du repérage de l'information, la classification de documents est un processus non supervisé, opéré sur un grand volume de données. Il est habituellement exécuté en utilisant un classifieur numérique tel que exploré dans (Meunier et al, 1997 ; Memmi et al 1998 ; Benhadid et al, 1998 ; Biskri et al, 1999). En tant que méthode de forage de données textuelles, ce processus souvent moins détaillé que les approches linguistiques et conceptuelles, permet une première exploration générale et rapide du corpus. Il identifie les classes de segments et groupes de lexèmes ayant des associations connues sous le nom de cooccurrence et détecte par conséquent, leurs réseaux sémantiques (Church et al, 1989 ; Lebart et al, 1988 ; Salton, 1989).

Diverses méthodes de classification automatique ont été explorées et appliquées avec succès au traitement automatique des documents. Parmi les méthodes les plus fréquemment citées, on retrouve la méthode des k-moyens, les diverses méthodes neuronales à savoir les cartes auto-organisatrices de Kohonen (Kohonen, 2001), le réseau neuronal ART et ses variantes (ART1 (Grossberg et al, 1987a), ART2 (Carpenter et al, 1987b), Fuzzy ART (Grossberg et al, 1991), ARTMAP (Carpenter et al, 1991), Fuzzy ARTMap (Grossberg et al, 1992), Gaussian ARTMAP (Williamson, 1995) etc.), la technique des plus proches voisins (« nearest neighbor clustering ») (Jain et al, 1999) (Yang, 1999 ; Yang et al, 1999), les « Support Vector Machines » (SVM) (Joachims, 2002), etc. Dans notre projet, nous avons utilisé le classifieur neuronal ART1 pour les raisons que nous détaillons plus loin.

Plusieurs chercheurs avaient traditionnellement favorisé l'utilisation de réseaux neuronaux à rétropropagation. Toutefois, les recherches dans ce domaine ont rapidement démontré que malgré l'efficacité d'une telle méthode, celle-ci n'est pas dynamique et se heurte rapidement

[23] Un segment ne peut faire partie que d'une seule classe.
[24] Les classes sont organisées d'une façon hiérarchique (une classe contient des sous-classes plus spécifiques)
[25] C'est la possibilité de changer les relations entre prototypes de classes durant l'exécution du processus de classification

à un important problème de plasticité. En effet, lorsque la classification est effectuée sur un ensemble dynamique d'intrants, ce type de réseaux ne permet pas d'intégrer de nouveaux vecteurs au processus de classification déjà entrepris. Selon cette approche, il est nécessaire de reprendre de nouveau le calcul déjà effectué pour tenir compte des modifications subies par les vecteurs intrants. Cette opération est computationnellement bien coûteuse quand il est question d'un corpus de grande taille et évolutif.

Il est tout de même possible de contrer cette limite des réseaux à rétropropagation et de faire en sorte qu'ils soient plastiques, en utilisant un réseau de neurones distinct dont la particularité est de ré-effectuer la phase d'entraînement sur les nouveaux vecteurs intrants. Ce dernier réseau pourra ainsi, s'adapter aux différents changements de son environnement et traiter de manière dynamique les intrants différents, mais se ressemblant. Toutefois, la majorité des réseaux plastiques ne peut satisfaire le critère de stabilité selon lequel, le système doit conserver dans le temps les structures reconnues (connaissances acquises) malgré la différence des stimuli intrants. Autrement dit, ces réseaux ne peuvent conserver dans le temps leur apprentissage et par conséquent, la qualité des résultats obtenus par ces réseaux diminue rapidement dans le temps, au fur et à mesure de l'ajout des nouveaux intrants.

Dans un cadre d'apprentissage permanent, se pose crucialement le dilemme *plasticité / stabilité* : il faut en effet que le même réseau apprenne constamment de nouvelles singularités sans oublier les anciennes ; il faut également pouvoir ignorer des singularités non pertinentes. C'est ce dilemme entre plasticité et stabilité qui est à la base des travaux de Grossberg sur la théorie de la résonance adaptative (*Adaptative Resonance Theory*) (Grossberg et al, 1987). Un système neuronal efficace doit constamment passer d'un mode plastique à un mode stable et vice versa. Il doit être en mesure de conserver l'information antérieure, à savoir les classes antérieures, mais en même temps de tenir compte de la nouveauté, c'est-à-dire les nouveaux intrants. Il doit donc stabiliser les classes qu'il découvre, mais aussi les changer, si cela est nécessaire, en regard de la réalité nouvelle qui se présente à lui. Pour sa part, le modèle ART1 adhère en effet à ce principe en cherchant à contrôler la qualité des intrants et donc à arriver à une meilleure classification.

Tous les modèles de la famille ART (**Tableau 3.1**) normalisent les intrants, réduisent le bruit et stabilisent les patrons dans le temps. La normalisation des intrants consiste à imposer un même seuil à tous les neurones pour garantir une certaine stabilisation des intrants, spécialement lorsque leur impact fluctue de stimulus en stimulus. La réduction du bruit consiste en la définition de paramètres d'intensité avec lesquels un intrant peut agir. Finalement, la stabilisation des patrons dans le temps consiste à considérer un intrant en regard des intrants antérieurs, faisant en sorte que le système ne perd pas l'information acquise antérieurement.

Nom du modèle ART	Description
ART1 (Carpenter, Grossberg, 1987a)	- Réseau classifieur (clustering network) - Apprentissage non supervisé - Patrons d'entrées binaires
ART2 (Carpenter, Grossberg, 1987b)	- Réseau classifieur (clustering network) - Apprentissage non supervisé - Patrons d'entrées analogiques ou binaires
Fuzzy ART (Carpenter, Grossberg, Rosen, 1991)	- Réseau classifieur (clustering network) - Apprentissage non supervisé - Incorporation des notions de la logique floue
ARTMAP (Carpenter, Grossberg, 1991)	- Réseau classifieur - Apprentissage supervisé - Agencement de deux unités ART pour former un réseau à apprentissage supervisé, patrons d'entrées binaires ou analogiques selon le type de réseaux ART qui le constitue ART1 ou ART2 respectivement
Fuzzy ARTMAP (Carpenter, Grossberg, Markuzon, Reynolds, Rosen, 1992)	- Comme ARTMAP mais les ARTs sont remplacés par des FuzzyARTs.
Gaussian ARTMAP (Williamson, 1995)	- Comme ARTMAP, mais la fonction de choix et la fonction de correspondance des modules ART sont définies comme des Gaussiennes et par conséquent, le réseau est plus performant et plus résistant au bruit

Tableau 3.1 : Description de la famille des modèles du réseau de neurones ART (Adaptive Resonance Theory)

Nous avons opté pour la version ART1 qui se distingue en partie, des autres versions (ART2, Fuzzy Art) par le fait qu'elle ne traite que des vecteurs intrants de nature binaire et ce de manière non supervisée. Nous nous attardons ici uniquement sur la version ART1.

3.2.2.1 Matrice UNIFs/DOMIFs

L'opération de classification nécessite une étape préliminaire de traduction du corpus selon un modèle vectoriel (Salton, 1989). L'objectif est donc de parvenir à une représentation matricielle du texte de départ. Il s'agit, à partir des résultats obtenus lors du processus de segmentation et de ceux obtenus lors de l'identification et du filtrage des unités d'information, de constituer une matrice formée d'une part, des segments de textes et d'autre part, des unités d'information. La matrice UNIFs/DOMIFs est de type binaire, c'est-à-dire 0 ou 1, en fonction de la présence ou de l'absence de l'unité d'information dans le segment. Cette matrice sera soumise en intrant au module de classification, ART1.

3.2.2.2 Classifieur neuronal ART1

Le modèle ART1 (*Adaptive Resonance Theory*) (Grossberg et al, 1987) est un algorithme de classification (« *clustering* ») ou de regroupement de type adaptatif, auto-associatif et non supervisé. Ce type d'algorithme possède la particularité de prendre comme intrants un ensemble de vecteurs et suite à une phase de traitement, de créer des ensembles sous forme de regroupements de vecteurs. Le regroupement s'effectue sur la base de certains critères de similarité. Plusieurs études ont démontré son efficacité à des fins de classification de documents (Massey, 2003a ; 2003b).

Le principal avantage du modèle ART1 réside dans sa capacité à traiter les intrants de manière dynamique. À cet égard, l'algorithme ART1 peut construire, par étapes, des classes mais aussi s'adapter à un corpus lui-même changeant. Il est d'ailleurs parmi les rares qui puissent apprendre dans un environnement variant constamment. Il s'agit en d'autres termes, d'un réseau de neurones doublement dynamique.

L'idée principale du modèle ART1 est celle d'un système d'interaction entre deux niveaux qui entrent en phase de résonance (**Figure 3.2**).

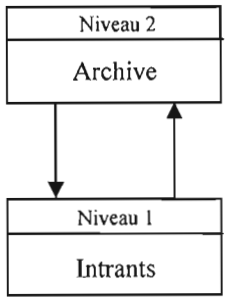


Figure 3.2 : L'interaction entre les intrants et l'archive dans ART1.

Le système reçoit au premier niveau N1, des stimuli intrants sous forme de vecteurs binaires représentant le premier élément à classer. Ces stimuli sont transformés selon une distribution et un poids particuliers et envoyés au niveau d'archivage N2 pour servir de gabarit (prototype) auquel les intrants suivants seront comparés. Le patron archivé au niveau N2 sert d'hypothèse de classe. Dans le cas où le nouvel intrant se distingue radicalement du patron initial (selon un critère ou paramètre de vigilance ρ déterminé par l'utilisateur du système), un nouveau patron sera à son tour créé et servira éventuellement de gabarit aux autres intrants. La comparaison se fait selon un critère ou paramètre de vigilance ρ prédéterminé par l'utilisateur du système. Dans le cas où le nouvel intrant se présente comme étant relativement comparable au patron initial, il est regroupé (selon des paramètres) avec ce même patron. C'est dans cette perspective qu'il importe de concevoir le phénomène de résonance. Il s'agit de la correspondance entre les patrons prototypes et les patrons intrants. Une consolidation émerge de cette résonance au fur et à mesure que se poursuit le processus d'apprentissage. L'adaptation se produit dans la modification constante des interconnexions entre les deux niveaux.

Les principaux mécanismes de fonctionnement de ce système se résument en cinq étapes :

1- **Initialisation** : Dans un premier temps, le paramètre de vigilance ρ est défini par l'utilisateur du système à l'aide d'une valeur comprise entre 0 et 1. Plus la valeur est proche de (1), plus la classification est stricte en terme de comparaison des stimuli intrants avec les patrons. La valeur choisie influe grandement sur la qualité ainsi que

la taille des classes à obtenir. Ainsi, plus la valeur est proche de zéro, plus les classes obtenues seront volumineuses mais peu nombreuses. Par contre, une valeur du paramètre de vigilance près de (1) résultera en une classification plus fine et en une multiplication du nombre de classes obtenues. En pratique, la valeur de ce paramètre est définie par l'utilisateur par essai et erreur. Le système est par la suite, entraîné afin de déterminer l'ensemble des vecteurs prototypes auxquels seront comparés les intrants suivants.

2- **Classification** : Un nouveau stimulus représenté par un vecteur est ensuite introduit afin d'être classé. Le système initialise alors l'entrée du vecteur suivant et le compare à l'ensemble des vecteurs prototypes candidats.

3- **Identification du patron le plus proche** : le système identifie le vecteur prototype le plus proche du vecteur intrant et calcule la distance entre le vecteur prototype sélectionné et le vecteur intrant.

4- **Vérification** : Si le vecteur intrant est suffisamment près du vecteur prototype (selon le seuil de vigilance ρ), le vecteur intrant est inséré dans la classe décrite par le vecteur prototype sélectionné et le vecteur prototype est ajusté. Si le vecteur intrant est trop différent, les vecteurs prototypes seront ajustés.

Le même mécanisme est repris de nouveau avec un nouvel intrant à partir de la deuxième étape.

3.2.2.3 Extraction de termes reliés

La classification utilise en tant qu'entrée un modèle vectoriel qui considère le texte dans sa totalité et vise à inférer à partir des textes, une structure sémantique implicite (Salton et al, 1983). Ce modèle traduit un texte sous forme d'espace matriciel qui associe les segments de textes avec les termes (ici les n-grammes) et produit par conséquent, des réseaux (classes) de termes correspondant à des thèmes traités dans le texte (Memmi, 2000). En effet, la cooccurrence des termes à l'intérieur de différentes parties de texte implique une couverture fort probable d'un même thème. Ainsi, les segments d'une même classe partagent un ensemble de termes communs ayant de fortes chances d'être reliés.

Le processus de classification génère finalement des classes de termes potentiellement reliés. Chaque classe de termes est construite à partir de l'intersection des termes qui cooccurrent ensemble à travers une classe de segments. Ce sont en fait ces termes en particulier qui ont fait en sorte que les segments de documents se regroupent dans une même classe.

Bien que les classes de termes renferment des relations sémantiques, elles ne sont pas dépourvues de bruit. Les étapes suivantes visent donc à déterminer des associations plus précises entre des couples de termes.

3.2.3 Indexation Sémantique Latente

À ce niveau, nous souhaitons extraire, à partir de ces classes de termes, ceux représentant un niveau élevé de corrélation. L'Indexation Sémantique Latente (ISL) est reconnue pour sa capacité à découvrir des relations sémantiques (Deerwester et al, 1990 ; Srivastava et al, 2002). Elle a été employée depuis les années 90 pour la recherche d'information sémantique à partir des textes, bien que les premiers travaux sur la cooccurrence aient commencé depuis les années 70. Cette technique a été utilisée, spécialement pour sa simplicité et sa justification par des fondements mathématiques assez précis.

Initialement, l'ISL a pour but d'indexer de manière automatique des documents, c'est-à-dire leur affecter des mots clés. L'idée directrice est de représenter les mots et les documents dans un même espace, de manière à pouvoir les comparer. L'ISL réduit la dimension des documents à analyser et restreint par conséquent, l'étendue du problème (Manning et al, 1999). Les nouvelles dimensions sont une meilleure représentation des documents et des termes. La métaphore exprimée par le terme « latent » est que les nouvelles dimensions constituent la vraie représentation.

L'ISL est une alternative aux mesures de similarité telles que « td-idf » (Rosario, 2000). Sa spécificité réside dans le fait qu'elle ne reflète pas la cooccurrence entre termes mais plutôt les relations sémantiques (Wade-Stein et al, 2004). Elle s'appuie sur des fondements mathématiques solides permettant d'inférer des relations sémantiques plus profondes (Landauer et al, 1998). Ces fondements mathématiques sont basés sur la réduction de la dimensionnalité d'un espace vectoriel. L'ISL utilise une méthode statistique appelée la

Décomposition en Valeurs Singulières (DVS) pour découvrir les associations de termes à travers les documents. La DVS est une méthode statistique des moindres carrés, similaire à la régression linéaire. L'objectif de cette technique est de regrouper ensemble dans l'espace de projection, les mots et les documents qui sont associés en vue de capturer la structure sémantique enchâssée dans les associations entre les termes et les documents de la collection.

Formellement, L'ISL utilise en tant qu'entrée, une matrice termes-documents correspondant aux poids des termes dans les documents du corpus. Nous précisons ici que le corpus est considéré comme une collection de documents. Chaque terme T_i se voit affecté un poids $w_{i,k}$ selon le nombre d'occurrences $C_{i,k}$ dans le document D_k. Ce poids donné par la formule suivante :

$$w_{i,k} = \frac{C_{i,k}}{\sum_{j=1}^{n_k} C_{j,k}}$$

n_k est le nombre total de termes dans le document. Ces termes sont limités à ceux filtrés et sélectionnés pour la classification.

Les statistiques pour chaque document individuel sont combinées en vue de produire une analyse statistique pour la totalité de la collection. Une normalisation de la longueur de documents, telle qu'expliquée dans (Greengrass, 1997), est utilisée pour éviter le fait qu'un terme peut avoir un poids élevé, simplement parce que le document, dans lequel il apparaît est court, plutôt qu'en raison de sa fréquence élevée à travers la collection de documents. Les poids des termes normalisés deviennent :

$$W_{i,k} = \frac{w_{i,k}}{\sqrt{\sum_{j=1}^{n_k} w_{j,k}^2}}$$

Nous appliquons l'ISL, non pas sur tous les termes du corpus, mais plutôt sur les termes faisant partie des classements générés par ART1 (Gargouri et al, 2003). Pour chaque classe de termes c, les poids correspondants forment la matrice W_c. Cette matrice de termes-

documents est constituée de lignes représentant m_c termes (T_i^c) appartenant à la classe c, et de colonnes représentant la collection des n documents (D_k).

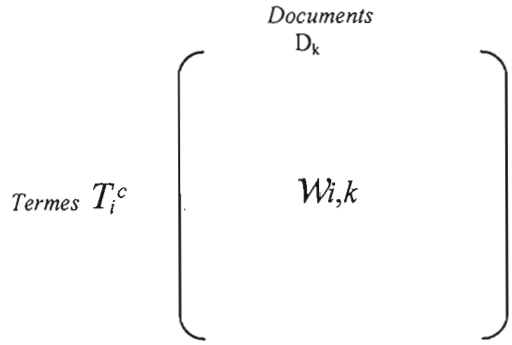


Figure 3.3 : La matrice terme-document W_c

3.2.4 Décomposition en Valeurs Singulières

À l'intérieur de chaque classe de termes, nous visons la détermination des couples de termes reliés. La méthode ISL relie des termes sous forme d'une structure sémantique intéressante, telle que détaillée dans (Berry et al, 1995). L'ISL représente les documents par des concepts qui sont réellement et statistiquement indépendants de telle sorte que les termes ne le sont pas. Un concept est considéré ici en tant qu'un ensemble de termes reliés. L'ISL implique principalement la décomposition de la matrice W_c en utilisant la Décomposition en Valeurs Singulières (DVS) (Golub et al, 1970), qui est un type de régression linéaire. Alors, W_c peut être décomposée comme suit :

$$W_c = U \Sigma V^T$$

où U est une matrice de termes ($m_c \times r$), V est une matrice de documents ($r \times n$) et Σ est une matrice ($r \times r$), où r est le rang de W_c. Σ est une matrice diagonale contenant les valeurs singulières de W_c. Dans cette décomposition, la valeur singulière σ_i correspond au vecteur u_j (la $j^{ème}$ colonne de U) et v_i (la $i^{ème}$ ligne de V). Les colonnes de U, les lignes de V et les valeurs diagonales de Σ sont arrangées de sorte que les valeurs singulières sont dans un ordre décroissant, en descendant la diagonale. Cette transformation de formule n'entraîne aucune perte de généralité.

Les valeurs singulières de Σ inférieures à un seuil de pourcentage de la valeur singulière la plus large, σ_1 sont éliminées (Deerwester et al, 1990 ; Nicholas et al, 1998). W_c^S représente en d'autres termes, une approximation de W_c dont l'exactitude s'accroît au fur et à mesure que *s* s'approche de *r* :

$$W_c^s = U^s \Sigma^s V^{sT}$$

où Σ^s est dérivée de Σ en éliminant toutes les valeurs sauf les *s* valeurs singulières les plus grandes, U^s est dérivée de U en éliminant toutes les valeurs sauf les *s* colonnes correspondant aux valeurs singulières les plus grandes, et V^s est dérivée de V en éliminant toutes les valeurs sauf les *s* lignes correspondant, où $s \leq r$. L'algorithme DVS (**Annexe 6**) détaille les itérations de cette transformation.

Le produit de deux vecteurs lignes de W_c évalue jusqu'à quelle mesure deux termes présentent un patron similaire d'occurrences à travers l'ensemble de documents. La matrice $W_c . W_c^T$ est une matrice symétrique carrée contenant tous ces produits de vecteurs termes à termes.

La décomposition en valeurs singulières donne une matrice orthonormée V. Par conséquent, $V^T.V=I$ où I est la matrice identité.

$$W_c . W_c^T = (U \Sigma V^T)(U \Sigma V^T)^T = (U \Sigma V^T)(V^{T^T} U^T \Sigma^T) = U \Sigma V^T V \Sigma^T U^T = U \Sigma \Sigma^T U^T$$

Le produit $\Sigma \Sigma^T$ est une matrice diagonale. U est alors faite des vecteurs propres de $W_c . W_c^T$ (Deerwester et al, 1990).

U^s est le composant le plus important pour nous. Cette matrice ($m_c \times s$) représente en effet, les corrélations terme à terme dans la collection de documents et appartenant à la classe *c*. Chaque colonne de la matrice, u_{i_n} est un vecteur dont les éléments sont associés à un certain concept. Les éléments de u_i indiquent la corrélation terme à terme. Le fait de mettre à zéro

tous les éléments dans u_i qui sont inférieurs à un seuil du pourcentage des termes les plus corrélés dans u_i, permet d'éliminer les termes qui sont faiblement associés (**Figure 3.4**).

Bien qu'en pratique les valeurs de U^s soient difficilement interprétables, elles indiquent tout de même la présence ou non d'une corrélation entre le terme et le concept en question.

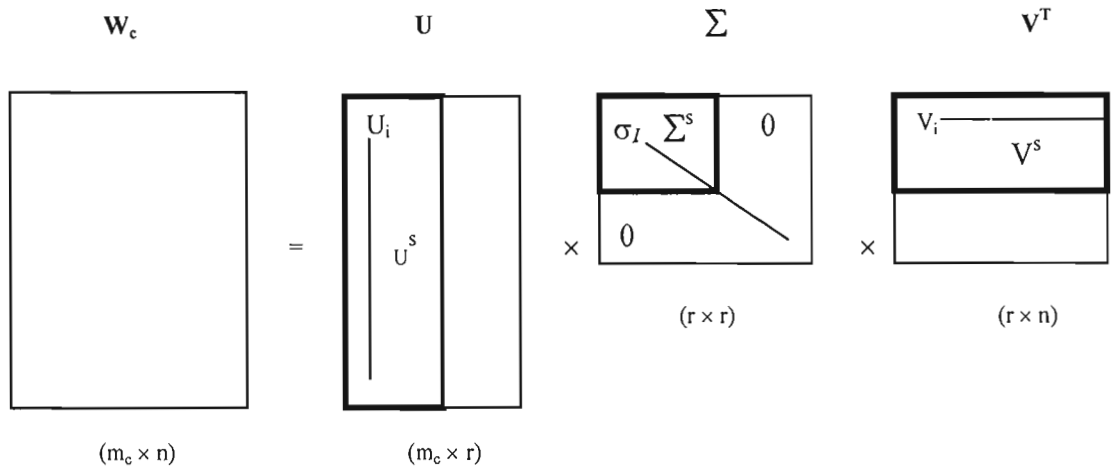


Figure 3.4 : Décomposition en Valeurs Singulières de la matrice terme/document W_c

L'application de la technique d'ISL sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du texte, possède l'avantage de réduire la matrice de cooccurrence de termes dans les documents à une dimension raisonnable. En effet, malgré la difficulté d'identifier dans la théorie une dimension adéquate et précise de cette matrice, il est facile de constater qu'une dimension immense pourrait empêcher l'émergence de suffisamment de relations sémantiques entre les termes, et aussi, une dimension trop réduite pourrait entraîner une grande perte d'information. (Deerwester et al, 1990) ont établi le nombre optimal de valeurs singulières à retenir, à 100 pour des collections de quelques milliers de documents. Ils ont démontré une amélioration moyenne de la précision de 13% par rapport au modèle vectoriel, sur tous les niveaux de rappel, sauf dans les deux plus bas niveaux. (Berry et al, 1995) ont obtenu des améliorations plus élevées (jusqu'à 30% sur la précision) au cours d'essais similaires.

En utilisant l'espace conceptuel, l'ISL résout deux principaux problèmes, auxquels la recherche d'information est souvent confrontée, à savoir, la polysémie et la synonymie :

- La polysémie ou le fait que la plupart des mots aient plus d'un sens et que ce dernier soit connu à partir du contexte du mot. Ainsi, une référence au mot « livre » aurait un sens différent dans des documents sur la bibliographie ou le marché financier ou le transport.

- La synonymie ou le fait d'avoir plusieurs possibilités pour décrire un même objet. Le problème de synonymie tend à altérer la performance en terme de taux de rappel, des systèmes de Recherche d'Information (Deerwester et al, 1990).

Toutefois, l'ISL a été critiquée en pratique à cause du fait que le traitement informatique de la DVS nécessite un énorme stockage en mémoire ce qui rend le traitement relativement long. Une autre critique de l'ISL est relative au fait que la DVS soit conçue pour des données présentant une distribution normale, alors que la matrice terme à terme (même si elle est pondérée) pour une collection de documents n'a nécessairement pas une distribution normale (Rosario, 2000). Pour remédier à cela, certains auteurs préconisent une réduction de la dimensionnalité basée sur une distribution de Loi de Poisson qui pourrait conduire à une meilleure approximation de la matrice terme/document (Manning et al, 1999).

À ce niveau du processus, tous les couples possibles parmi les termes restants, sont présumés être potentiellement reliés. Il convient toutefois, de procéder à un filtrage manuel de ces termes pour éliminer ceux jugés non spécifiques au domaine, dans la mesure où seules les relations entre termes propres au domaine sont pertinentes pour l'ontologie. À cet effet, le bon sens de l'expert du domaine est retenu comme critère principal de filtrage. Nous présentons dans la section suivante l'apport des thésaurus et l'approche des Vecteurs Conceptuels pour assister ce jugement.

3.2.6 Vecteurs Conceptuels et thésaurus

Le processus itératif de raffinement de relations entre termes se poursuit avec les groupes de termes retenus. À ce niveau, l'objectif est d'identifier des paires de termes possédant une certaine relation sémantique. Bien que l'extraction de connaissances à partir d'analyses

textuelles soit importante pour refléter la réalité du domaine (Gargouri et al, 2003), nous croyons que ce processus ne peut se limiter à la seule analyse statistique (ni même linguistique) pour exprimer la richesse sémantique du langage naturel. En effet, les connaissances implicites du domaine ne peuvent exclusivement être repérées via une analyse du corpus. Par conséquent, nous supportons l'idée que de telles connaissances peuvent en partie, être accédées avec l'utilisation de ressources terminologiques complémentaires telles que les thésaurus (comme Wordnet). Il est toutefois irréaliste de prétendre qu'un thésaurus puisse englober toutes connaissances d'un domaine d'expertise. Les thésaurus sont spécialement utiles pour offrir des réseaux lexicaux et de l'information additionnelle reliée à la signification de termes (utilisation, définition, synonymie, etc.).

Il est important de noter que les ressources terminologiques représentent davantage des outils linguistiques reflétant la langue plutôt que le domaine. Par conséquent, les connaissances implicites du domaine que l'approche du thésaurus tente de repérer, sont spécialement de nature linguistique.

La combinaison des deux techniques (la classification et l'ISL d'une part, et les vecteurs conceptuels et les thésaurus d'autre part) constitue un choix méthodologique pertinent pour découvrir des relations entre termes. Ces deux techniques sont en effet complémentaires (tel que nous le verrons dans le **chapitre 4** (Implémentation et expérimentation)).

Les concepteurs d'ontologies ont fait usage dans certains projets, de dictionnaires électroniques pour enrichir leurs ontologies (Hearst, 1992 ; Mitra et al, 1999). La performance de ces travaux est principalement attribuée à des analyses sémantiques et linguistiques pour extraire de nouveaux termes et relations entre termes à partir de dictionnaires. Nous avons privilégié l'approche numérique des vecteurs conceptuels en orientant le choix de la composition de ces vecteurs vers des relations types caractéristiques des ontologies (synonymie, hyperonymie et antonymie).

Un Vecteur Conceptuel est construit à partir des items lexicaux associés au terme en question, selon des liens de synonymie, d'hyperonymie ou d'antonymie, extraits à partir du thésaurus. La technique proposée est en fait fondée sur un mariage entre, d'une part, l'approche des réseaux sémantiques et des thésaurus, associée au domaine de la

représentation des connaissances, et d'autre part, *l'approche vectorielle* issue des « *représentations saltoniennes* » (Salton, 1968) et de la recherche d'information. Il s'agit en d'autres termes, de représenter les termes par des vecteurs conceptuels.

Le modèle vectoriel n'est pas récent, puisqu'il a été introduit par (Salton, 1968) en informatique documentaire. Ce modèle est basé sur la représentation des significations de concepts par un espace vectoriel. Sa réhabilitation dans les recherches en TALN a été essentiellement motivée par la mise à disposition des chercheurs de grandes bases de textes grâce au Web en particulier, alors que précédemment, ces recherches passaient par des phases ardues de constitution de corpus d'expérience (Jalabert et al, 2004).

Les vecteurs conceptuels s'inspirent également d'une façon ultime de l'Analyse Formelle de Concepts (AFC), détaillée dans la **section 2.5.3.2**. Le modèle de vecteurs conceptuels s'appuie sur la projection dans un modèle mathématique de la notion linguistique de champ sémantique. Un vecteur correspond donc à une combinaison linéaire d'autres idées, termes, sens, sous forme de réseau et d'interdépendances. Ce vecteur, bien qu'il représente des mots extraits d'un thésaurus sert ultimement à une interprétation conceptuelle. Cette représentation homogène des interdépendances, quelle que soit la granularité, est très avantageuse pour la classification des textes, l'indexation, la recherche évoluée d'information et spécialement pour notre cas, la recherche de relations sémantiques entre les termes.

La façon choisie de représentation du contenu des vecteurs conceptuels dépend des fins pour lesquelles ces vecteurs sont utilisés. L'approche des vecteurs conceptuels que nous avons adoptée s'inspire du modèle vectoriel de Salton, mais elle en diffère en ce que nous faisons l'hypothèse qu'il existe un jeu de concepts prédéterminé par les lexicologues quand ils réalisent un thésaurus. Par exemple, pour découvrir des relations de synonymie ou d'antonymie, les vecteurs conceptuels peuvent être construits à partir du lexique représentant le sens de chacun des termes. Ainsi, ce genre de relations est découvert en calculant la distance sémantique entre les couples de vecteurs (Hearst, 1992 ; Schwab et al, 2002). Le rôle du sens joue dans ce cas, un rôle certes important, mais insuffisant pour découvrir des relations de synonymie ou d'antonymie (Jalabert et al, 2004).

Avant de détailler notre choix de représentation du contenu des vecteurs conceptuels, nous discutons dans ce qui suit, du choix de Wordnet comme exemple de thésaurus servant de source terminologique pour nos Vecteurs Conceptuels.

3.2.6.1 Wordnet comme base de données lexicale

En tant que base de données lexicale, Wordnet (Fellbaum, 1998 ; Miller et al, 1993) est considérée comme une des ressources les plus importantes qui soit disponible pour les chercheurs de la linguistique computationnelle, l'analyse textuelle ainsi que d'autres domaines reliés. Son architecture est inspirée des théories psycholinguistique et computationnelle de la mémoire lexicale de l'humain. Cette base de données lexicale consiste en plusieurs réseaux de lexèmes où chaque nœud correspond à un concept et est représenté par un ensemble de mots, constituant ainsi un « *synset* ». Elle fournit une brève définition (*gloss*) de chacun des termes. Les *synsets* sont définis selon quatre principales catégories syntaxiques (nom, verbe, adjectif et adverbe).

L'application Wordnet, disponible en ligne, présente les différentes définitions d'un terme, mais aussi les termes qui lui sont sémantiquement associés selon certaines relations spécifiques. Ainsi, il est possible d'obtenir les différents termes partageant l'une ou l'autre des relations sémantiques avec le terme initial. Plusieurs relations générant les réseaux sémantiques peuvent être explorées. Parmi celles-ci, on retrouve entre autres, les relations de synonymie, hyperonymie, hyponymie, holonymie, méronymie, antonymie, causalité, etc.

Nous rappelons dans ce qui suit les définitions de ces relations :

- Un *synonyme* est un mot ou expression dont le sens est identique ou semblable à celui d'un autre mot ou d'une autre expression. (Par exemple, *heureux* et *content*)
- L'*hyperonyme* qualifie un terme dont le sens inclut le sens (ou les sens) d'un ou de plusieurs autres termes, appelés alors hyponymes. (Par exemple, *animal* est un hyperonyme de *cheval*). L'*hyponyme*, par opposition à l'hyperonyme, est un terme plus spécifique. (Par exemple, *fer* est un hyponyme de *métal*).

- L'*holonyme* est un terme lié à un autre terme par une relation sémantique de tout à partie. Par contre, le *méronyme* est un terme qui fait partie d'un tout. Par exemple, le *bras* (méronyme) est une partie du *corps* (holonyme).
- Un *antonyme* est un mot dont le sens est opposé à un autre. (Par exemple, *chaud* est un antonyme de *froid*).

Dans plusieurs travaux, Wordnet est utilisée comme base lexicale pour évaluer l'association sémantique entre des couples de termes. Parmi les mesures les plus fréquemment citées, on note celles de (Banerjee et al, 2002 ; Hirst et al, 1998 ; Leacock et al, 1998 ; Lin, 1998 ; Jiang et al, 1997; Resnik, 1995). Toutes ces mesures permettent de calculer l'association sémantique entre des couples de termes par une valeur numérique.

(Leacock et al, 1998) se basent sur la longueur du chemin entre deux concepts c1 et c2 à travers la hiérarchie *est-un* (*is-a*) de Wordnet. La longueur du chemin est ainsi mesurée par la profondeur de la hiérarchie dans laquelle les termes existent pour obtenir leur degré d'association sémantique.

(Resnik, 1995) a introduit une mesure basée sur le *contenu informationnel*, qui est une valeur indiquant la spécificité du terme. Ces valeurs sont extraites à partir du corpus et sont utilisées pour enrichir les termes dans la hiérarchie *is-a* de Wordnet. La mesure d'association sémantique entre deux termes est le contenu informationnel que le terme le plus spécifique partage avec l'autre terme. Jiang et ses collaborateurs ont développé davantage la mesure de Resnik (Resnik, 1995) pour combiner les contenus informationnels des deux termes ainsi que leur terme commun le plus spécifique (Jiang et al, 1997). Lin a également étendu la mesure de Resnik en considérant le ratio du contenu informationnel partagé par rapport à ceux des termes individuels (Lin, 1998).

(Banerjee et al, 2003) ont introduit la notion de «*Extended Gloss Overlaps*», qui est une mesure déterminant le lien de parenté sémantique entre deux termes en calculant le nombre de termes communs dans chacune de leurs définitions. Les définitions extraites de Wordnet sont enrichies par celles des termes qui sont directement liés par une relation (*synset*) de Wordnet. (Patwardhan et al, 2006) ont également utilisé cette même technique.

3.2.6.2 Contenu des Vecteurs Conceptuels

Les relations caractéristiques des ontologies sont principalement celles d'hyperonymie et d'hyponymie, mais aussi de synonymie, antonymie, holonymie, méronymie. Nous avons ainsi, envisagé ce genre de relations pour construire nos vecteurs conceptuels plutôt que de se limiter au sens de chaque terme. Nous nous sommes toutefois limités aux relations de synonymie, d'hyperonymie et d'antonymie pour des raisons que nous détaillons plus loin.

La représentation des termes sous forme de vecteurs conceptuels a été réalisée de la façon suivante :

Pour chaque couple de termes (T_1, T_2), deux vecteurs conceptuels VC_{T1} et VC_{T2} sont construits à partir des synonymes (S_{T1} et S_{T2}), hyperonymes (H_{T1} et H_{T2}) et antonymes (A_{T1} et A_{T2}) de chacun des termes T_1 et T_2. Ainsi :

VC_{T1} : $[S_{T1}, H_{T1}, A_{T1}]$
VC_{T2} : $[S_{T2}, H_{T2}, A_{T2}]$

Par exemple, les vecteurs conceptuels du couple (Wireless, Network) sont composés des lexiques suivants :

	Synonymes	**Hyperonymes**	**Antonymes**
Wireless	*radio, radiocommunication, radio_receiver, receiving_set, radio_set, tuner*	*broadcasting, telecommunication, telecom, medium, means,* ***instrumentality, instrumentation, artifact, artefact, object, physical_object, entity, whole, whole_thing, unit,*** *receiver, receiving_system, set, electronic_equipment, equipment*	Ø
Network	*web, net, mesh, meshing, meshwork, electronic_network*	*system, scheme, group, grouping, communication_system, facility, installation, communication_equipment,* ***artifact, artefact, object, physical_object, entity, whole, whole_thing, unit,*** *fabric, cloth, material, textile,* ***instrumentality, instrumentation***	Ø

Tableau 3.2 : Composition des vecteurs conceptuels de (Wireless, Network)

Ainsi, les VC respectifs sont les suivants :

- **Wireless**: [*radio, radiocommunication, radio_receiver, receiving_set, radio_set, tuner, broadcasting, telecommunication, telecom, medium, means,* ***instrumentality, instrumentation, artifact, artefact, object, physical_object, entity, whole, whole_thing, unit****, receiver, receiving_system, set, electronic_equipment, equipment*]
- **Network**: [*web, net, mesh, meshing, meshwork, electronic_network, system, scheme, group, grouping, communication_system, communication_equipment, facility, installation,* ***artifact, artefact, object, physical_object, entity, whole, whole_thing, unit,*** *fabric, cloth, material, textile,* ***instrumentality, instrumentation***]

Ensuite, un vecteur $VC_{T1,T2}$ est construit à partir de l'union des lexiques des vecteurs VC_{T1} et VC_{T2} :

$$VC_{T1,T2} : [S_{T1}, H_{T1}, A_{T1}, S_{T2}, H_{T2}, A_{T2}]$$

Par exemple : Vecteur Union (*Wireless, Network*): [*radio, radiocommunication, radio_receiver, receiving_set, radio_set, tuner, broadcasting, telecommunication, telecom, medium, means, instrumentality, instrumentation, artifact, artefact, object, physical_object, entity, whole, whole_thing, unit, receiver, receiving_system, set, electronic_equipment, equipment, network, web, net, mesh, meshing, meshwork, electronic_network, system, scheme, group, grouping, communication_system, communication_equipment, facility, installation, fabric, cloth, material, textile*].

Le terme T_1 (respectivement T_2) est mathématiquement représenté par un vecteur $VC_{T1,T2}$ (respectivement $VC_{T2,T1}$), composé d'une suite de 0 et 1 dépendamment de l'existence ou non des composants de $VC_{T1,T2}$ (respectivement $VC_{T2,T1}$) parmi les synonymes, hyperonymes et antonymes du terme T_1 (respectivement T_2).

En pratique, plus $VC_{T1,T2}$ est grand, plus fine sera la description de termes reliés, offerte par le vecteur, mais plus sa manipulation informatique peut être lourde, spécialement si l'on traite un grand volume de données. Ainsi, d'autres relations entre termes comme l'holonymie, la méronymie, et l'hyponymie n'ont pas été considérées dans nos vecteurs conceptuels. En effet, bien que ces relations puissent enrichir ces vecteurs, elles alourdissent leur traitement

informatique. De plus, dans la mesure où une relation sémantique entre deux termes T_1 et T_2 sera découverte en se basant sur la coexistence d'un ensemble de termes communs à l'intérieur de $V_{T1,T2}$ et $V_{T2,T1}$, le nombre de termes communs serait relativement « dilué » si les vecteurs conceptuels sont de grande taille.

Les vecteurs conceptuels sont généralement utilisés avec la mesure du cosinus de l'angle entre les deux vecteurs pour prendre des décisions par rapport à la qualité d'association entre les termes.

3.2.6.3 Mesures d'association sémantique

Dans la mesure où les termes sont représentés par des vecteurs conceptuels, les mesures d'association sémantique sont mieux conceptualisées à l'aide de mesures de similarité vectorielle. Plusieurs mesures peuvent être considérées. Nous commentons dans ce qui suit les mesures suivantes :

- Coefficient d'apariement $= VC_{T1,T2} \cap VC_{T2,T1}$

Cette mesure compte simplement le nombre d'entrées non nulles dans chacun des vecteurs. Contrairement aux autres mesures, elle ne tient pas compte de la taille des vecteurs et du nombre total des entrées non nulles dans chacun.

- Cœfficient de « Dice » $= \frac{2 | VC_{T1,T2} \cap VC_{T2,T1} |}{| VC_{T1,T2} | + | VC_{T2,T1} |}$

Ce coefficient normalise la taille des vecteurs en divisant par le nombre total des entrées non nulles. Cette mesure est multipliée par 2 pour avoir une valeur comprise entre 0 et 1. La valeur « 1 » indique que les vecteurs sont identiques.

- Coefficient de Jacquard (ou Tanimoto) $= \frac{| VC_{T1,T2} \cap VC_{T2,T1} |}{| VC_{T1,T2} \cup VC_{T2,T1} |}$

Cette mesure pénalise davantage que le cœfficient de « Dice », un petit nombre d'entrées communes par rapport à la proportion de toutes les entrées non nulles.

- Coefficient de chevauchement (Overlap) $= \frac{| VC_{T1,T2} \cap VC_{T2,T1} |}{\min(| VC_{T1,T2} |, | VC_{T2,T1} |)}$

Cette mesure est caractérisée par la spécificité de l'inclusion. La valeur de cette mesure est de « 1 » si toute entrée non nulle du premier vecteur est aussi non nulle pour le second vecteur et vice versa (en d'autres termes $VC_{T1,T2} \subseteq VC_{T2,T1}$ ou $VC_{T2,T1} \subseteq VC_{T1,T2}$).

- Cosinus : $\cos(VC_{T1,T2}, VC_{T2,T1}) = \frac{| VC_{T1,T2} \cap VC_{T2,T1} |}{\sqrt{| VC_{T1,T2} \times VC_{T2,T1} |}}$

Le cosinus est identique au coefficient de « Dice » pour les vecteurs ayant le même nombre d'entrées non nulles, mais il pénalise moins les cas où le nombre des entrées non nulles est très différent. Par exemple, si on compare un vecteur ayant une seule entrée non nulle avec un autre ayant 1000 entrées non nulles et une seule entrée commune, alors le coefficient de « Dice » est de 2*1/(1+1000) = 0.002 et le cos est de $1/\sqrt{1000*1} = 0.03$

Cette propriété du cosinus est importante pour notre cas dans la mesure où nous comparons souvent des vecteurs conceptuels de dimensions différentes.

De plus, le repérage de proximité sémantique entre vecteurs conceptuels se base principalement sur la présence d'entrées communes dans les deux vecteurs. Cette propriété est spécifiquement bien représentée par la formule du Cosinus. La valeur de cette mesure est comprise entre 0 et 1. La valeur 0 correspond à deux vecteurs orthogonaux, et la valeur 1 correspond à deux vecteurs identiques.

Nous avons ainsi opté pour cette dernière mesure comme coefficient pour repérer la proximité sémantique entre les termes. Soit $RS(T_1, T_2)$ la mesure de la relation sémantique entre les termes T_1 et T_2.

$$RS(V_{T1,T2}, V_{T2,T1}) = \cos(V_{T1,T2}, V_{T2,T1}) = \frac{V_{T1,T2} \cdot V_{T2,T1}}{|V_{T1,T2}| \times |V_{T2,T1}|} = \frac{\sum_{i=1}^{n} V_{T1,T2_i} \times V_{T2,T1_i}}{\sqrt{\sum_{i=1}^{n} V_{T1,T2_i}^{2}} \times \sqrt{\sum_{i=1}^{n} V_{T2,T1_i}^{2}}}$$

Par exemple :

$$\text{Cos (Wireless, Network)} = \frac{10}{26 \times 28} = 0{,}013$$

Ces deux vecteurs conceptuels partagent en effet, 10 entrées lexicales communes (*instrumentality, instrumentation, artifact, artefact, object, physical_object, entity, whole, whole_thing, unit*) sur un total de 45. La valeur du cosinus (0,013) confirme effectivement la relation sémantique entre *Wireless* et *Network*. Une valeur supérieure à un seuil zéro permet de constater une relation potentielle. Ce seuil peut être établi à un niveau plus élevé si le type de relations visées est relativement fort.

En utilisant cette mesure de relation sémantique entre les vecteurs conceptuels, nous identifions les couples de termes présentant une forte probabilité d'être sémantiquement associés. Toutes ces relations entre termes doivent être vérifiées et analysées par un expert pour confirmer leur pertinence par rapport à l'ontologie courante.

Les itérations relatives à l'application de l'ISL et des VC sont appliquées pour chacune des classes de termes identifiées lors de l'étape de la classification textuelle.

3.2.7 Mise à jour de l'ontologie

Finalement, les nouveaux termes, ne figurant pas dans l'ontologie courante, ainsi que leurs relations, sont intégrés dans l'ontologie de domaine. Tel que défini dans les objectifs de la thèse, nous nous limitons dans notre modèle au processus de conceptualisation de l'ontologie. L'ontologisation et l'opérationnalisation ne seront pas traitées. Nous présentons toutefois, une réflexion sur les conséquences des changements touchant les ontologies. Lors d'un

processus de mise-à-jour, il est primordial de focaliser spécifiquement sur la consistance continue, tant de l'ontologie que des systèmes utilisant cette ontologie.

Dans l'objectif de faciliter l'intégration des nouveaux termes dans l'ontologie, il est important d'avoir une vue d'ensemble plus claire sur les termes figurant dans l'ontologie courante, en vue de mieux visualiser les interdépendances des anciens termes avec les nouveaux. Le processus de maintenance doit en effet, évoluer en fonction de la situation actuelle de l'ontologie. En d'autres termes, un outil d'assistance à la maintenance doit nécessairement prévoir une visualisation superposée, intégrant tant les termes et les relations de l'ontologie courante que ceux proposés comme candidats pour la maintenance.

La variété des causes et des conséquences de l'évolution de l'ontologie fait de sa maintenance un processus particulièrement complexe. Les changements élémentaires susceptibles d'être apportés à une ontologie consistent en l'ajout, la modification ou la suppression d'éléments ; à savoir un terme, une relation entre termes, un axiome, l'étendu d'une propriété, etc.

Les ontologies sont construites pour des usages multiples. Outre la prise en compte du contexte spécifique à l'application, son environnement, ses propres contraintes et son champ d'application, l'implication de l'utilisateur dans le processus d'évolution permet de garantir une ontologie plus adaptée aux besoins des utilisateurs. Par ailleurs, pour les applications à utilisation étendue, telles que celles du Web sémantique, les ontologies sont construites et maintenues dans des environnements dynamiques et distribués. Par conséquent, l'environnement de maintenance doit supporter la collaboration entre plusieurs utilisateurs distribués dans l'espace[26].

Une stratégie de gestion de l'évolution d'une ontologie doit prévoir un mécanisme d'inférence cohérent régissant l'ensemble des termes, des relations entre termes et des axiomes en vue de formaliser cette évolution. Ce mécanisme assiste l'expert chargé de la maintenance, en proposant l'application de certaines règles, dont nous en citons quelques unes ci-après. La conformité automatique des changements ontologiques avec ces règles

[26] Cette dimension est supportée par une architecture de type client/serveur

garantit la cohérence de l'ontologie, sans se préoccuper d'une vérification explicite (Franconi et al, 2000).

Nous présentons ci-après, les principales opérations de changements qui peuvent être apportées à une ontologie et discutons de leurs effets, spécialement, sur les instances de classes annotées suivant l'ancienne ontologie[27]. Nous rappelons ici qu'une *classe*[28] réfère à un concept ou un terme du domaine. Elle est définie par des propriétés, formulées sous forme de restrictions, c'est-à-dire des formules de la logique de description. Les classes sont structurées de façon hiérarchique, de sorte qu'une classe hérite les propriétés de sa super-classe.

Les effets des changements sur les instances de classes peuvent donc être résumés comme suit :

- La suppression d'une classe C déplace ses instances vers la super-classe. Les instances deviennent moins spécifiques.
- La création d'une nouvelle classe C n'a généralement pas d'effet particulier sur les données de l'ontologie, sauf si des instances existantes sont des instances de C.
- L'ajout d'une propriété P à la classe C n'affecte pas les données.
- La suppression d'une propriété P de la classe C nécessite la suppression des valeurs de cette propriété pour toutes les instances.
- Le rattachement d'une propriété P à une classe C ne détruit pas les données.
- L'ajout d'un lien de spécialisation entre une sous-classe SousC et une super-classe SuperC conduit à la création de nouvelles propriétés au niveau de SousC, héritées de SuperC. Par contre, la suppression de ce lien entraîne celle des propriétés héritées de SuperC et la suppression des valeurs de ces propriétés pour les instances de SousC.
- La transformation d'une instance I en une classe C n'affecte pas les données. Une transformation inverse rend les instances de C moins spécifiquement typées.

---

[28] Dans le domaine des ontologies

- La déclaration de deux classes C1 et C2 comme disjointes rend leurs instances invalides.
- La déclaration d'une propriété P comme transitive ou symétrique pourrait remettre en cause certaines des valeurs de P ne respectant pas ces deux caractéristiques.
- Le déplacement d'une propriété P de la sous-classe SousC à la super-classe SuperC n'affecte pas le lien d'héritage de P dans SousC et les instances préservent leurs valeurs de P. Par contre pour un déplacement inverse, les valeurs de P pour les instances de SuperC sont détruites.
- Le déplacement d'une super-classe d'une classe C à un niveau plus élevé (c'est-à-dire plus général) dans la hiérarchie conduit à une situation où C possède encore des propriétés qui héritent directement de cette super-classe. Par conséquent, les valeurs des propriétés, pour les instances de C, sont perdues. Par contre, le déplacement d'une super-classe d'une classe C à un niveau plus bas dans la hiérarchie (c'est-à-dire plus spécifique), pourrait conduire à l'enrichissement de C par des propriétés additionnelles héritées. Aucune des données n'est perdue.
- Les valeurs d'une propriété P restent valides si une restriction[29] relative à P se trouve élargie. Par exemple : augmenter le nombre de valeurs possibles, diminuer le nombre de valeurs exigées, rajouter une classe à l'étendue ou remplacer une classe existante dans l'étendue par sa super-classe, etc. Par contre, si la restriction se trouve plus contraignante (moins élargie), les valeurs de P, ne respectant pas ces restrictions, deviennent invalides. Par exemple : diminuer le nombre de valeurs possibles, augmenter le nombre de valeurs exigées, supprimer une classe de l'étendue ou la remplacer par une sous-classe, etc.
- La fusion de classes (super-classes et sous-classes) n'affecte pas les valeurs des propriétés si ces dernières sont transférées vers la classe de fusion. Toutefois, la règle du rajout d'un lien de sous-classe doit être respectée.
- La division d'une classe en différentes classes n'entraîne pas une perte de données.

[29] Une restriction est une condition que la valeur de la propriété en question doit satisfaire.

3.2.8 Conclusion

Nous avons présenté un modèle et des outils d'assistance au développement et à la maintenance des ontologies en se basant sur une approche semi-automatique, qui a l'avantage d'être indépendante de la langue et est applicable à de larges données textuelles. Notre contribution vise en fait, à remédier aux lacunes méthodologiques et au manque d'outils, intégrant la fonctionnalité de maintenance dans le domaine de l'ingénierie des ontologies.

CHAPITRE IV

IMPLÉMENTATION ET EXPÉRIMENTATION

Dans l'objectif d'évaluer le modèle proposé, une expérimentation a été réalisée en vue de valider l'approche proposée. Le modèle a été en partie implémenté puis testé sur un corpus. Le domaine d'expérimentation est celui des télécommunications sans fil.

Nous présentons dans ce chapitre les logiciels utilisés. Nous expliquons par la suite, comment le corpus a été constitué. La troisième section traite du processus d'expérimentation et de l'évaluation des résultats. Enfin, la dernière section détaille la validation des résultats à la lumière des hypothèses de recherche préalablement établies.

4.1 Logiciels utilisés

L'expérimentation a été accomplie sur le corpus constitué en utilisant des modules de la plateforme SATIM ; Gramexco et ONTOLOGICO. Le premier module est une chaîne de traitements déjà existante. Nous avons implémenté le deuxième module pour les besoins de cette thèse.

4.1.2 Plateforme SATIM

La plateforme SATIM (Système d'ANalyse et de Traitement de l'Information Multidimensionnelle) (Biskri et al, 2002) est un outil informatique générique qui est le résultat de plusieurs années de recherche portant sur le traitement de l'information. Elle fût développée au Laboratoire d'ANalyse de l'Information (LANCI), sous la direction de Jean-Guy Meunier. La version actuelle fût programmée sous la direction d'Ismail Biskri de l'Université du Québec à Trois-Rivières.

En tant que système de traitement d'information, SATIM permet l'exploration et l'expérimentation de différents types d'analyses grâce à sa modularité, sa flexibilité, ses diverses fonctions d'analyse et sa capacité d'adaptation par rapport à la croissance des

données textuelles. Cet outil permet, par l'organisation et l'interaction entre différents modules indépendants, la construction de différentes chaînes de traitements pouvant effectuer diverses tâches liées au traitement de l'information en fonction de la nature des données à analyser (texte, image, son ...). Les chaînes de traitements construites à partir de SATIM sont destinées entre autres à la classification (Meunier et al, 2005), la production de résumés automatiques, d'index, de thésaurus, d'ontologies (Gargouri et al, 2003), de liens hypertextes (Nault, 2001), d'analyse lexicale, d'extraction des connaissances et d'analyses thématiques (Forest et al, 2001).

L'organisation des modules dans une chaîne de traitement doit respecter l'homogénéité des données communiquées entre les modules. Ainsi, les outputs des modules sont dans un format compatible avec celui des intrants des modules suivants. Il relève de l'utilisateur d'agencer de manière optimale les différents modules qu'il aura choisis en fonction de l'objectif souhaité. Une fois confirmée et validée par l'utilisateur, la chaîne de traitements pourra être figée afin d'en faire un agent autonome réutilisable, tel que Numexco et Gramexco qui sont destinées au traitement des données textuelles.

4.1.3 Chaîne de traitement ONTOLOGICO

Le modèle proposé ONTOLOGICO (détaillé dans la **section 3.2**) a été en partie implémenté sous forme d'une chaîne de traitements au sein de la plate-forme SATIM. Cette chaîne de traitements utilise celle de Gramexco qui sert à la classification. Ainsi, ONTOLOGICO procède à l'extraction de termes reliés en traitant le lexique généré par Gramexco.

Dans l'objectif d'évaluer le modèle proposé dans cette thèse, deux modules ont été implémentés au niveau de la chaîne de traitement ONTOLOGICO, à savoir le module **ISL**, associé avec la DVS et le module **Vecteurs Conceptuels**. L'**annexe 5** présente l'algorithme DVS utilisé. Le langage de programmation choisi est Java. Ce choix est justifié tout simplement par des objectifs de portabilité, de continuité et d'extensibilité future, au niveau de la plateforme SATIM compte tenu de l'aspect modulaire qu'offrent les langages Orientés Objets. Cette modularité constitue un choix pratique particulièrement puissant, permettant entre autres, la possibilité de réutilisation de certains de ces modules pour d'autres objectifs

d'analyses textuelles, ainsi que l'intégration de nouveaux modules répondant à des fonctionnalités complémentaires.

L'architecture de la chaîne de traitements ONTOLOGICO (**Figure 4.1**) est détaillée comme suit :

4.1.3.1 Gramexco

La chaîne de traitements Gramexco vise à procéder à la classification textuelle. Dans une première étape, un extracteur de termes génère le lexique du corpus et les n-grams correspondant à ce lexique. Ensuite, un lemmatiseur est appliqué pour remplacer les mots par leurs lemmes correspondants. L'utilisateur procède par la suite, au filtrage du lexique. Le corpus est ensuite segmenté selon une taille choisie par l'utilisateur. En se basant sur le lexique retenu (UNIFs) et les segments de documents (DOMIFs), un classifieur ART1 génère des classes de segments.

Le processus de classification génère finalement des classes de termes potentiellement reliés, construits à partir de l'intersection des termes qui cooccurrent ensemble à travers chacune des classes de segments. Ce sont en fait, ces termes en particulier qui ont fait en sorte que les segments de documents se regroupent dans une même classe. Le résultat de la classification est finalement stocké dans une base de données Access (*classes.mdb*).

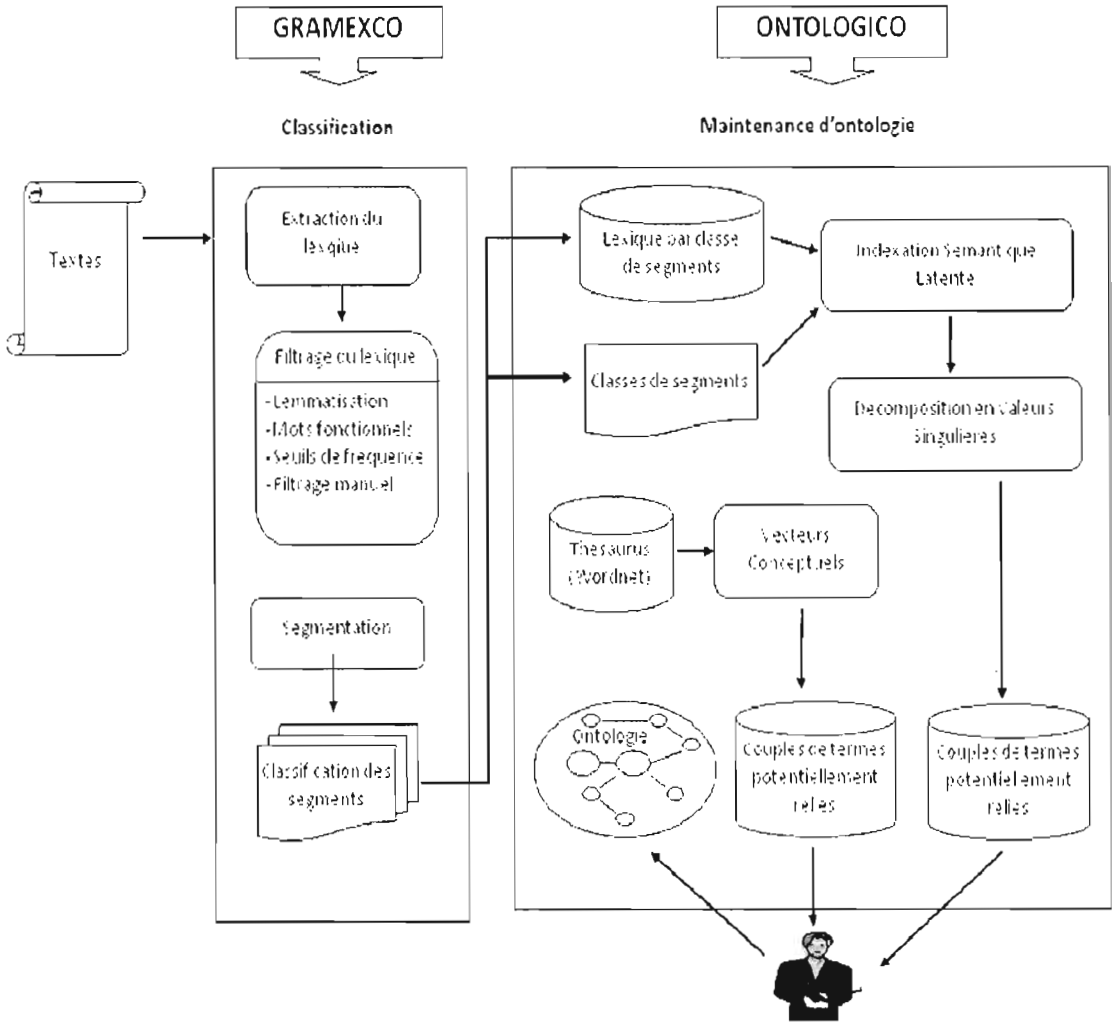


Figure 4.1 : Architecture de la chaîne de traitements ONTOLOGICO

4.1.3.2 Modules ISL et Vecteurs Conceptuels

Le module ISL génère, à partir de chaque classe de termes, les couples de termes qui sont potentiellement reliés en se basant sur l'Indexation Sémantique Latente ainsi que sur la Décomposition en Valeurs singulières (tel que détaillé dans la **section 3.2.3**). Le module Vecteurs Conceptuels génère les vecteurs conceptuels de chaque couple de termes et calcule la valeur du cosinus de l'angle entre ces vecteurs.

L'interface graphique, illustrée par la figure (**Figure 4.2**), permet à l'utilisateur de réaliser les fonctionnalités suivantes :

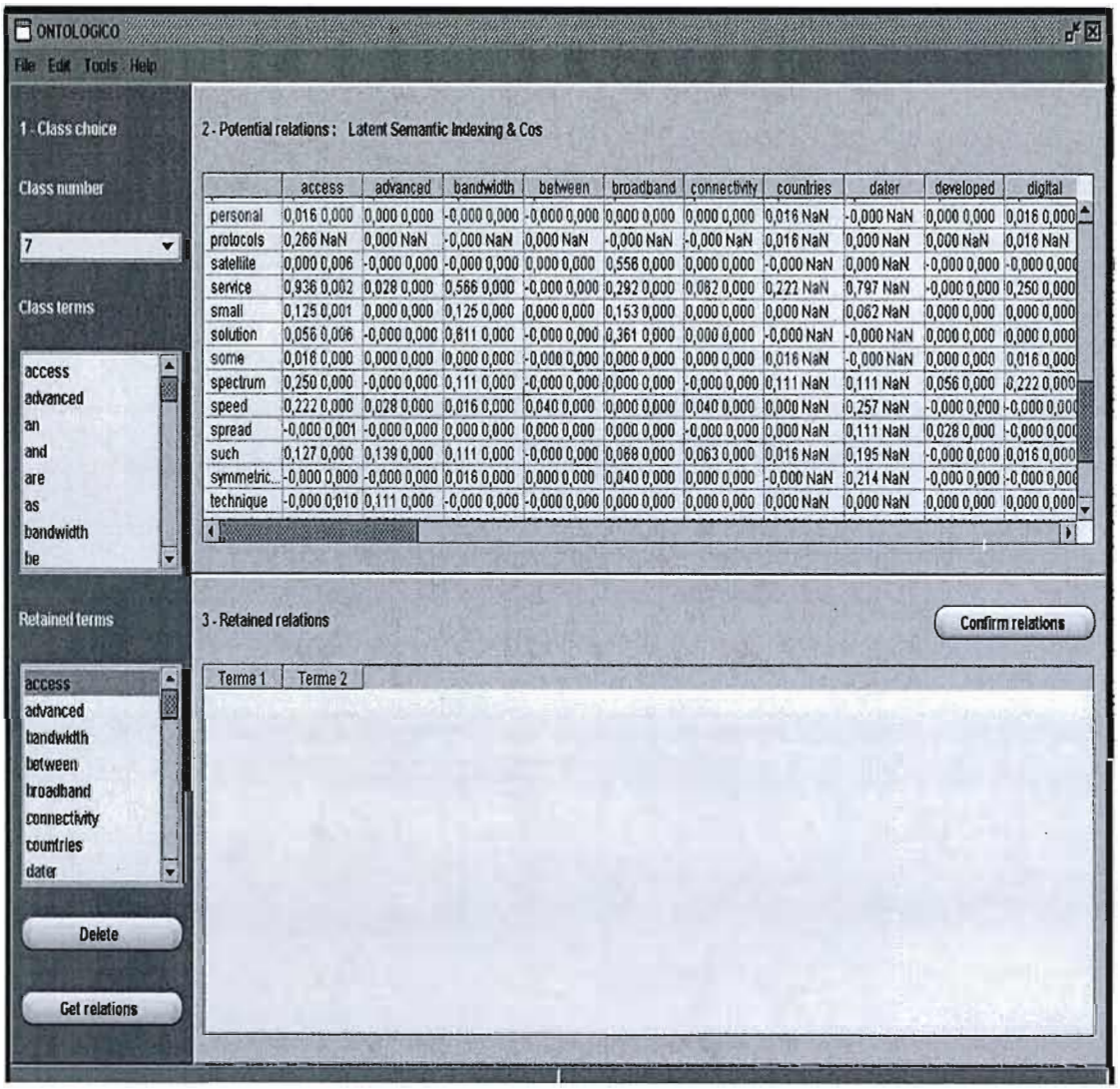


Figure 4.2 : Interface graphique d'ONTOLOGICO

- **Connexion à la base de données** : la première étape consiste à ouvrir une connexion à la base de données *classes.mdb*, en choisissant le menu *File* et le sous-menu *OpenDatabase*.

- **Chargement des classes** : la liste de classes de termes, générées par *Gramexco*, est automatiquement chargée. Ces classes sont identifiées par des numéros

d'identification. En sélectionnant un numéro, la liste de termes de la classe correspondante est affichée.

- **Filtrage de termes** : cette fonctionnalité permet à l'utilisateur de filtrer les termes d'une classe donnée, en éliminant ceux considérés comme bruit, par rapport à l'ontologie de domaine. L'utilisateur sélectionne les termes à filtrer (en utilisant la touche *Ctrl*) puis clique sur le bouton *Delete*. La liste de termes retenus est affichée plus bas.

- **Génération des poids d'association**: en cliquant sur *Get relations*, les modules *ISL* et *Vecteurs conceptuels* sont exécutés. Le premier module génère des poids d'association entre des couples de termes. Ces valeurs sont affichées dans un tableau, tel qu'illustré à la figure (**Figure 4.2**). Chaque cellule de ce tableau affiche deux valeurs ; celle de gauche représente le poids d'association, selon l'ISL, d'un couple de termes et celle de droite indique la valeur du cosinus de l'angle entre les vecteurs conceptuels relatifs à ce même couple de termes.

  Les résultats générés peuvent être classés dans 2 catégories de relations entre termes : la **catégorie A**, regroupant les relations très fortes ; il s'agit des relations qui sont confirmées tant par l'ISL que par les Vecteurs Conceptuels. Les valeurs correspondant à ces relations sont en rouge. La **catégorie B**, regroupant des relations fortes qui sont repérées exclusivement par l'une ou l'autre des deux techniques (ISL ou vecteurs conceptuels) ; les valeurs correspondant à ces relations sont en bleu.

- **Confirmation des relations** : sur la base de ces deux valeurs données, l'utilisateur confirme la relation pour un couple de termes donnés, en double cliquant sur la cellule correspondante. Ce couple de terme est par conséquent, rajouté à la liste des relations retenues (au dessous).

- **Chargement des relations dans la base de données** : une fois la liste des relations retenues achevée, l'utilisateur clique sur le bouton « Confirm relations ». La liste est enfin chargée dans la base de données.

4.2 Constitution de corpus

Le corpus choisi est une collection de documents techniques relatifs au domaine des télécommunications sans fil[30]. Il contient 6985 lignes et 840 000 mots. Le corpus est en anglais. Le choix de la langue est tout simplement motivé par la disponibilité en ligne du thésaurus Wordnet en anglais.

Les textes, objets de notre expérimentation, sont sélectionnés à partir d'une documentation technique relative au domaine des télécommunications sans fil. Cette sélection est élaborée en faisant recours à des experts du domaine. En effet, un domaine n'est pas seulement défini par le champ de connaissances qu'il couvre, mais aussi par le point de vue sous lequel les utilisateurs de l'ontologie considèrent ce champ de connaissances (Uschold et al, 1995). Ainsi, le corpus constitué pour construire une ontologie des connaissances médicales ne sera pas le même si le public visé est constitué de médecins ou si le public visé n'a pas de connaissances particulières dans ce domaine, en particulier en ce qui concerne la granularité du corpus et des connaissances modélisées.

Les textes collectés proviennent principalement du Web. Bien que le Web soit une source considérable de textes relatifs à un domaine précis, il n'est toutefois pas dépourvu de défauts. Par exemple, il est fréquent de trouver des phrases qui apparaissent de manière récurrente dans plusieurs pages d'un même site. On trouve aussi des phrases, provenant de citations ou de dépêches journalistiques, qui se retrouvent sur plusieurs sites différents, parfois telles quelles, parfois légèrement reformulées. Ces phrases récurrentes faussent les statistiques de fréquence utilisées pour la classification ainsi que l'ISL. Il a donc fallu identifier et éliminer automatiquement les phrases identiques ou trop similaires.

4.3 Expérimentation et évaluation des résultats

L'expérimentation a été réalisée en appliquant le modèle ONTOLOGICO et les outils associés sur le corpus sélectionné. Ce dernier a été segmenté sous forme de paragraphes et soumis à un processus de classification en utilisant GRAMEXCO. Ce processus a généré 140 classes de segments.

---

[30] Le corpus a été constitué principalement de documents collectés sur Internet.

Ensuite, nous avons extrait pour chaque classe de segments, le lexique représentant l'intersection des termes appartenant aux segments de cette classe. Cette étape génère ainsi des classes de termes qui sont potentiellement reliés.

Le tableau suivant (**Tableau 4.1**) illustre un extrait de quelques classes de termes obtenues suite à ce processus de classification. L'**annexe 3** illustre d'autres exemples de classes de termes générées par Gramexco. La classe numéro 19 par exemple, renferme les termes suivants : *access, broadband, dater, inc, integrated, international, local, market, PC, room, wireless*. La cooccurrence de ces termes à travers le corpus laisse entendre que les segments de documents, renfermant ces termes, parlent des mêmes sujets ou du moins de sujets associés.

Cette cooccurrence explique d'ailleurs des relations sémantiques entre des couples de termes, que nous sommes déjà en mesure de repérer. Par exemple : (*access, wireless*), (*local, market*), (*international, local*)...

L'ensemble de ces classes de termes est ensuite traité par ONTOLOGICO. L'**annexe 4** montre un exemple de ce qu'ONTOLOGICO génère comme résultats pour la classe numéro 7. Dans cette annexe, les tableaux présentent les valeurs de l'ISL (qui se trouvent dans la partie gauche de chaque cellule) ainsi que le cosinus (dans la partie droite) de chaque couple de termes de cette classe.

Reprenons l'exemple du couple (Wireless, Network). Nous obtenons la valeur de cosinus suivante :

$$\text{Cos}(\text{Wireless}, \text{Network}) = \frac{\sum_{i=1}^{n} V_{T1,T2_i} \times V_{T2,T1_i}}{\sqrt{\sum_{i=1}^{n} V_{T1,T2_i}{}^2} \times \sqrt{\sum_{i=1}^{n} V_{T2,T1_i}{}^2}} = \frac{10}{26 \times 28} = 0{,}013$$

N. Classe	Termes de la classe
6	advance, broadband, enable, end, fixed, network, new, service, technique, technology, user, wireless
7	access, advanced, bandwidth, broadband, connectivity, countries, dater, developed, digital, enables, facility, fixed, frequency, high, home, hopping, information, large, links, most, needs, network, networks, operators, pcs, personal, protocols, satellite, service, small, solution, spectrum, speed, spread, symmetrical, technique, technology, used, user, users, very, video, wireless
9	access, also, alternative, available, bandwidth, business, copper, dater, devoir, enhanced, expensive, fiber, high, internet, kbps, mbps, new, pair, speed, symmetrical, twisted
10	additional, allocated, allocation, alternative, available, communication, countries, currently, dater, different, europe, frequencies, most, offered, public, required, service, spectrum, wireless
11	access, building, expansion, figure, new, office, over, quickly, voice, wireless
13	airlan, allows, analog, based, being, cable, called, code, connected, dater, developed, direct, fiber, figure, large, line, most, multiple, new, radio, see, sequence, spectrum, spread, technology, telephone, used, what, which, wireless
15	line, modems, multiple, pair, phone, table, twisted
16	access, adapter, alternative, although, area, business, cdma, cellular, code, communication, countries, dater, digital, division, fiber, fixed, frequency, gsm, have, ieee, issu, its, lan, lans, latin, links, mobile, multiple, network, networks, personal, ranger, systems, time, transport, used, user, various, video, wireless, world
17	asymmetrical, broadband, copper, downstream, kbps, line, low, mbps, medium, phone, satellite, small, upstream
18	cost, hopping, lans, networks, provide, requires, scalable, sequence, service, solution, transmitters, two, used, wireless
19	access, broadband, dater, inc, integrated, international, local, market, pc, room, wireless
20	CISCO, content, LANS, links, wireless
21	currently, direct, frequency, hopping, infrared, microwave, radio, sequence, spectrum, spread, technique, transmission, transmitter, user
26	access, additional, available, business, communication, digital, europe, has, ieee, local, made, mobile, multiple, networks, spectrum, using, wireless
27	access, based, communication, hill, incorporated, international, mcgraw, medium, mobile, personal, providing, satellite, system, wireless
29	access, alternative, asymmetrical, available, broadband, cisco, company, dater, fixed, frequency, incorporated, infrared, international, lan, local, low, networking, networks, product, products, provide, rat, scalable, shared, solution, standard, symmetrical, systems, through, used, wireless
31	allocation, american, analog, countries, digital, europe, frequencies, latin, like, mhz, spectrum, vs
33	american, analog, cellular, digital, latin, most, now, operators, primarily, quality, systems, using, voice

Tableau 4.1 : Extrait de quelques classes de termes générées par Gramexco.

Ces deux vecteurs conceptuels partagent en effet, 10 entrées lexicales communes (*instrumentality, instrumentation, artifact, artefact, object, physical_object, entity, whole, whole_thing, unit*) sur un total de 45. La valeur du cosinus (0.013) confirme effectivement la relation sémantique entre *Wireless* et *Network.*

Par ailleurs, la valeur d'ISL (0,377) du couple de termes *user* et *bandwith* montre qu'il existe une relation sémantique entre ces deux termes en se basant sur les régularités constatées à travers le corpus. Cependant, la valeur nulle du cosinus relatif aux vecteurs conceptuels respectifs n'indique aucune relation entre ces deux termes. Inversement, le couple de termes *frequency* et *video* présente une relation sémantique selon les vecteurs conceptuels (cosinus = 0,007), mais, aucune relation selon l'ISL.

Les relations générées par ONTOLOGICO peuvent être classées dans 2 catégories de relations entre termes :

- la catégorie (**A**), regroupant les relations très fortes ; il s'agit des relations qui sont confirmées tant par l'ISL que les vecteurs conceptuels. Les valeurs correspondant à ces relations sont colorées en rouge par ONTOLOGICO.

- La Catégorie (**B**) regroupe des relations fortes qui sont repérées exclusivement par l'une ou l'autre des deux techniques (ISL ou vecteurs conceptuels). Les valeurs correspondant à ces relations sont colorées en bleu par ONTOLOGICO.

Le tableau suivant (**Tableau 4.2**) illustre un exemple de relations générées par ONTOLOGICO pour une dizaine de classes de termes. Dans ce tableau, la flèche (→) est interprétée comme « relié à ». Par exemple, « New→ service, *technology*, *user*, *wireless* » est interprété comme suit : *New* est relié à *service*, *New* est relié à *technology*…

N. Classe	Termes de la classe	Relations - Catégorie B	Relations - Catégorie A
6	advance, broadband, enable, end, fixed, network, new, service, technique, technology, user, wireless	Advance→ broadband, enable, end, fixed network, new, service, technique, technology, user, wireless Broadband→ enable, end, fixed, network, new, service, technology, user, wireless Enable→ end, fixed, service, technology, user, wireless End→ fixed, technique Fixed→ network, new, , technology, user Network→ new, service, technology, user, wireless New→ service, technology, user, wireless Service→ technique Technology→user, wireless	End→ network, service, technology, user, wireless Fixed→ service, wireless Network→ service, user, wireless Service→ technology, user, wireless User→ wireless
9	access, alternative, available, bandwidth, business, copper, dater, devoir, enhanced, expensive, fiber, high, internet, kbps, mbps, new, pair, speed, symmetrical, twisted	Access→ available, bandwidth, copper, dater, devoir, enhanced, internet, new, speed Alternative→ business, dater, devoir, expensive, fiber, high, new, speed Available→copper, dater, fiber, high, new, pair, speed, symmetrical, twisted Bandwidth→business, dater, high, mbps, new, speed, symmetrical business→ devoir, expensive, fiber, new, pair Cooper→dater, enhanced, fiber, kbps, mbps, pair, speed, symmetrical, twisted Dater→ high, mbps, new, pair, speed, symmetrical, twisted Devoir→ expensive, fiber, high, speed Enhanced→ high, kbps, mbps, pair, symmetrical, twisted Expensive→ fiber, high, speed Fiber→ new, pair, symmetrical High→ internet, kbps, mbps, new, symmetrical, twisted Internet→ new, speed Kbps→ mbps, pair, symmetrical, twisted Mbps→ pair, speed, symmetrical, twisted New→ speed Pair→ speed, symmetrical, twisted Speed→ symmetrical Symmetrical→ twisted	Access→ business, fiber, high, business→high, speed cooper→ high fiber→ high, speed high→pair, speed
11	access, building, expansion, figure, new, office, over, quickly,	Access→ building, expansion, new, voice Building→ expansion, figure, new, office, quickly, voice expansion→ figure, voice, wireless Figure→ over, voice New→ office, quickly, wireless Office→ voice Over→ voice, wireless	Access→ figure, office, wireless Building→ wireless expansion→ office Figure→ office, wireless Office→ wireless Voice→ wireless

	voice, wireless	Quickly→ wireless	
15	line, modem, multiple, pair, phone, table, twisted	Line→modem, multiple, twisted Modem→ multiple, pair, phone, table, twisted Multiple→ pair, phone, table, twisted Pair→ phone, twisted Phone→ twisted Table→twisted	Line→ pair, phone, table Pair→ table Phone→ table
17	asymmetrical, broadband, copper, downstream, kbps, line, low, mbps, medium, phone, satellite, small, upstream	Asymmetrical→ broadband, copper, downstream, kbps, line, low, mbps, medium, phone, satellite, small, upstream Broadband→ downstream, low, mbps, medium, satellite, small, upstream Copper→ downstream, kbps, line, mbps, medium, phone, satellite, upstream Downstream→ kbps, line, low, mbps, medium, phone, satellite, small Kbps→ line, mbps, phone, upstream Line→ , medium, satellite, small, upstream Low→ mbps, medium, phone, small, upstream Mbps→ medium, satellite, upstream Medium→ phone, upstream Phone→ satellite, upstream Satellite→ upstream Small→ upstream	Cooper→ low Downstream→ upstream Line→ phone Low→ satellite Medium→ satellite, small Phone→ small Satellite→ small
18	cost, hopping, LANS, networks, provide, requires, scalable, sequence, service, solution, transmitters, two, used, wireless	Cost→ provide, scalable, wireless Hopping→ sequence, transmitters, two LANS→ networks, used, wireless Networks→ provide, service, used, wireless Provide→ scalable, service, solution, wireless Requires→ service, used Scalable→ service, solution, wireless Sequence→ service, solution, transmitters, two, wireless Service→ solution, two, used Solution→ used Transmitters→ two Used→ wireless	Cost→ sequence, service, solution Service→ wireless Solution→ wireless
19	access, broadband, dater, integrated, international, local, market, pc, room, wireless	Access→ broadband, dater, integrated, international, room Broadband→ dater, local, market, wireless Dater→ market, wireless Integrated→ pc, room International→ wireless Local→ market, room Market→ wireless PC→ room, wireless Room→ wireless	Access→ local, market, wireless International→ local Local→ wireless
20	CISCO content,	Content→ wireless LANS→ links, wireless	Cisco→ content, wireless

	LANS, links, wireless	Links→ wireless	
31	allocation, American, analog, countries, digital, Europe, frequencies, Latin, like, MHz, spectrum, vs	Allocation→ American, analog, countries, digital, Europe, frequencies, Latin, like, MHz, spectrum, vs American→ countries, Europe, spectrum Analog→ countries, spectrum, vs Countries→ digital, frequencies, Latin, spectrum, vs Digital→ Europe, MHz, spectrum, vs Europe→ Europe, frequencies, Latin, like, MHz, Frequencies→ like, MHz, spectrum Latin→ MHz, spectrum Like→ MHz, spectrum MHz→ spectrum Spectrum→ vs	American→ Latin Analog→ digital Europe→ spectrum
33	American, analog, cellular, digital, Latin, operators, primarily, quality, systems, voice	American→ operators, primarily, quality, systems, voice Analog→ cellular, operators, quality, systems, voice Cellular→ digital, quality, systems, voice Digital→ operators, quality, systems, voice Latin→ operators, primarily, quality, systems, voice Operators→ primarily, quality, systems, voice Primarily→ systems Quality→ systems Systems→ voice	American→ Latin Analog→ digital Quality→ voice

Tableau 4.2 : Exemple de relations (fortes et très fortes) générées par ONTOLOGICO

L'ensemble de ces classes est constitué de 113 termes et de 657 relations possibles entre les couples de termes appartenant à une même classe. Pour des fins de simplification, nous considérons les relations bidirectionnelles entre deux termes comme étant une seule relation. ONTOLOGICO a retenu 54% des relations (soit 358 relations) comme étant des relations fortes et 10% comme étant des relations très fortes (soit 65 relations). Enfin, parmi les relations retenues, 85% des relations sont découvertes par l'une ou l'autre des techniques de l'ISL ou des vecteurs conceptuels. Cette grande proportion montre en partie, la complémentarité entre les deux techniques ; En effet, si une relation échappe à une technique, elle est découverte par l'autre et vice versa. Alors que 15% des relations sont confirmées par les deux techniques. Il reste toutefois à évaluer le taux de rappel et de précision de ces relations retenues. C'est ce que nous étudions plus loin dans cette section.

Le nombre moyen de termes par classe est de 11,3. L'analyse de la distribution moyenne des différents types de relations à l'intérieur de chaque classe de termes montre que le nombre moyen de relations très fortes est de 6,5 et celui des relations fortes est de 35,8.

Numéro Classe	Nombre termes	Nombre Relations possibles	Nombre Relations Fortes	Nombre Relations Très fortes
6	12	66	44	14
9	20	190	92	10
11	10	45	22	9
15	7	21	16	5
17	13	78	57	9
18	14	91	33	5
19	10	45	20	5
20	5	10	4	2
31	12	66	40	3
33	10	45	30	3
Total	113	657	358	65
Moyenne	11,30	65,70	35,80	6,50
% par rapport aux relations possibles			54%	10%
% par rapport aux relations retenues			85%	15%

Tableau 4.3 : Statistiques pour un extrait de 10 classes de termes sur les relations générées par ONTOLOGICO

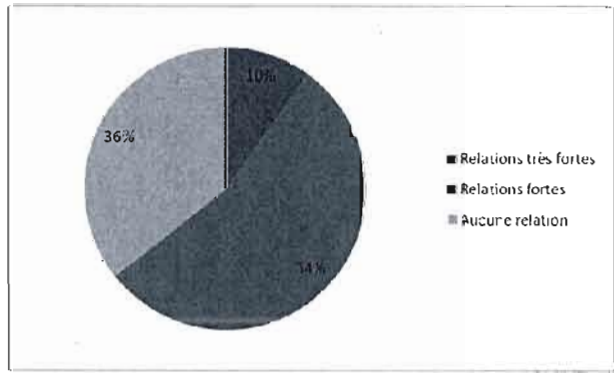


Figure 4.3 : Proportions des relations trouvées par rapport au nombre des relations possibles

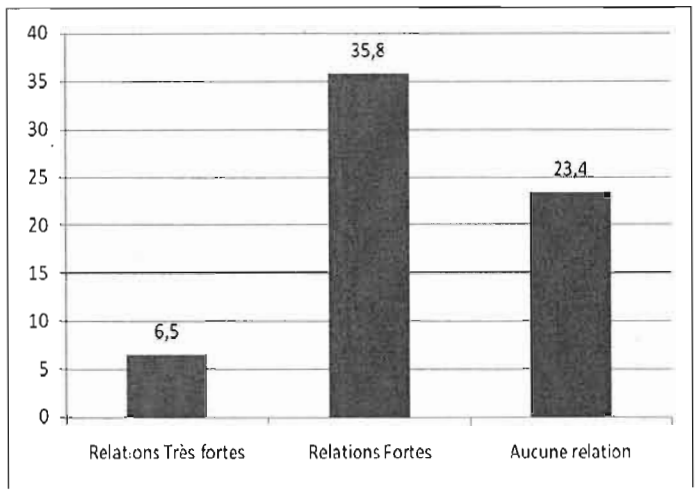


Figure 4.4 : Nombre moyen de relations par classe de termes

Le modèle génère effectivement des résultats impressionnants en termes de qualité des relations trouvées. Ces résultats doivent bien entendu être validés par un expert de domaine pour ne retenir que les relations pertinentes à l'ontologie du domaine.

Ce processus de découverte de relations permet le filtrage d'une grande proportion des relations potentielles. La tâche de validation pour un expert est par conséquent, bien facilitée. Ce dernier apportera son jugement sur des relations potentielles beaucoup moins nombreuses et dont la qualité est « confirmée » par une analyse documentaire et/ou et un thésaurus.

Par ailleurs, pour évaluer en partie, le degré d'assistance de l'outil ONTOLOGICO, il revient d'estimer le temps épargné par l'expert en utilisant cet outil plutôt que d'évaluer manuellement tous les couples de termes possibles à partir de tous les termes du corpus. En nous limitant aux 113 termes générés par le processus de classification (ce qui représente d'ores et déjà un filtrage relativement important), le nombre total de couples de termes potentiels est de 6328 (soit le nombre de combinaisons sans répétitions de 2 termes parmi 113 : C_{113}^{2}) et le nombre de couples de termes proposés par ONTOLOGICO est de 423 (soit 368 + 65). Ainsi, le temps épargné par l'expert est estimé à 93,32% ($\frac{6328 - 423}{6328} * 100$).

La découverte de relations se heurte, comme les autres tâches de construction d'ontologie, à la difficulté de l'évaluation du résultat obtenu. L'appréciation chiffrée de ces résultats (relatifs à la découverte de relations entre termes) est une entreprise bien difficile. En effet, le processus de validation, par un expert, comporte une certaine subjectivité qui nous empêche d'évaluer, avec exactitude, les taux de rappel et de précision des termes et des relations entre termes qui sont finalement retenus pour l'ontologie. Outre ce problème de subjectivité, une relation peut être retenue pour un contexte particulier d'ontologie et non pour un autre, dépendamment des besoins de l'application associée à cette ontologie. Aussi, le degré de granularité recherché pour l'ontologie peut également varier d'un contexte à un autre.

Étant conscients de ces défis, nous avons tenté de procéder à une évaluation des relations proposées par ONTOLOGICO par rapport à un repérage manuel de relations. L'appréciation des résultats est fondée sur des mesures standards d'évaluation à savoir les *taux de rappel* et de *précision*. La précision est une mesure standard de la qualité ; c'est la mesure de la proportion des relations correctes, sélectionnées tant par ONTOLOGICO (Onto) que par repérage manuel (Référentiel), par rapport à celles proposées par ONTOLOGICO. Sa formule est la suivante :

$$Taux\ de\ Précision = \frac{Onto \cap Référentiel}{Onto}$$

Le *rappel* est une mesure standard de la quantité de relations collectées. Sa formule est la suivante :

$$Taux\ de\ Rappel = \frac{Onto \cap Référentiel}{Référentiel}$$

Pour que l'évaluation du système ne soit biaisée par la subjectivité d'un seul juge, nous avons procédé à la soumission du système aux regards de trois juges. Ces derniers sont des experts du domaine des télécommunications et sont familiers avec les modèles sémantiques. Pour rapporter leur jugement, ils disposent d'un tableau croisé contenant les termes en lignes et en colonnes. L'exercice d'évaluation consiste à confirmer les relations entre les couples de termes en cochant les cases correspondantes.

Dans l'objectif de réduire la tâche des juges à une taille raisonnable, nous avons réduit l'échantillon d'évaluation à 10 classes de termes, choisies au hasard et contenant 113 termes et 657 relations possibles entre les couples de termes à l'intérieur de chaque classe. La consolidation des résultats des trois juges consiste à retenir les relations confirmées par au moins 2 juges.

Le tableau ci-dessous (Tableau 4.4) récapitule les résultats des relations repérées par les juges, par classe de termes.

Numéro Classe	Nbr termes	Nbr Relations possibles	Nbr Relations Ontologico	Nbr Relations Ontologico	Nbr Relations Juges	Nbr Relations Trouvées	Taux de Rappel	Taux de Précision
6	12	66	58	58	52	43	83%	74%
9	20	190	102	102	89	84	94%	82%
11	10	45	31	31	30	23	77%	74%
15	7	21	21	21	19	17	89%	81%
17	13	78	66	66	61	54	89%	82%
18	14	91	38	38	36	27	75%	71%
19	10	45	25	25	28	18	64%	72%
20	5	10	6	6	5	4	80%	67%
31	12	66	43	43	40	32	80%	74%
33	10	45	33	33	29	26	90%	79%
Total	113	657	423	423	389	328	84%	78%

Tableau 4.4 : Tableau Taux de rappel et de précision par classe de termes

En se basant sur cet échantillon, les résultats montrent que le modèle ONTOLOGICO possède des performances élevées en termes de rappel (84%) et de précision (78%). Ces résultats sont considérés comme relativement élevés, spécialement dans le domaine du TALN. Le taux de rappel par classe varie entre 64% et 94%. Celui de précision varie entre 67% et 82% (**Figure 4.5**). Ainsi, les taux moyens sont relativement représentatifs de l'ensemble des classes.

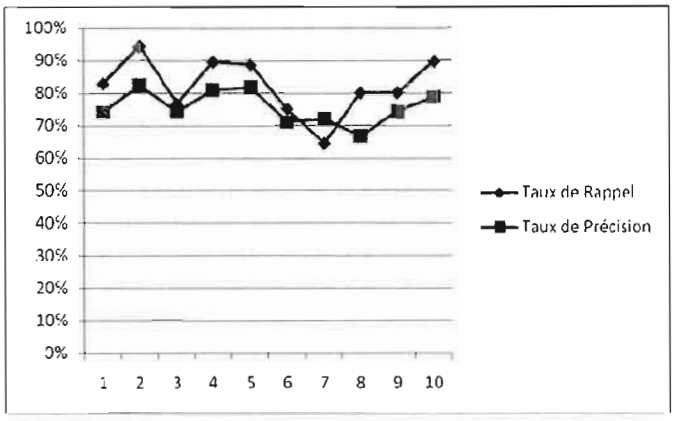


Figure 4.5 : Taux de rappel et de précision par classe de termes

Il est à noter toutefois, que ces résultats concernent plutôt le processus d'extraction de relations à l'intérieur des classes de termes, qui sont rappelons-le, générées par le réseau ART1. Les relations entre termes appartenant à des classes différentes ne sont toutefois pas négligées. En effet, le même terme peut se trouver dans une ou plusieurs classes, faisant en sorte que les relations entre classes sont également considérées d'une façon indirecte.

L'expérimentation réalisée utilise Wordnet comme base terminologique pour concevoir les VC. L'approche de VC telle que nous l'avons conçue permet d'extraire des relations, non seulement à partir d'un seul thésaurus, mais également à partir d'un ensemble de thésaurus, auquel cas, cette technique devient encore plus puissante. La performance du modèle proposé en termes de rappel et de précision peut ainsi être améliorée en suivant cette piste.

4.4 Validation des résultats

Notre démarche, fondée sur un ensemble d'hypothèses, détaillées dans la **section 1.5.2**, vise à confirmer une complémentarité des différentes approches abordées et démontre comment ces travaux permettent de mettre en œuvre les processus de repérage de termes et de relations entre termes. L'approche de validation consiste à confirmer (ou à infirmer) les hypothèses de recherche à la lumière des résultats de l'expérimentation. La base théorique de notre approche repose fondamentalement sur l'hypothèse qui stipule que les textes propres à un domaine constituent la source principale à considérer pour enrichir un modèle de l'ontologie. L'utilisation de corpus de textes non structurés offre différents avantages. D'abord, les textes

sont généralement facilement accessibles. La prolifération de documents électroniques sur le Web facilite en effet le traitement automatique. Ensuite, les textes non structurés contiennent des connaissances de domaine qui sont pertinentes pour les ontologies de domaine.

L'extraction de connaissances à partir de textes a été réalisée en se basant sur un processus de classification en vue de découvrir certaines régularités sémantiques entre des segments de texte. La proximité sémantique entre les termes repose principalement sur la cooccurrence d'un ensemble de termes à travers différents segments d'un corpus. Les classes de segments contiennent un certain type d'information similaire et servent, par conséquent, à détecter des indices précieux pour des associations entre termes. En effet, des termes qui ont des distributions comparables ont souvent un élément de sens commun.

Grâce à l'approche numérique de classification employée, un premier traitement rapide sur des données textuelles nous a permis de générer des classes de termes et de focaliser la suite des traitements sur ces classes. Les regroupements de termes se rapportent à des thèmes, généralement associés au domaine en question. Cependant, ces regroupements ne sont pas dépourvus de bruit, nécessitant ainsi, un traitement supplémentaire à savoir l'ISL.

- **L'Indexation Sémantique Latente présente de meilleurs résultats quand elle s'applique sur des classes de termes plutôt que sur tous les termes d'un document :**

L'ISL analyse l'ensemble d'un corpus pour en représenter les termes dans un espace sémantique multidimensionnel. Cette analyse statistique permet de faire ressortir les relations sémantiques entre termes. Deux termes peuvent être considérés sémantiquement proches s'ils sont utilisés dans des contextes similaires. Le contexte d'un terme est ici défini comme l'ensemble des termes qui apparaissent conjointement. Ainsi, les termes «*écrivain*» et «*auteur*» sont considérés comme sémantiquement proches puisqu'ils apparaissent tous les deux avec des termes tels que « *roman* », « *livre* », etc. et ils n'apparaissent que rarement avec des mots comme « *voiture* », « *cellulaire* », etc. Cette notion de cooccurrence est principalement de nature statistique et l'analyse sémantique ne fonctionne que si un nombre suffisant de textes est utilisé. Mais, il ne s'agit pas simplement de comptage, il faut aussi

disposer d'une procédure pour établir les liaisons sémantiques. Cette procédure est la réduction de la matrice.

L'une des solutions intuitives visant à réduire la matrice, consiste à subdiviser le corpus en un ensemble de textes de taille « raisonnable » et à appliquer l'ISL sur chacun de ces textes. Cette solution se heurte d'une part, à la difficulté empirique d'identifier une taille raisonnable, et d'autre part, au problème de la subdivision aléatoire du corpus et à la perte substantielle d'information engendrée par un traitement indépendant (de l'ISL) sur des parties de corpus qui sont sémantiquement reliés et qui, lorsque traités ensemble, auraient pu générer plus de relations sémantiques entre termes.

L'application de la technique d'Indexation Sémantique Latente sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du corpus, possède l'avantage de réduire la matrice de cooccurrence de termes dans les documents à une dimension raisonnable. Ce choix méthodologique constitue en effet, un remède à la difficulté d'identifier dans la théorie, une dimension adéquate et précise de cette matrice. Une dimension de trop grande taille pourrait en effet, empêcher l'émergence de suffisamment de relations sémantiques entre les termes, et aussi, une dimension de taille trop réduite pourrait entraîner une grande perte d'information.

- **Interprétation des scores d'association :**

En général, les scores d'association obtenus à partir de mesures différentes ne peuvent pas être comparés directement. Dans plusieurs cas aussi, la valeur numérique en tant que telle ne peut pas être interprétée. Typiquement, leur utilisation a plutôt pour but d'ordonnancer des paires de mots candidats en fonction de leur degré d'association. Tous les traitements et comparaisons effectués par la suite se basent sur des listes de n meilleures relations ou sur des seuils, et la valeur numérique en tant que telle n'est plus considérée.

L'approche des VC, telle que nous l'avons définie, permet de découvrir des relations, sans pour autant permettre la comparaison chiffrée des relations sémantiques. Notre objectif n'est pas d'attribuer une valeur numérique « exacte » à la relation sémantique entre deux termes, mais plutôt d'identifier des régularités associatives qui peuvent être interprétées comme des

associations sémantiques. Par ailleurs, une relation peut être pertinente pour une ontologie, mais pas pour une autre. Par conséquent, l'évaluation exacte de la relation sémantique n'est utile que lorsqu'elle est définie par rapport à un contexte particulier d'ontologie. Le contexte de l'ontologie est difficile à définir pour servir de référence à tout terme et spécialement pour les termes ne faisant pas partie de l'ontologie.

Les mesures de l'ISL et du cosinus identifient des régularités associatives que l'on interprète comme des associations sémantiques. Pour chacune des deux mesures, deux termes n'ont pas nécessairement besoin d'entretenir une relation sémantique directe pour être jugés sémantiquement reliés : ils peuvent se contenter de partager un ensemble de mots avec lesquels ils entretiennent une telle relation.

- **Complémentarité entre l'ISL et les vecteurs conceptuels** :

L'expérimentation réalisée montre que l'extraction de connaissances à partir de textes ne peut se contenter d'un traitement statistique (ni même linguistique) de données textuelles pour accaparer toute leur richesse sémantique. Le fondement cognitif est comparable à celui d'un processus de lecture de textes par un expert du domaine lors duquel le lecteur fait souvent usage de certaines de ses propres connaissances du domaine ou les connaissances extraites d'un dictionnaire ou d'un thésaurus (qu'on ne retrouve nécessairement pas dans les textes), pour repérer des relations d'associations conceptuelles entre termes.

Les connaissances implicites du domaine que l'approche du thésaurus tente de repérer, sont spécialement de nature linguistique, alors que l'analyse textuelle traite plutôt les connaissances du domaine. Par exemple, l'application de l'ISL sur notre corpus d'expérimentation génère une relation entre « *wireless* » et « *frequency* » avec un score de 0,25. Cette relation, propre au domaine de télécommunication sans fil, n'a pu être découverte par l'approche des VC. Et vice versa, l'approche des VC a permis de retrouver la relation entre « *information* » et « *access* » avec un score de cosinus de 0,005. Une telle relation n'a pu par contre, être extraite par l'analyse textuelle (basée sur l'ISL).

L'application du modèle ONTOLOGICO sur notre corpus d'expérimentation montre que la majorité des relations trouvées (soit 71%) proviennent de l'usage de l'ISL. En effet, 71% des

relations sont découvertes exclusivement par l'ISL, 18% sont extraites par l'usage exclusif de l'approche des VC, et 11% sont confirmées par les deux approches (**Figure 4.6**).

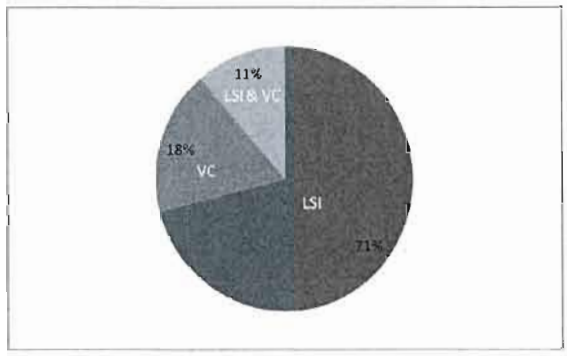


Figure 4.6 : Proportions des relations trouvées en fonction de l'approche utilisée

Les deux approches se complètent effectivement d'une façon remarquable. La grande proportion de relations trouvées à l'aide de l'analyse textuelle (71%) est particulièrement intéressante dans la mesure où ces relations sémantiques sont plutôt reliées au domaine. Bien qu'en théorie, les relations repérées par les VC soient plutôt de nature linguistique, ces relations ne sont pas en pratique, totalement indépendantes du domaine. En effet, l'application des VC est faite sur des classes de termes, provenant du processus de classification textuelle. Ces classes se rattachent à des thèmes qui sont généralement reliés d'une façon directe ou indirecte au domaine.

- **Pertinence des relations par rapport au domaine :**

Les relations entre termes générées par notre approche ne sont pas totalement indépendantes de l'ontologie du domaine en question. En effet, la pertinence de ces relations a été relativement guidée par le choix même du corpus qui est représentatif du domaine. Les classes de termes générées par le processus de classification sont généralement rattachées à des thématiques, reliées directement ou indirectement au domaine. Ainsi, le choix du corpus et sa pertinence par rapport au domaine influencent considérablement le résultat de l'analyse textuelle.

En examinant de plus près la composition des VC, nous constatons que leurs items lexicaux (synonymes, antonymes et hyperonymes) appartiennent à différents domaines. Ainsi, deux vecteurs conceptuels, relatifs à deux termes T_1 et T_2 peuvent avoir une relation sémantique parce que ces termes sont reliés dans d'autres domaines. Dans un tel cas, cette relation candidate n'est généralement pas confirmée par l'ISL et fera ainsi partie de la catégorie B (telle que définie dans l'évaluation de résultats).

Par exemple, les termes « *Technique* » et « *information* », ayant les VC suivants :

> Technique : [*technique, proficiency, method, know-how, ability, power,* ***cognition, knowledge, noesis, psychological_feature***]
>
> Information : [*information, info, data, selective_information, entropy, message, content, subject_matter, substance, communication, social_relation, relation, abstraction, collection, aggregation, accumulation, assemblage, group, grouping,* ***cognition, knowledge, noesis, psychological_feature,*** *information_measure, system_of_measurement, metric, measure, quantity, amount*]

Ces deux termes ont une valeur de cosinus de 0,014 grâce à la coexistence des termes *cognition, knowledge, noesis* et *psychological_feature* dans les deux vecteurs. *Technique* et *information* sont reliés par rapport au domaine de la cognition et non celui des télécommunications. Malgré cela, la valeur de l'ISL est nulle. Ainsi, cette relation a été découverte par l'approche des vecteurs conceptuels, malgré une faible régularité sémantique dans le corpus. Notons tout de même, que cette régularité a été faible mais non nulle, dans la mesure où elle a été décelée par le processus de classification. Ces deux termes figurent en effet dans la classe numéro 7.

- **Résolution du problème d'ambiguïté :**

La polysémie constitue l'un des problèmes majeurs du traitement automatique du langage naturel. Quand il est question d'un mot polysémique, la conception du vecteur conceptuel est entravée par le problème d'ambigüité sémantique. Quel sens choisir pour ce terme ? Nous avons opté pour une composition de VC qui soit relative à tous les sens possibles du terme.

Autrement dit, le VC est construit à partir des synonymes, des antonymes et des hyperonymes pour les différents sens du terme. Par conséquent, quand il est question de comparer deux VC, les items lexicaux communs se trouvent « dilués » par un tel bruit. Ce fait réduit relativement la valeur du cosinus sans pour autant l'annuler. Toutefois, cette ambiguïté sémantique est indirectement résolue par l'ISL et la classification par l'effet de la cooccurrence d'un couple de termes à travers le corpus. Ce dernier étant relatif au domaine.

L'application de la technique de classification permet d'identifier des groupes de termes qui apparaissent ensemble et qui ont des relations sémantiques ou, du moins, des similarités sémantiques lorsque utilisés dans des contextes comparables. La cooccurrence de termes à travers différentes parties de textes implique une couverture fort probable d'un même thème. Ceci permet d'identifier le contexte qu'un terme possède dans le texte, et par conséquent, de préciser son environnement sémantique correspondant dans lequel sa signification est employée, pour résoudre, dans une certaine mesure, le problème d'ambiguïté lexicale. Par exemple, le sens du terme « *souris* » est directement mis au clair, lorsque ce terme est présenté avec des cooccurrents tels que « *cliquer* », « *pointeur* », « *clavier* », etc., et est, par conséquent, distingué de la souris « *animal* » ou du verbe « *sourire* ».

Il existe donc un effet de compensation dans le modèle entre l'ISL et les vecteurs conceptuels en ce qui a trait au problème d'ambiguïté sémantique.

- **Choix de thésaurus :**

Les VC sont conçus à partir d'informations extraites d'un thésaurus. Bien que le thésaurus présente des définitions et des relations sémantiques se rapportant à des différents domaines, le thésaurus est avant tout un outil linguistique, reflétant la langue plutôt que le domaine d'application. Certaines spécificités relatives aux domaines d'expertise ne sont pas nécessairement couvertes par le thésaurus. Il est à noter que l'approche des VC, telle que nous l'avons définie dans notre modèle, est malgré cela indépendante de la langue. Il s'agit simplement d'utiliser le thésaurus de la langue en question et l'approche demeure toujours valable. D'ailleurs, l'approche proposée supporte également l'utilisation d'une combinaison de thésaurus en vue d'assurer une complémentarité plus élargie.

Dans notre expérimentation, nous utilisons Wordnet comme base de données lexicale. L'application Wordnet, disponible en ligne, présente les différentes définitions d'un terme, mais aussi les termes qui lui sont sémantiquement associés selon certaines relations spécifiques.

De par son mode de construction, essentiellement manuel, Wordnet, comme les principales ressources sémantiques exploitables sous forme électronique, ne se démarque pas fondamentalement des dictionnaires sous forme papier. Il s'appuie avant tout sur une formalisation et une systématisation des pratiques lexicographiques existantes. Les critiques formulées quant à leur inadéquation vis-à-vis du traitement automatique des langues ne sont pas surprenantes. Ces critiques portent à la fois sur la nature des sens qu'il distingue et sur leur caractérisation. Ces sens sont jugés à la fois trop fins et incomplets. Par exemple, le verbe « *give* » (donner) n'a pas moins de 44 sens. Une telle profusion ne facilite pas une tâche de désambiguïsation lexicale. Par contre, Wordnet ne contient que des informations limitées sur l'usage des mots. En effet, la caractérisation des sens, réalisée pour l'essentiel au travers des relations de synonymie, d'hyperonymie et d'hyponymie, manque d'éléments définissant leur contexte d'usage.

Par ailleurs, Wordnet ne définit pas des relations pragmatiques et ne matérialise pas d'une façon formelle tout le sens contenu dans les définitions des termes. Par exemple, l'information qu' « un chat ne rugit pas » figure dans la définition, mais ne se retrouve formalisée dans aucune relation. De même, des relations pragmatiques telles que « *soap / bath* » (*savon / bain*) sont absentes de Wordnet. De telles relations pragmatiques sont utiles pour les ontologies de domaine.

Les termes extraits de l'analyse textuelle ne sont pas tous présents dans ce thésaurus, spécialement les acronymes. Dans un tel cas, le VC correspondant n'est pas considéré. Seule l'ISL est prise en considération. Toutefois, l'approche que nous proposons pour constituer les vecteurs conceptuels favorise l'utilisation d'un ensemble de thésaurus. Ceci améliore effectivement la qualité de la composition des vecteurs conceptuels et remédie à l'incomplétude des thésaurus.

- **L'intervention d'un expert est une opération incontournable**

Le modèle proposé ne vise pas à remplacer complètement l'humain pour la découverte de relations entre termes, mais plutôt à assister les experts du domaine dans cette tâche. Nous avons ainsi, opté pour un modèle qui soit un compromis acceptable mais suffisamment performant et fonctionnel. Le modèle se base ainsi, sur différentes sources de connaissances, à savoir les textes relatifs au domaine, mais aussi, des données en vue d'accaparer le mieux possible, les connaissances du domaine ainsi que certaines connaissances reliées plutôt à la langue, au sein de l'ontologie.

Il est relativement facile de remettre en cause des systèmes automatiques prétendant accomplir cette tâche sans biais ou imperfection. Il semble plus raisonnable et plus réaliste de suivre un processus plutôt semi-automatique, impliquant une simple intervention d'un expert du domaine, à travers certaines étapes spécialement pour la validation des résultats. En effet, l'expert du domaine fait appel à d'autres connaissances relatives à son expérience, son expertise ainsi qu'au « bon sens » qui ne peuvent être extraits à partir d'un corpus ou d'un thésaurus. Ce type de connaissances est de nature purement humaine et ne peut être reproduit par la machine.

CHAPITRE V

CONCLUSION ET PERSPECTIVES

Ce chapitre est une synthèse des principales idées présentées dans la thèse et des résultats que le modèle proposé a pu atteindre. Nous concluons enfin, par l'énumération des différentes contributions originales de notre projet sur le plan scientifique, ainsi que les défis et les obstacles que nous avions à affronter. Dans la mesure où ce travail se veut une ouverture sur d'autres projets de recherche, nous proposons des voies de recherche susceptibles de déboucher sur des environnements d'assistance à la maintenance des ontologies.

5.1 Synthèse et contributions originales

Cette thèse présente des réflexions sur les différents thèmes de recherche abordés dans les sciences de la cognition, qui pourraient trouver une oreille attentive dans le domaine de la maintenance d'ontologie et ouvrir de nouvelles perspectives dans la recherche de solutions à certains problèmes rencontrés dans le domaine (extraction de relations sémantiques entre concepts à partir d'analyses textuelles, la cooccurrence, la classification de textes, la représentation de termes par des vecteurs conceptuels, l'extraction de relations sémantiques à partir de thésaurus). Pour chacun de ces thèmes, le lien avec les activités de maintenance et en particulier, la conceptualisation, est mis en exergue et sa place occupée dans la thèse en est précisée.

Cette thèse s'inscrit dans plusieurs disciplines inter-reliées : l'apprentissage machine, l'intelligence artificielle, le TALN, la psychologie cognitive... Au plan informatique, le développement et le déploiement des logiciels de repérage de l'information mettent à contribution les processus du génie logiciel. En particulier, les modèles de repérage développés mettent à contribution l'ingénierie des algorithmes, la modélisation des connaissances et la programmation des modèles dans un paradigme orienté objet.

Au plan cognitif, l'élaboration de systèmes d'assistance à la maintenance des ontologies nécessite l'intégration de processus de raisonnement et d'apprentissage, principalement

dédiés à la découverte de relations sémantiques ou d'association entre termes à partir de données textuelles. Les motivations cognitives se justifient entre autres, par le fait que certaines des difficultés du TALN ne sont pas dues à des difficultés d'ordre informatique, mais plutôt à des conceptions théoriques sur le traitement de données textuelles. Cette thèse met en évidence l'importante de la contribution des sciences de la cognition pour s'attaquer aux difficultés recensées dans notre problématique.

Notre travail a proposé des contributions originales que nous résumons dans ce qui suit :

- **Remèdes aux lacunes méthodologiques :**

  Nous présentons une *méthodologie* et des *outils* pour maintenir une ontologie à partir d'analyses textuelles. Dans la mesure où les outils d'édition des ontologies, disponibles actuellement, n'intègrent pas cette fonctionnalité de conceptualisation, il s'agit de l'aspect le plus important de notre contribution, remédiant spécialement aux lacunes méthodologiques.

- **Application de l'ISL sur une matrice réduite :**

  L'application de la technique d'ISL sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du texte, constitue un choix original qui a l'avantage de réduire la matrice de cooccurrence de termes dans les documents à une dimension raisonnable (Gargouri et al, 2003). En effet, une très grande dimension peut empêcher l'émergence de suffisamment de relations sémantiques entre les termes et à l'inverse, une dimension trop réduite peut entraîner une grande perte d'information.

- **Composition des Vecteurs Conceptuels :**

  En se basant sur la notion de Vecteurs Conceptuels, nous avons proposé une façon « optimale » de composition de ces vecteurs afin de permettre la découverte de relations sémantiques entre termes à partir de relations extraites d'un thésaurus. Cette technique constitue en effet, un processus cohérent de raffinement graduel de relations conceptuelles entre termes. Elle s'inspire d'une façon étroite de l'approche d'AFC dans sa représentation des concepts.

- **Utilisation potentielle d'un ensemble de thésaurus :**

  L'approche de VC telle que nous l'avons conçue permet d'extraire des connaissances, non seulement à partir d'un seul thésaurus, mais également à partir d'un ensemble de thésaurus. Ce qui rend cette technique encore plus performante en remédiant au problème d'incomplétude des thésaurus.

- **Contribution en extraction de connaissances à partir d'analyses textuelles:**

  Cette recherche se place au cœur des échanges entre terminologie et acquisition de connaissances. Elle amène par conséquent, une réflexion sur les divers paliers à envisager dans une telle démarche de modélisation de connaissances textuelles pour des objectifs de maintenance d'une ontologie. Les réflexions apportées à l'occasion du repérage de relations sémantiques entre termes, constituent une contribution aux travaux dans le domaine de l'extraction de connaissances à partir d'analyses statistiques de textes. Le modèle proposé assiste les terminologues chargés de naviguer à travers de vastes données textuelles pour extraire et normaliser la terminologie. Il facilite également la tâche des ingénieurs en connaissances chargés de modéliser des domaines.

- **Modèle indépendant de la langue et du domaine** :

  L'analyse textuelle proposée n'est pas spécifique à une langue particulière. Par conséquent, elle ne prend pas avantage des connaissances linguistiques, à l'exception de certaines procédures simples (telles que la lemmatisation, le filtrage de termes, etc.). De telles procédures, peuvent être adaptées à la langue en question sans pour autant remettre en cause les autres démarches méthodologiques.

  Bien que notre approche soit fondée au départ sur un corpus textuel, notre processus méthodologique de la maintenance d'ontologie demeure indépendant du corpus d'expérimentation et reste valable pour d'autres domaines. Ceci n'est pas le cas pour des approches linguistiques, basées par exemple, sur l'utilisation de marqueurs.

- **Complémentarité des sources de connaissances :**

  La découverte de relations entre termes pour les besoins d'une conceptualisation de domaine s'avère le résultat d'une complémentarité de traitements appliqués tant sur des

textes de domaine que sur un thésaurus. D'une part, les analyses textuelles fondées principalement sur l'application de l'Indexation Sémantique Latente sur des classes de termes génèrent des relations sémantiques précises. D'autre part, l'extraction de relations sémantiques à partir d'un thésaurus, en se basant sur une représentation par des Vecteurs conceptuels, constitue un choix théorique judicieux et performant. Ces deux processus jouent en effet, un rôle important dans la complétude des relations. L'application du modèle ONTOLOGICO sur notre corpus d'expérimentation montre que la majorité des relations trouvées (71%) proviennent de l'usage exclusif de l'ISL, 18% sont extraites par l'usage exclusif de l'approche des VC, et 11% sont confirmées par les deux approches. Ces résultats montrent donc la complémentarité entre les deux approches.

5.2 Problèmes et défis rencontrés

Nous sommes confrontés à la complexité de la terminologie, qui se trouve d'ailleurs au cœur des préoccupations dans diverses applications reliées à la représentation des connaissances, la construction d'ontologies, la gestion des connaissances, la veille scientifique, la traduction automatique, la classification ou la fouille de textes. Le problème majeur est celui de l'automatisation partielle ou totale des processus de repérage, de structuration des termes et de représentation des connaissances dans un domaine en s'appuyant sur la terminologie textuelle. D'autres défis plus spécifiques sont détaillés dans ce qui suit :

- L'ingénierie des ontologies ne peut être un processus linéaire, mais plutôt un processus nécessitant beaucoup d'interaction avec les textes ainsi que les thésaurus, mais aussi, des connaissances d'experts humains. Les textes et les thésaurus se complètent pour apporter à l'expert du domaine l'assistance nécessaire à la conceptualisation du domaine. Le défit est de mettre à profit, de la façon la plus optimale possible, ces deux sources de connaissance en vue d'apporter le plus d'assistance possible.

- L'ontologie du domaine n'est pas uniquement descriptive mais elle est également censée être utilisée pour produire un raisonnement. C'est ainsi que les relations entre termes ne se limitent pas à des relations de type synonymie ou hyperonymie. D'autres relations sémantiques sont également nécessaires pour permettre le

processus de raisonnement. Par exemple, la relation « *has* » que « *BANDWIDTH* » peut avoir avec « *SPEED* » intervient dans des raisonnements faisant intervenir ces deux termes pour inférer certains résultats. Il ne s'agit pas d'une relation intuitive qu'on peut facilement proposer comme relation candidate. Elle peut toutefois émerger d'une intelligente analyse textuelle. Notre défit était donc d'accaparer la richesse textuelle et d'en extraire les connaissances utiles à la modélisation conceptuelle.

- Les expérimentations réalisées montrent que la modélisation est plus facile lorsque les tâches de l'ontologie et le domaine sont bien délimités. La conceptualisation d'un domaine de connaissance ne peut se faire de manière non ambiguë que dans un champ d'application précis. Un même terme peut désigner deux concepts différents, c'est à dire désigner deux objets (qui peuvent être physiquement les mêmes) avec des sémantiques différentes dans deux cadres applicatifs différents. Ainsi, la même sémantique ne sera pas associée à l'objet « MÉMOIRE » dans le cas où la mémoire est la possibilité d'enregistrer des informations constituées par des expériences ou des événements, de les conserver et de pouvoir les utiliser (contexte applicatif de la médecine) et dans le cas où le mémoire est un exposé écrit d'une problématique de recherche (contexte applicatif de l'éducation). Modéliser des connaissances ne peut donc se faire que dans un domaine de connaissance donné et souvent dans un contexte applicatif donné. Cette restriction est nécessaire sans forcément être suffisante pour garantir l'unicité de la sémantique associée aux termes du domaine. Un terme ou une relation peuvent être pertinents pour un contexte d'utilisation dans une ontologie et non pour un autre. Lorsque la tâche n'est pas délimitée, la construction d'une ontologie ressemble plus à la construction d'une terminologie ou à de la lexicologie : la description d'un terme dans toute sa généralité devient une entreprise particulièrement difficile.

- Dans la mesure où notre modèle se fonde sur un processus semi-automatique, l'intervention d'un expert du domaine pour la validation des résultats au niveau de différentes étapes de notre processus se trouve confrontée à un problème de subjectivité. L'appréciation chiffrée de nos résultats est, par conséquent, une

entreprise bien difficile. Conscients de ces défis, nous avons procédé à une évaluation des relations proposées par ONTOLOGICO par rapport à un repérage manuel de relations réalisé par 3 juges. De cette façon, nous avons réduit l'effet du biais que l'évaluation aurait pu avoir si elle avait était conduite par un seul juge. Par ailleurs, et dans l'objectif de réduire la tâche des juges à une taille raisonnable, nous avons réduit la taille de notre échantillon de test en sélectionnant au hasard les classes de termes à évaluer.

- L'identification complète et précise des termes propres à un domaine à partir d'un corpus spécifique est considérée comme un traitement d'une grande importance pour la construction d'ontologie. Le problème est d'ailleurs partagé par d'autres travaux de recherche reliés à la recherche documentaire, l'indexation, la traduction, le résumé automatique, la terminologie, etc. Lors du processus de classification, le filtrage de termes est un prétraitement important pour améliorer la qualité des classes générées. Il s'agit entre autres de filtrer les termes très fréquents (dont la fréquence est supérieure à un certain seuil) en vue d'améliorer le caractère discriminatoire des segments de textes à classer. Toutefois, compte tenu du fait qu'après la classification, la suppression des mots très fréquents élimine définitivement des termes potentiellement pertinents pour l'ontologie à maintenir, le défit est donc de concilier la qualité de la classification et les étapes ultérieures. La solution choisie était de conserver les termes fréquents et d'accepter une classification d'une qualité relativement moins élevée. Le principal avantage de ce choix consiste dans le fait qu'un terme fréquent a la possibilité de faire partie de plusieurs classes de termes. Ainsi, les mots fréquents ont plus de chance d'être comparés avec plus de termes appartenant à différentes classes, ce qui est logiquement et sémantiquement justifié. Le second avantage est relatif au fait que des termes spécifiques à un domaine sont susceptibles d'être plus présents dans les corpus spécialisés.

- Le repérage de termes qui sont particulièrement spécifiques au domaine en question est un élément clé pour maintenir une ontologie. Le défit de la maintenance est d'éviter d'alourdir l'ontologie par des termes non pertinents par rapport au domaine. Nous appuyons l'idée qu'un corpus représentatif du domaine résout en grande partie

ce problème. Le filtrage final de ces termes se fait manuellement à la dernière étape de la méthodologie proposée. À cette étape, la sélection de termes spécifiques au domaine est considérablement réduite à une taille relativement raisonnable.

- Comme toute recherche en TALN, nous sommes confrontés au problème d'ambiguïté sémantique qui se pose au niveau de différentes étapes de notre processus itératif de maintenance. Les logiciels automatiques de désambiguïsation sémantique ne sont pas totalement précis et nécessitent une intervention humaine. Par rapport à un domaine de spécialité, un terme désigne un objet ou une opération, habituellement sans ambiguïté. Toutefois, la composition d'un VC tel que nous l'avons conçu est relative aux différents sens possibles du terme. Cette ambiguïté sémantique est indirectement résolue par l'ISL et la classification par l'effet de la cooccurrence d'un couple de termes à travers le corpus (voir **section 4.4**).

- Plusieurs approches statistiques et quantitatives sont utilisées pour extraire des termes complexes. Ces derniers n'ont pas été considérés dans notre expérimentation pour des fins de simplification. Les termes complexes sont tout simplement traités de la même façon que les termes simples, dans tous les processus de traitements.

- Enfin, les données textuelles que nous traitons sont de large taille. Nous sommes par conséquent confrontés à une contrainte de temps d'exécution, qui est croissante en fonction de la taille du corpus. Ainsi, sur le plan informatique, nous avions à surmonter plusieurs défis, spécialement reliés à l'optimisation du code java, étant donné que la complexité des algorithmes utilisés (en particulier, l'ISL) nécessite un processeur de grande puissance. Nos expérimentations ont été réalisées sur une machine de 1.8Mhz, 2Go de mémoire RAM. Nous avons alors opté pour un corpus de taille raisonnable par rapport à ces ressources.

5.3 Conclusion et perspectives de recherche

La conceptualisation d'un domaine met en jeu une variété de ressources et de traitements. Il importe donc que des efforts de recherche soient dirigés vers le développement d'applications visant à assister rigoureusement la construction et la mise à jour d'ontologies. Nous avons

présenté dans ce travail une méthodologie et des modules qui contribuent à atteindre ces objectifs.

Bien que la conceptualisation soit le processus le plus délicat dans une perspective de maintenance d'une ontologie de domaine, celle-ci nécessite également des étapes importantes d'ontologisation, d'opérationnalisation et d'évaluation. Ce travail ouvre ainsi la voie vers d'autres pistes de recherche se rapportant à ces différents processus.

Nous croyons que dans l'objectif de faciliter l'intégration de nouveaux concepts dans l'ontologie, il est important d'avoir une vue d'ensemble plus claire sur les concepts figurant dans l'ontologie courante, en vue de mieux visualiser les interdépendances des concepts existants avec les nouveaux candidats. Ainsi, un outil d'assistance à la maintenance doit nécessairement prévoir une visualisation graphique intégrant tant les termes et les relations de l'ontologie courante que ceux proposés comme candidats pour la maintenance. Il est évident que plus l'ontologie est élargie, moins l'humain est capable d'avoir une vue globale de celle-ci et de juger de la pertinence des nouvelles connaissances à intégrer et des modifications à apporter à celles déjà existantes.

En présence d'une action de maintenance, il existe généralement, plusieurs façons de préserver la consistance d'une ontologie. Par exemple, si un concept à l'intérieur d'une hiérarchie est supprimé, tous les sous-concepts peuvent être, soit supprimés, soit rattachés à d'autres concepts. Si les sous-concepts sont préservés, alors les propriétés du concept supprimé peuvent être propagées, leurs instances peuvent être distribuées,... Ainsi, pour chaque changement dans l'ontologie, il est possible de générer différents scénarios de changements additionnels, conduisant à divers états finaux cohérents. La plupart des systèmes actuels de développement d'ontologies fournissent une seule possibilité de changement qui est généralement la plus simple. Notre objectif futur, à cet effet, est d'assister le cogniticien, chargé de la maintenance, dans la génération de différents états plausibles de changements et de l'orienter dans ses choix.

Dans une perspective de construction d'une ontologie, il importe également d'assurer une certaine continuité entre le processus de construction d'ontologie et celui de la maintenance. La construction initiale d'une ontologie respecte des normes, des procédures d'élaboration,

des règles et des outils de formalisation. Par conséquent, il est inconcevable de séparer ces deux processus. Les actions de maintenance doivent être alignées vers les règles et normes régissant l'ontologie courante en vue de conserver la cohérence globale.

La gestion des erreurs potentielles dans l'ontologie est une difficulté supplémentaire de grande importance. Cette gestion constitue l'une des préoccupations de base lors du processus de maintenance. Un agrandissement excessif de l'ontologie entraîne l'accroissement de la complexité de la gestion des changements et empêche par conséquent, l'expert du domaine de parcourir efficacement les différents concepts et de procéder manuellement aux corrections. Par conséquent, il est important d'explorer des alternatives efficaces pour assister l'utilisateur lors de sa gestion des erreurs.

Dans plusieurs domaines de spécialité, les termes complexes sont importants dans une ontologie. Une fois que les termes complexes sont extraits d'un corpus, les différents traitements réalisés dans une perspective ONTOLOGICO demeurent identiques à celui des termes simples ; pour des fins de simplification, les résultats de nos expérimentations n'en tiennent pas compte

Nous avons présenté les nombreux avantages du réseau neuronal ART1 pour la classification de données textuelles. Cependant, le domaine de la classification est actuellement un important lieu de recherche. Il est possible que d'autres méthodes de classification, comme les cartes auto-organisatrices de Kohonen (Kohonen, 2001), puissent s'avérer encore plus performantes. La méthodologie proposée pourra même supporter un choix entre l'une ou l'autre des techniques de classification ou même la consolidation des résultats de classifications générées par une combinaison de techniques.

Par ailleurs, nous sommes d'avis que l'ajout de certaines fonctionnalités au modèle ONTOLOGICO visant entre autres, à identifier automatiquement les termes spécifiques à un domaine, pourrait être une amélioration bien intéressante. À cet effet, les thésaurus représentent la source de connaissances la plus importante à considérer.

Bref, le développement d'outils informatiques visant à assister à la maintenance d'ontologies demeure un lieu de recherche des plus complexes. Les problématiques reliées à l'extraction

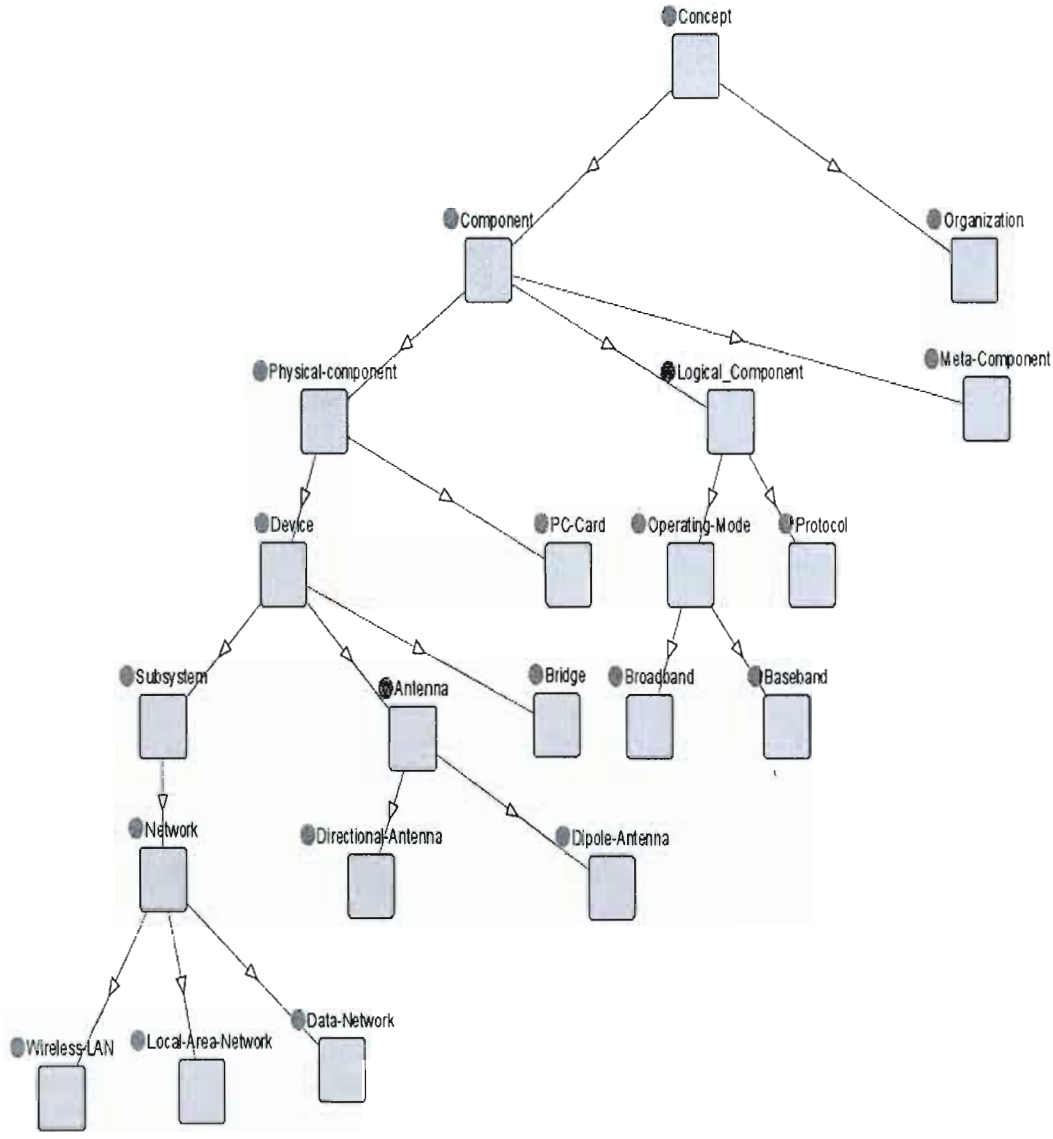
de termes et de relations entre termes soulèvent des enjeux théoriques et des défis techniques importants qui font et feront l'objet de plusieurs initiatives de recherche durant les prochaines années. Nous croyons que le développement d'une méthodologie de maintenance d'ontologie ne peut que bénéficier d'une synergie de plusieurs travaux de recherche. Nous espérons que notre travail contribuera à atteindre cet objectif.

ANNEXES

ANNEXE 1

ONTOLOGIE DE DOMAINE

Cette représentation graphique illustre un extrait de l'ontologie de domaine conçue dans le cadre du projet GDST.



ANNEXE 2

ONTOLOGIE DE DOMAINE

Ce document (*domainOntology.daml*) représente un extrait de l'ontologie de domaine conçue dans le cadre du projet GDST. Cette ontologie est le résultat des travaux de toute l'équipe de recherche du projet GDST.

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE uridef[
  <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns'>
  <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema'>
  <!ENTITY xsd 'http://www.w3.org/2000/10/XMLSchema'>
  <!ENTITY dc 'http://orlando.drc.com/daml/ontology/DC/3.2/dces-
ont'>
  <!ENTITY daml 'http://www.daml.org/2001/03/daml+oil'>
  <!ENTITY dom
'http://www.gdst.uqam.ca/Documents/Ontologies/domainOntology.daml'>
]>

<rdf:RDF
  xmlns:rdf = '&rdf;#'
  xmlns:rdfs ='&rdfs;#'
  xmlns:daml ='&daml;#'
  xmlns:dc ='&dc;#'
  xmlns:xsd ='&xsd;#'
  xmlns:dom ='&dom;#'
  xmlns  ='&dom;#'
  >

 <daml:Ontology rdf:about=''>
   <daml:versionInfo>Modified at date 2002/10/08 </daml:versionInfo>
   <rdfs:comment>An ontology for the IEEE 802.11b (Broadband
Wireless) Protocol.</rdfs:comment>
  <daml:imports rdf:resource='&daml;'/>
 </daml:Ontology>

 <daml:Class rdf:ID='Concept'>
  <rdfs:label>Concept</rdfs:label>
  <rdfs:comment>A concept is a Thing which can be quoted or ... in a
part of a document intersection of: All the things that linked to a
member of this class by the mean of the  isQuotedIn property (role)
are elements of Citation
  </rdfs:comment>

</daml:Class>
```

```
<daml:Class rdf:ID='Component'>
   <rdfs:comment xml:lang='en'>
      A component of a system
   </rdfs:comment>
   <rdfs:subClassOf rdf:resource='#Concept' />
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
   <rdfs:label xml:lang='en'> component </rdfs:label>
   <rdfs:label xml:lang='fr'> composante </rdfs:label>
 </daml:Class>

 <daml:Class rdf:ID='Physical-Component'>
   <rdfs:comment xml:lang='en'>
      Any tangible component of a system.
   </rdfs:comment>
   <rdfs:label xml:lang='en'> physical component</rdfs:label>
   <rdfs:label xml:lang='fr'> composante physique </rdfs:label>
   <rdfs:subClassOf rdf:resource='#Component' />

  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
 </daml:Class>

 <daml:Class rdf:ID='Logical-Component'>
   <rdfs:comment xml:lang='en'>
      An intangible component of a system.
   </rdfs:comment>
   <rdfs:label xml:lang='en'>logical component</rdfs:label>
   <rdfs:label xml:lang='fr'> composante logique </rdfs:label>
   <rdfs:subClassOf rdf:resource='#Component' />

  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
 </daml:Class>

<daml:Class rdf:ID='Meta-Component'>
 <rdfs:comment xml:lang='en'>
      Something that is about a system rather than of this system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Component' />
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>meta-component</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Device'>
  <rdfs:comment xml:lang='en'>
```

```
      A mechanical invention or contrivance for some specific
purpose.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Physical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>device</rdfs:label>
  <rdfs:label xml:lang='en'>apparatus</rdfs:label>
  <rdfs:label xml:lang='fr'>dispositif</rdfs:label>
  <rdfs:label xml:lang='fr'>appareil</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Subsystem'>
  <rdfs:subClassOf rdf:resource='#Device'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
 <rdfs:comment xml:lang='en'>
      A catch-all category for subsystems of a IEEE 802.11
networking system. (Note: Will have to be reconceptualized.)
  </rdfs:comment>
  <rdfs:label xml:lang='en'>subsystem</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Protocol'>
  <rdfs:comment xml:lang='en'>
      A set of rules formulated to control the exchange of data
between two communicating parties.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>protocol</rdfs:label>
  <rdfs:label xml:lang='fr'>protocole</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Station'>
  <rdfs:comment xml:lang='en'>Any device that contains an IEEE
802.11 conformant medium access control (MAC) and physical layer
interface (PHY) to the wireless medium (WM).
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Device'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>station</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Access-Point'>
```

```
  <rdfs:comment xml:lang='en'>Any entity that has station
functionality and provides access to the distribution services, via
the wireless medium for associated stations.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Station'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>access point</rdfs:label>
  <rdfs:label xml:lang='en'>AP</rdfs:label>
  <rdfs:label xml:lang='fr'>point d'accès</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Antenna-Subsystem'>
  <rdfs:comment xml:lang='en'>
      A wireless communication subsystem used for the transmission
and reception of a radio signal; it takes place at the interface
between the signal and its wired correspondent. An antenna system
comprises
numerous components, including the antenna, mounting hardware,
connectors, antenna cabling, and in some cases a lightning arrestor.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Subsystem'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>antenna subsystem</rdfs:label>
  <rdfs:label xml:lang='en'>antenna system</rdfs:label>
  <rdfs:label xml:lang='en'>antenna</rdfs:label>
  <rdfs:label xml:lang='fr'>antenne</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Antenna'>
  <rdfs:label>Antenna</rdfs:label>
  <daml:subClassOf rdf:resource='#Device'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <daml:subClassOf>
    <daml:Restriction daml:cardinalityQ='1'>
          <daml:onProperty rdf:resource='#hasSignalGain'/>
          <daml:hasClassQ rdf:resource='&xsd;#decimal'/>
    </daml:Restriction>
  </daml:subClassOf>

  <daml:subClassOf>
    <daml:Restriction daml:cardinalityQ='1'>
          <daml:onProperty rdf:resource='#hasSignalPolarization'/>
          <daml:hasClassQ rdf:resource='&xsd;#decimal'/>
    </daml:Restriction>
  </daml:subClassOf>
```

```
  <daml:subClassOf>
    <daml:Restriction daml:cardinalityQ='1'>
          <daml:onProperty rdf:resource='#hasSignalDirection'/>
          <daml:hasClassQ rdf:resource='&xsd;#integer'/>
    </daml:Restriction>
  </daml:subClassOf>
</daml:Class>

<daml:DatatypeProperty rdf:ID='hasSignalGain'>
  <rdfs:comment xml:lang='en'>
      An increase in the power of a signal.
  </rdfs:comment>
  <rdfs:label xml:lang='en'>gain</rdfs:label>
  <rdfs:label xml:lang='en'>signal gain</rdfs:label>
  <rdfs:label xml:lang='fr'>gain</rdfs:label>
</daml:DatatypeProperty>

<daml:DatatypeProperty rdf:ID='hasSignalDirection'>
  <rdfs:comment xml:lang='en'>
      Direction is the shape of the transmission pattern.
  </rdfs:comment>
  <rdfs:label xml:lang='en'>direction</rdfs:label>
  <rdfs:label xml:lang='en'>signal direction</rdfs:label>
  <rdfs:label xml:lang='fr'>direction</rdfs:label>
</daml:DatatypeProperty>

<daml:DatatypeProperty rdf:ID='hasSignalPolarization'>
  <rdfs:comment xml:lang='en'>
      No definition available
  </rdfs:comment>
  <rdfs:label xml:lang='en'>polarization</rdfs:label>
  <rdfs:label xml:lang='fr'>polarisation</rdfs:label>
</daml:DatatypeProperty>

<daml:Class rdf:ID='Dipole-Antenna'>
  <rdfs:comment xml:lang='en'>An antenna that is a single linear
conductor separated at the center by a transmission line feed or
other source or receiver of radio signals. The dipole radiation
pattern is 360 degrees in the horizontal plane and 75 degrees in the
vertical plane (assuming the dipole antenna is standing vertically)
and resembles a donut in shape. </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Antenna'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>dipole antenna</rdfs:label>
  <rdfs:label xml:lang='en'>dipole</rdfs:label>
  <rdfs:label xml:lang='fr'>antenne dipôle</rdfs:label>
  <rdfs:label xml:lang='fr'>dipôle</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Omni-Directional-Antenna'>
```

```
  <rdfs:comment xml:lang='en'>
      An omni-directional antenna is designed to provide a 360
degree radiation pattern. This type of antenna is used when coverage
in all directions from the antenna is required.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Antenna'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>omni-directional antenna</rdfs:label>
  <rdfs:label xml:lang='fr'>antenne omnidirectionnelle</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Directional-Antenna'>
  <rdfs:comment xml:lang='en'>An antenna that sends or receives a
radio signal in one direction only. </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Antenna'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>directional antenna</rdfs:label>
  <rdfs:label xml:lang='fr'>antenne directionnelle</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Patch-Antenna'>
  <rdfs:comment xml:lang='en'>A type of directional
antenna.</rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Directional-Antenna'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>patch antenna</rdfs:label>
  <rdfs:label xml:lang='en'>directional patch antenna</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Yagi-Antenna'>
  <rdfs:comment xml:lang='en'>A type of directional antenna.
</rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Directional-Antenna'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>yagi</rdfs:label>
  <rdfs:label xml:lang='en'>yagi antenna</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Parabolic-Dish'>
  <rdfs:comment xml:lang='en'>A type of directional antenna shaped
as a dish and having a very narrow RF energy path.</rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Directional-Antenna'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
```

```
  </rdf:type>
  <rdfs:label xml:lang='en'>parabolic dish</rdfs:label>
  <rdfs:label xml:lang='fr'>antenne parabolique</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Area'>
  <rdfs:comment xml:lang='en'>A region of space.</rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>area</rdfs:label>
  <rdfs:label xml:lang='fr'>aire</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Band'>
  <rdfs:comment xml:lang='en'>A specified range of radio
wavelengths.</rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>band</rdfs:label>
  <rdfs:label xml:lang='en'>frequency band</rdfs:label>
  <rdfs:label xml:lang='fr'>bande</rdfs:label>
  <rdfs:label xml:lang='fr'>bande de fréquence</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Bandwidth'>
  <rdfs:comment xml:lang='en'>The difference between the highest and
lowest sinusoidal frequency signals that can be transmitted across a
transmission line or through a network. It is measured in hertz (Hz)
and also defines the maximum information-carrying capacity of the
line or network.</rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>bandwidth</rdfs:label>
  <rdfs:label xml:lang='fr'>largeur de bande</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Base-Station'>
  <rdfs:comment xml:lang='en'>A radio transmitter placed at the
fixed-wire termination point and providing the cordless link between
each computer and the central site.</rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Station'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>base station</rdfs:label>
  <rdfs:label xml:lang='en'>base</rdfs:label>
```

```
  <rdfs:label xml:lang='fr'>station de base</rdfs:label>
  <rdfs:label xml:lang='fr'>base</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Operating-Mode'>
  <rdfs:comment xml:lang='en'>A mode in which a system operates.
</rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>operating mode</rdfs:label>
  <rdfs:label xml:lang='en'>mode</rdfs:label>
  <rdfs:label xml:lang='fr'>mode d'opération</rdfs:label>
  <rdfs:label xml:lang='fr'>mode</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Baseband'>
  <rdfs:comment xml:lang='en'>A particular operating mode of a
transmission line in which all the available bandwidth is used to
derive a single high bit rate (10 Mbps or higher) transmission path
(channel). The opposite of 'broadband'. </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Operating-Mode'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>baseband</rdfs:label>
  <rdfs:label xml:lang='en'>baseband mode</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Basic-Service-Set'>
  <rdfs:comment xml:lang='en'>A set of stations controlled by a
single coordination function. </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Subsystem'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>basic service set</rdfs:label>
  <rdfs:label xml:lang='en'>BSS</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Bridge'>
  <rdfs:comment xml:lang='en'>A device used to link two homogeneous
local area subnetworks, that is, two subnetworks utilizing the same
physical and medium access control method. </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Device'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>bridge</rdfs:label>
  <rdfs:label xml:lang='fr'>pont</rdfs:label>
</daml:Class>
```

```
<daml:Class rdf:ID='Broadband'>
  <rdfs:comment xml:lang='en'>
      A particular mode of operation of a coaxial cable in which the
available bandwidth is divided to derive a number of lower bandwidth
subchanneles (and hence transmission paths) on one cable. The
opposite of 'baseband'.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Operating-Mode'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>broadband</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Network'>
  <rdfs:subClassOf rdf:resource='#Subsystem'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:comment xml:lang='en'>
      A set of stations interconnected so as to be able to exchange
information.
  </rdfs:comment>
  <rdfs:label xml:lang='en'>network</rdfs:label>
  <rdfs:label xml:lang='fr'>réseau</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Data-Network'>
  <rdfs:comment xml:lang='en'>
      A network dedicated to the transmission of digital data.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Network'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>data network</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Local-Area-Network'>
  <rdfs:comment xml:lang='en'>
      A data communication network used to interconnect a community
of digital devices distributed over a localized area of up to, say,
10 square kilometers.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Data-Network'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>local area network</rdfs:label>
  <rdfs:label xml:lang='en'>LAN</rdfs:label>
  <rdfs:label xml:lang='fr'>réseau local</rdfs:label>
```

```
  <rdfs:label xml:lang='fr'>LAN</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Wireless-LAN'>
  <rdfs:comment xml:lang='en'>
      A LAN that uses either radio or infrared as the transmission
medium.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Local-Area-Network'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>wireless LAN</rdfs:label>
  <rdfs:label xml:lang='en'>WLAN</rdfs:label>
  <rdfs:label xml:lang='fr'>réseau local sans fil</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Broadband-Fixed-Wireless'>
  <rdfs:comment xml:lang='en'>
      A broadband wireless network that revolves around a fixed
access point.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Wireless-LAN'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>broadband fixed wireless</rdfs:label>
  <rdfs:label xml:lang='en'>BBFW</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Broadband-Radio-Access-Network'>
  <rdfs:comment xml:lang='en'>
      A broadband wireless network.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Wireless-LAN'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>broadband radio access
network</rdfs:label>
  <rdfs:label xml:lang='en'>BRAN</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Addressing-Mode'>
  <rdfs:comment xml:lang='en'>
      The means whereby a message may be addressed to a single or to
multiple devices on a network.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
```

```
  <rdfs:label xml:lang='en'>addressing mode</rdfs:label>
  <rdfs:label xml:lang='fr'>mode d'adressage</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Broadcast'>
  <rdfs:comment xml:lang='en'>
      A means of transmitting a message to all devices connected to
a network.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Addressing-Mode'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>broadcast</rdfs:label>
  <rdfs:label xml:lang='fr'>transmission en diffusion</rdfs:label>
  <rdfs:label xml:lang='fr'>diffusion</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Channel'>
  <rdfs:comment xml:lang='en'>
      A narrow band of frequencies used in radio transmissions as a
path for transmitting a signal or data.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>channel</rdfs:label>
  <rdfs:label xml:lang='fr'>canal</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Broadcast-Channel'>
  <rdfs:comment xml:lang='en'>
      A channel used for broadcasting.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Channel'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>broadcast channel</rdfs:label>
  <rdfs:label xml:lang='fr'>canal de diffusion</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Transmission-Medium'>
  <rdfs:comment xml:lang='en'>
      The communication path linking two communicating devices.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Physical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>transmission medium</rdfs:label>
```

```
  <rdfs:label xml:lang='fr'>support de transmission</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Cable'>
  <rdfs:comment xml:lang='en'>
      An encased group of insulated wires for transmitting
electricity or telecommunications.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Transmission-Medium'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>cable</rdfs:label>
  <rdfs:label xml:lang='en'>wire</rdfs:label>
  <rdfs:label xml:lang='fr'>câble</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='PC-Card'>
  <rdfs:comment xml:lang='en'>
      An adapter containing a radio transmitter/receiver that allows
a PC to send and receive messages wirelessly.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Physical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>PC card</rdfs:label>
  <rdfs:label xml:lang='en'>PCMCIA adapter</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Clear-Channel-Assessment'>
  <rdfs:comment xml:lang='en'>
      An assessment of the current state of use of the wireless
medium.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>clear channel assessment</rdfs:label>
  <rdfs:label xml:lang='en'>CCA</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Feature'>
  <rdfs:comment xml:lang='en'>
      A distinct or outstanding part, quality or characteristic of
something.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Concept' />
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
```

```
  <rdfs:label xml:lang='en'>feature</rdfs:label>
  <rdfs:label xml:lang='en'>characteristic</rdfs:label>
  <rdfs:label xml:lang='fr'>caractéristique</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Client-Organization'>
  <rdfs:comment xml:lang='en'>
      An organization that acts as a client, specifically in the
context of the provision of networking services.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Organization'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>client organization</rdfs:label>
  <rdfs:label xml:lang='en'>client</rdfs:label>
  <rdfs:label xml:lang='en'>customer</rdfs:label>
  <rdfs:label xml:lang='fr'>client</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Configuration'>
  <rdfs:comment xml:lang='en'>
      A specific set-up of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>configuration</rdfs:label>
  <rdfs:label xml:lang='fr'>configuration</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Servicing-Phase'>
  <rdfs:comment xml:lang='en'>
      Any of the stages in the servicing relationship between a
service provider and its client.
  </rdfs:comment>
  <rdfs:label xml:lang='en'>servicing phase</rdfs:label>
  <rdfs:label xml:lang='en'>servicing stage</rdfs:label>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
</daml:Class>

<daml:Class rdf:ID='Deployment'>
  <rdfs:comment xml:lang='en'>
      The servicing phase in which a system is put to effective
action.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Servicing-Phase'/>
  <rdf:type>
```

```
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>deployment</rdfs:label>
  <rdfs:label xml:lang='fr'>mise en service</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Property'>
  <rdfs:comment xml:lang='en'>
      A property of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Meta-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>property</rdfs:label>
  <rdfs:label xml:lang='fr'>propriété</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Design'>
  <rdfs:comment xml:lang='en'>
      The general arrangement or layout of a system or part thereof.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Property'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>design</rdfs:label>
  <rdfs:label xml:lang='en'>layout</rdfs:label>
  <rdfs:label xml:lang='fr'>design</rdfs:label>
  <rdfs:label xml:lang='fr'>conception</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Limitation'>
  <rdfs:comment xml:lang='en'>
      A limitation of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Property'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>limitation</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Capability'>
  <rdfs:comment xml:lang='en'>
      A capability of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Property'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>capability</rdfs:label>
```

```
</daml:Class>

<daml:Class rdf:ID='Technical-Specification'>
  <rdfs:comment xml:lang='en'>
      The technical specification for a system or part thereof.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Meta-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>technical specification</rdfs:label>
  <rdfs:label xml:lang='en'>specification</rdfs:label>
  <rdfs:label xml:lang='fr'>spécification technique</rdfs:label>
  <rdfs:label xml:lang='fr'>spécification</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Accessory'>
  <rdfs:comment xml:lang='en'>
      An accessory part of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Physical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>accessory part</rdfs:label>
  <rdfs:label xml:lang='en'>accessory</rdfs:label>
  <rdfs:label xml:lang='fr'>accessoire</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Scenario'>
  <rdfs:comment xml:lang='en'>
      A description of how to proceed about something.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Meta-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>scenario</rdfs:label>
  <rdfs:label xml:lang='fr'>scénario</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Installation-Scenario'>
  <rdfs:comment xml:lang='en'>
      A description of how to proceed with installation of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Scenario'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>installation scenario</rdfs:label>
  <rdfs:label xml:lang='fr'>scénario d'installation</rdfs:label>
</daml:Class>
```

```
<daml:Class rdf:ID='Diagram'>
  <rdfs:comment xml:lang='en'>
      A diagrammatic representation of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Meta-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>diagram</rdfs:label>
  <rdfs:label xml:lang='fr'>diagramme</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Requirement'>
  <rdfs:comment xml:lang='en'>
      Something that must be present in the definition of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Capability'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>requirement</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Coverage'>
  <rdfs:comment xml:lang='en'>
      The area covered or reached.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Area'/>
  <rdfs:subClassOf rdf:resource='#Property'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>coverage</rdfs:label>
  <rdfs:label xml:lang='fr'>couverture</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Performance'>
  <rdfs:comment xml:lang='en'>
      The performance of a system.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Property'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>performance</rdfs:label>
  <rdfs:label xml:lang='fr'>performance</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Regulatory-Information'>
  <rdfs:comment xml:lang='en'>
      Information about the regulations that pertain to a system.
```

```
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Meta-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>regulatory information</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Organization'>
  <rdfs:comment xml:lang='en'>
      The base class for all organizations, whether corporations,
professional associations, governements or other.
  </rdfs:comment>
  <rdfs:label xml:lang='en'>organization</rdfs:label>
  <rdfs:label xml:lang='en'>organisation</rdfs:label>
  <rdfs:label xml:lang='fr'>organisation</rdfs:label>
  <rdfs:label xml:lang='fr'>organisme</rdfs:label>
  <daml:subClassOf rdf:resource='#Concept'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
</daml:Class>

<daml:Class rdf:ID='Service-Provider'>
  <rdfs:comment xml:lang='en'>
      A provider of networking services.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Organization'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>service provider</rdfs:label>
  <rdfs:label xml:lang='en'>Internet service provider</rdfs:label>
  <rdfs:label xml:lang='en'>ISP</rdfs:label>
  <rdfs:label xml:lang='fr'>fournisseur d'accès</rdfs:label>
  <rdfs:label xml:lang='fr'>fournisseur de services</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Manufacturer'>
  <rdfs:comment xml:lang='en'>
      Specifically, a manufacturer of networking equipment.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Organization'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>manufacturer</rdfs:label>
  <rdfs:label xml:lang='fr'>manufacturier</rdfs:label>
  <rdfs:label xml:lang='fr'>fabricant</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Standards-Organization'>
```

```
  <rdfs:comment xml:lang='en'>
      An organization that issues standards.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Organization'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>standards organization</rdfs:label>
  <rdfs:label xml:lang='en'>standards organisation</rdfs:label>
  <rdfs:label xml:lang='en'>standards board</rdfs:label>
  <rdfs:label xml:lang='en'>standards institute</rdfs:label>
  <rdfs:label xml:lang='en'>standards institution</rdfs:label>
  <rdfs:label xml:lang='en'>standards body</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Regulation-Organization'>
  <rdfs:comment xml:lang='en'>
      A para-governmental organization called upon to regulate the
telecommunications field.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Organization'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
</daml:Class>

<daml:Class rdf:ID='Territory'>
  <rdfs:comment xml:lang='en'>
      A politically-defined geographical territory.
  </rdfs:comment>
  <rdfs:label xml:lang='en'>territory</rdfs:label>
  <rdfs:label xml:lang='fr'>territoire</rdfs:label>
  <rdfs:subClassOf rdf:resource='#Concept' />
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
</daml:Class>

<daml:ObjectProperty rdf:ID='has_territorial_jurisdiction'>
  <rdfs:comment xml:lang='en'>
      The property of an organization having jurisdiction over a
territory.
  </rdfs:comment>
  <rdfs:domain rdf:resource='#Organization'/>
  <rdfs:range rdf:resource='#Territory'/>
  <rdfs:label xml:lang='en'>jurisdiction</rdfs:label>
  <rdfs:label xml:lang='fr'>juridiction</rdfs:label>
</daml:ObjectProperty>

<daml:Class rdf:ID='Network-Topology'>
  <rdfs:comment xml:lang='en'>
      The class of all network configurations (LAN, WAN and so on).
```

```
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>network</rdfs:label>
  <rdfs:label xml:lang='fr'>réseau</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Wireless-Transmission-Medium'>
  <rdfs:comment xml:lang='en'>
      Any transmission medium that isn't based on wires.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Transmission-Medium'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>wireless medium</rdfs:label>
  <rdfs:label xml:lang='en'>wireless transmission
medium</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Transmission-Scheme'>
  <rdfs:comment xml:lang='en'>
      The specific transmission technique used to transmit data over
a transmission medium.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Logical-Component'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>transmission scheme</rdfs:label>
  <rdfs:label xml:lang='en'>transmission technique</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Wireless-Transmission-Scheme'>
  <rdfs:comment xml:lang='en'>
      A transmission scheme used with wireless media.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Transmission-Scheme'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>wireless transmission
scheme</rdfs:label>
  <rdfs:label xml:lang='en'>wireless transmission
technique</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Radio-Transmission-Scheme'>
  <rdfs:comment xml:lang='en'>
      A wireless transmission scheme used with radio transmission.
```

```
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Wireless-Transmission-Scheme'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>radio transmission scheme</rdfs:label>
  <rdfs:label xml:lang='en'>radio transmission
technique</rdfs:label>
</daml:Class>

<daml:Class rdf:ID='Spread-Spectrum-Transmission-Scheme'>
  <rdfs:comment xml:lang='en'>
      A radio transmission scheme in which the signal is spread over
a proportionately wider frequency band than the original source data
bandwidth, which makes the signal appear as (pseudo) noise to other
users of the same frequency band, thus alleviating co-channel
interference rejection.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource='#Radio-Transmission-Scheme'/>
  <rdf:type>
    <daml:Class rdf:about="#Concept"/>
  </rdf:type>
  <rdfs:label xml:lang='en'>spread spectrum</rdfs:label>
</daml:Class>
</rdf:RD
```

ANNEXE 3

LISTE DE CLASSES

Le tableau suivant illustre quelques exemples de classes de termes générées par Gramexco.

N. Classe	Termes de la classe
6	advance, broadband, enable, end, fixed, network, new, service, technique, technology, user, wireless
7	access, advanced, bandwidth, broadband, connectivity, countries, dater, developed, digital, enables, facility, fixed, frequency, high, home, hopping, information, large, links, most, needs, network, networks, operators, pcs, personal, protocols, satellite, service, small, solution, spectrum, speed, spread, symmetrical, technique, technologie, used, user, users, very, video, wireless
9	access, also, alternative, available, bandwidth, business, copper, dater, devoir, enhanced, expensive, fiber, high, internet, kbps, mbps, new, pair, speed, symmetrical, twisted
10	additional, allocated, allocation, alternative, available, communication, countries, currently, dater, different, europe, frequencies, most, offered, public, required, service, spectrum, wireless
11	access, building, expansion, figure, new, office, over, quickly, voice, wireless
13	airlan, allows, analog, based, being, cable, called, code, connected, dater, developed, direct, fiber, figure, large, line, most, multiple, new, radio, see, sequence, spectrum, spread, technology, telephone, used, what, which, wireless
15	line, modems, multiple, pair, phone, table, twisted
16	access, adapter, alternative, although, area, business, cdma, cellular, code, communication, countries, dater, digital, division, fiber, fixed, frequency, gsm, have, ieee, issu, its, lan, lans, latin, links, mobile, multiple, network, networks, personal, ranger, systems, time, transport, used, user, various, video, wireless, world
17	asymmetrical, broadband, copper, downstream, kbps, line, low, mbps, medium, phone, satellite, small, upstream
18	cost, hopping, lans, networks, provide, requires, scalable, sequence, service, solution, transmitters, two, used, wireless
19	access, broadband, dater, inc, integrated, international, local, market, pc, room, wireless
21	currently, direct, frequency, hopping, infrared, microwave, radio, sequence, spectrum, spread, technique, transmission, transmitter, user
26	access, additional, available, business, communication, digital, europe, has, ieee, local, made, mobile, multiple, networks, spectrum, using, wireless
27	access, based, communication, hill, incorporated, international, mcgraw, medium, mobile, personal, providing, satellite, system, wireless
29	access, alternative, asymmetrical, available, broadband, cisco, company, dater, fixed, frequency, incorporated, infrared, international, lan, local, low, networking, networks, product, products, provide, rat, scalable, shared, solution, standard, symmetrical, systems, through, used, wireless
31	allocation, american, analog, countries, digital, europe, frequencies, latin, like, mhz,

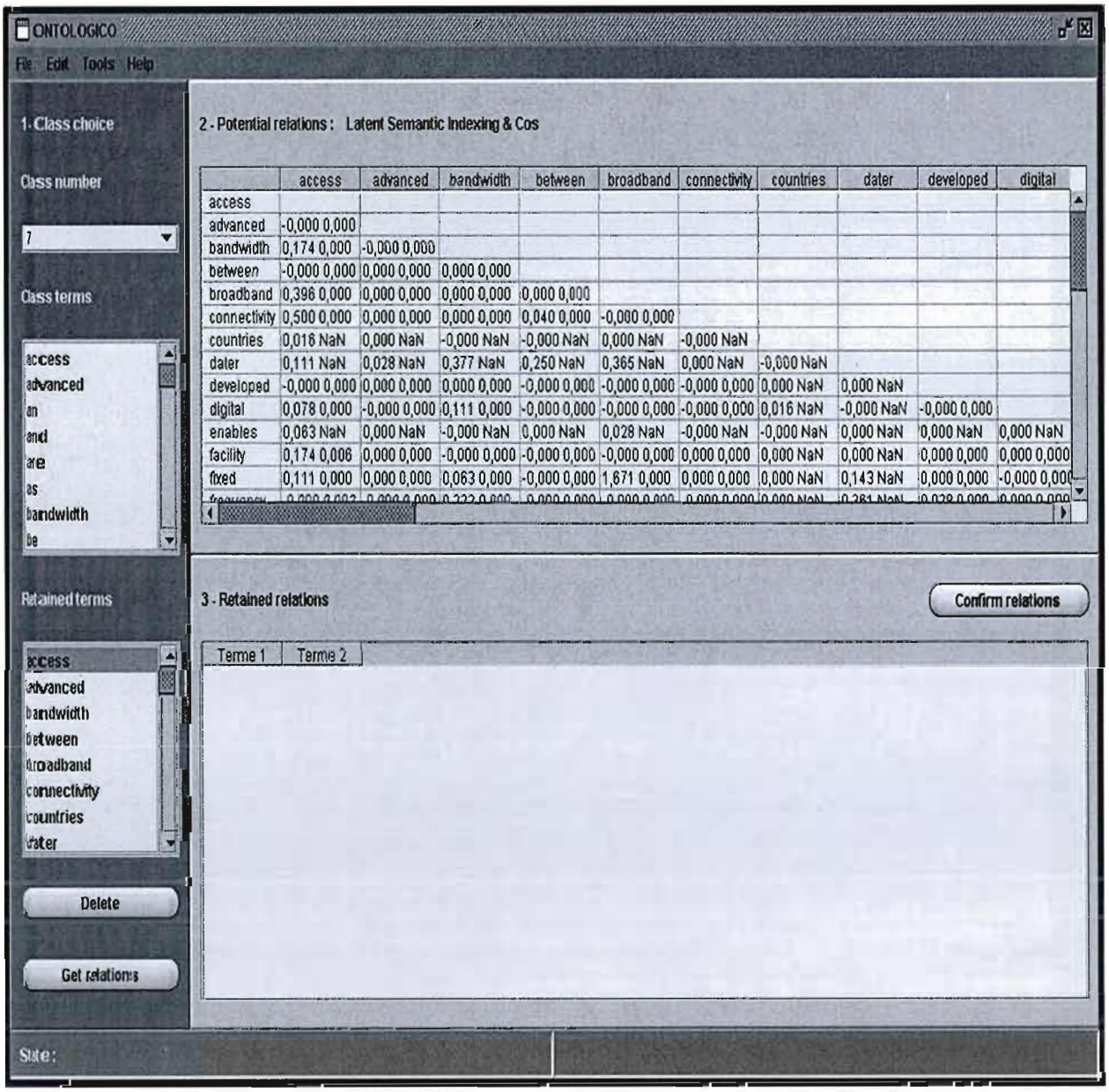
	spectrum, vs
33	american, analog, cellular, digital, latin, most, now, operators, primarily, quality, systems, using, voice
46	advances, broadband, enables, end, fixed, network, new, service, technique, technology, user, wireless
47	area, background, band, bands, capacity, common, communication, deployments, ghz, interference, lan, market, network, networks, propagation, shot, shown, systems, table, technology, unlicensed, wireless
48	access, actual, appendix, application, aps, areas, available, band, bandwidth, calculations, characteristics, comparison, conclusion, consideration, dater, dense, environments, equal, frequencies, frequency, ghz, gofdm, gpbcc, high, highly, important, including, intensive, mbps, modulation, network, power, propagation, provide, ranger, rat, require, signal, spectrum, support, table, technologie, technology, throughput, times, transmettre, user, wlan
51	benefit, characteristics, different, end, many, models, operating, usage, user, valoir
52	across, dater, dsss, gofdm, gpbcc, however, line, mbps, mhz, mpdu, networks, pbcc, rat, second
54	application, aps, area, areas, at, bandwidth, based, capacity, common, companies, consideration, constant, corresponding, costs, coverage, dater, distance, each, efficient, either, environment, environments, example, f, for, free, g, gofdm, greater, high, higher, in, its, likely, market, modulation, more, most, network, not, office, one, operating, operational, other, outside, over, overall, pbcc, rat, receiver, require, same, several, small, space, spectrum, standard, technologie, that, this, transmettre, two, users, using, variable, well, will, would]
56	all, also, an, and, band, bluetooth, by, commercial, conclusion, dater, distance, f, hills, how, however, important, in, including, interference, is, issu, it, its, legal, many, market, mhz, model, more, networks, not, one, only, operating, paper, phase, physics, potential, primarily, products, propagation, radio, ranger, rat, section, see, several, signal, significant, similar, simultaneous, small, still, systems, table, that, there, they, this, traffic, two, used, wave, when, with, would
57	a, and, based, because, become, costs, could, development, each, expensive, four, from, generally, has, innovation, is, market, next, process, spectrum, technical, that, the, thus, to, user, very, will, with
58	addition, an, and, application, are, band, be, believe, by, equipment, existing, good, in, many, models, more, potential, several, speed, standard, technologie, that, there, this, unlicensed, usage, we, will, wlan, worldwide
60	application, europe, hiperlan, radio, ranger, short, systems, wireless
63	an, and, are, as, at, be, because, believe, both, by, can, client, could, deployment, e, esn, have, high, hills, in, internet, issu, it, link, many, o, ofdm, out, power, provide, reilly, same, speed, system, technologie, that, those, transmettre, used, wave, we, who, wlan
64	a, all, appendix, as, b, be, calculations, corporation, devices, do, for, further, however, interference, lan, likely, market, mobilian, networks, not, power, reserved, rights, section, should, similar, support, that, the, this, throughput, using, will, wireless, wlan, world
67	about, again, and, any, application, aps, are, as, available, bands, bandwidth, become, been, between, both, by, characteristics, client, common, communication, companies, conclusion, considerations, cost, costs, current, deployment, deployments, devices, differences, do, driving, e, efficiency, environment, environments, expected, fact, g, given, good, greater, has, have, high, homerf, how, in, include, innovation,

	interference, large, legal, levels, likely, loss, lower, many, model, modulation, must, mw, network, not, office, operating, other, out, paper, path, potential, power, qos, significant, so, spectral, spectrum, standard, such, system, systems, technical, technique, technology, that, these, this, transmettre, transmitter, unlicensed, usage, valoir, we, well, what, while, who, will, with, wlan
69	common, deployments, likely, mhz, network, power, simultaneously, support, that, video, will
70	application, are, as, at, attractif, be, become, benefit, compared, db, every, extension, figure, given, higher, in, into, log, lx, more, new, only, per, ranger, shown, similar, three, user, with, would]
74	ack, average, backoff, dater, duration, micro, mpdu, overhead, ppdu, rat, sec, sifs

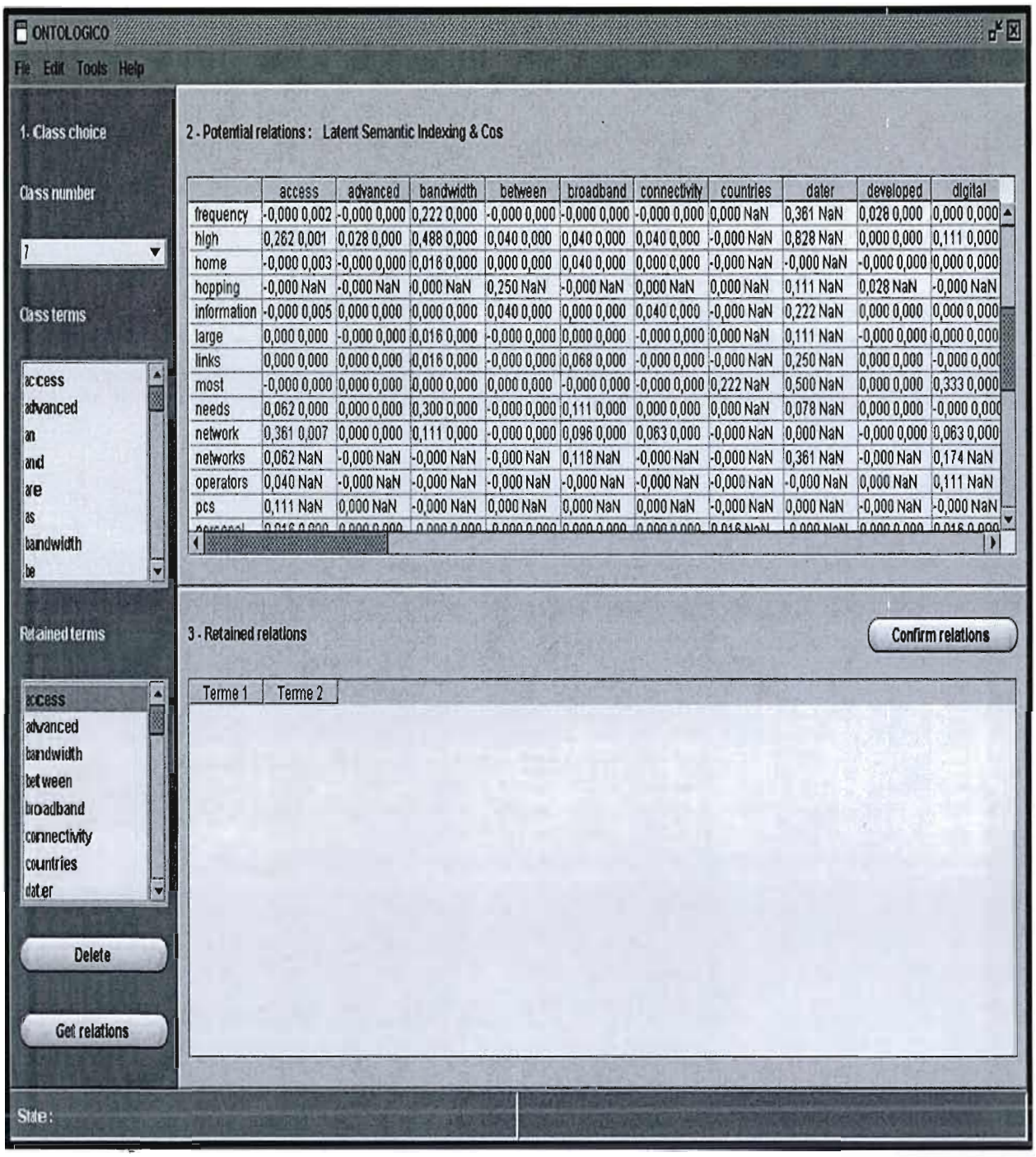
ANNEXE 4

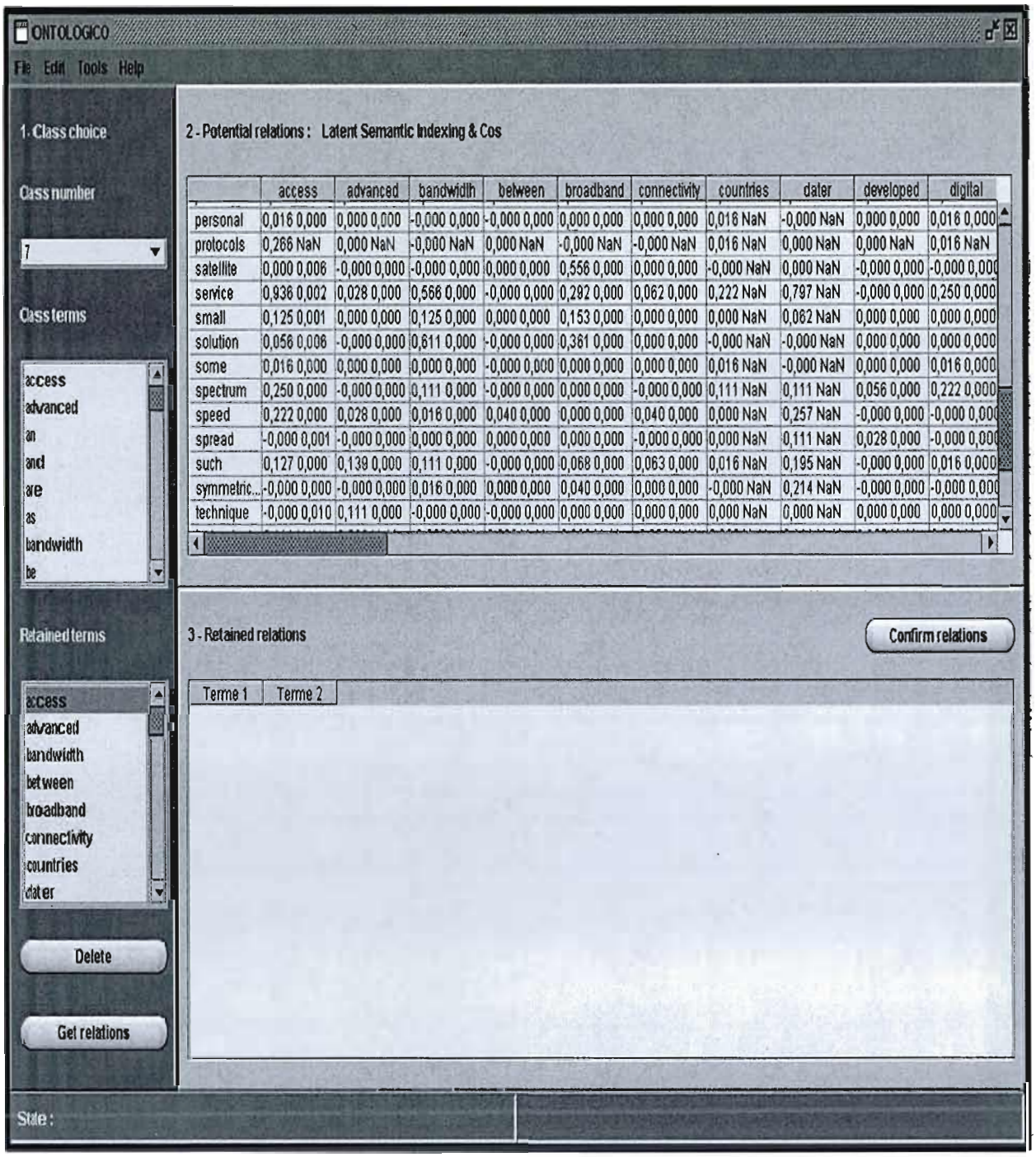
UN EXEMPLE GÉNÉRÉ PAR ONTOLOGICO (CLASSE DE TERMES N. 7)

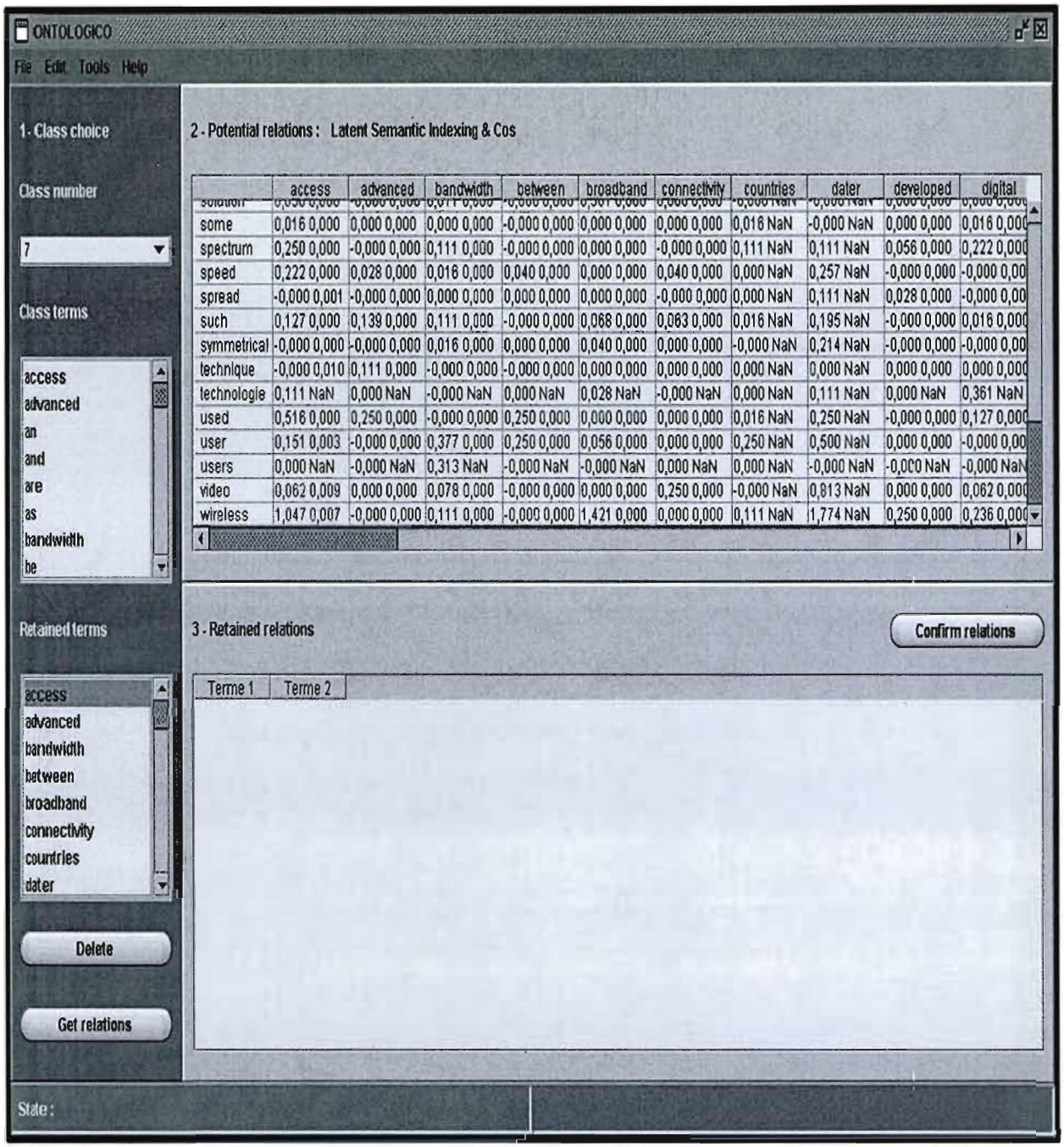
Sept captures d'écran sont présentées pour mieux visualiser les résultats du tableau croisé relatif aux relations potentielles à l'intérieur de la classe de termes N. 7. Les relations potentielles sont répertoriées dans le ***tableau 4.2***.

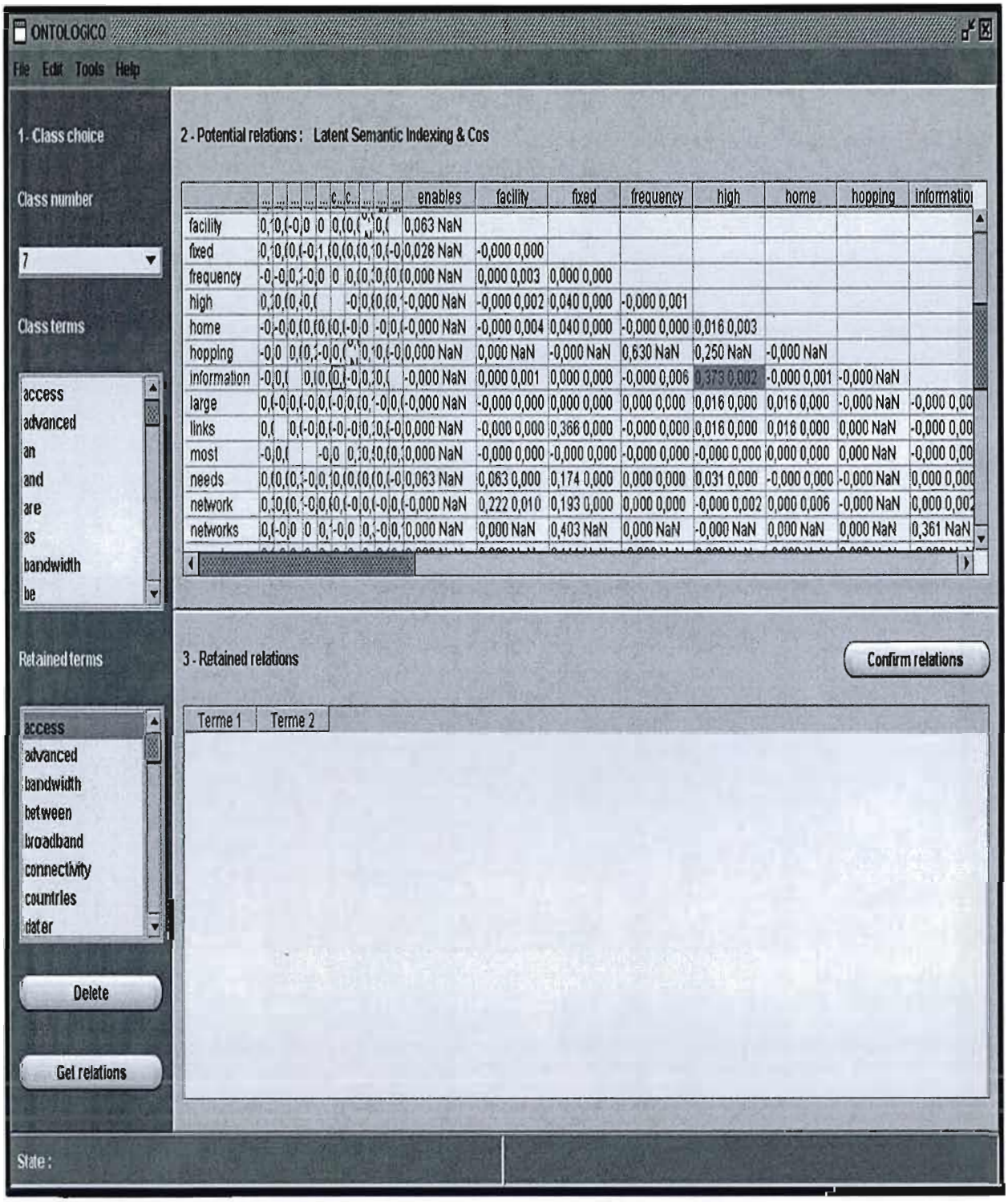


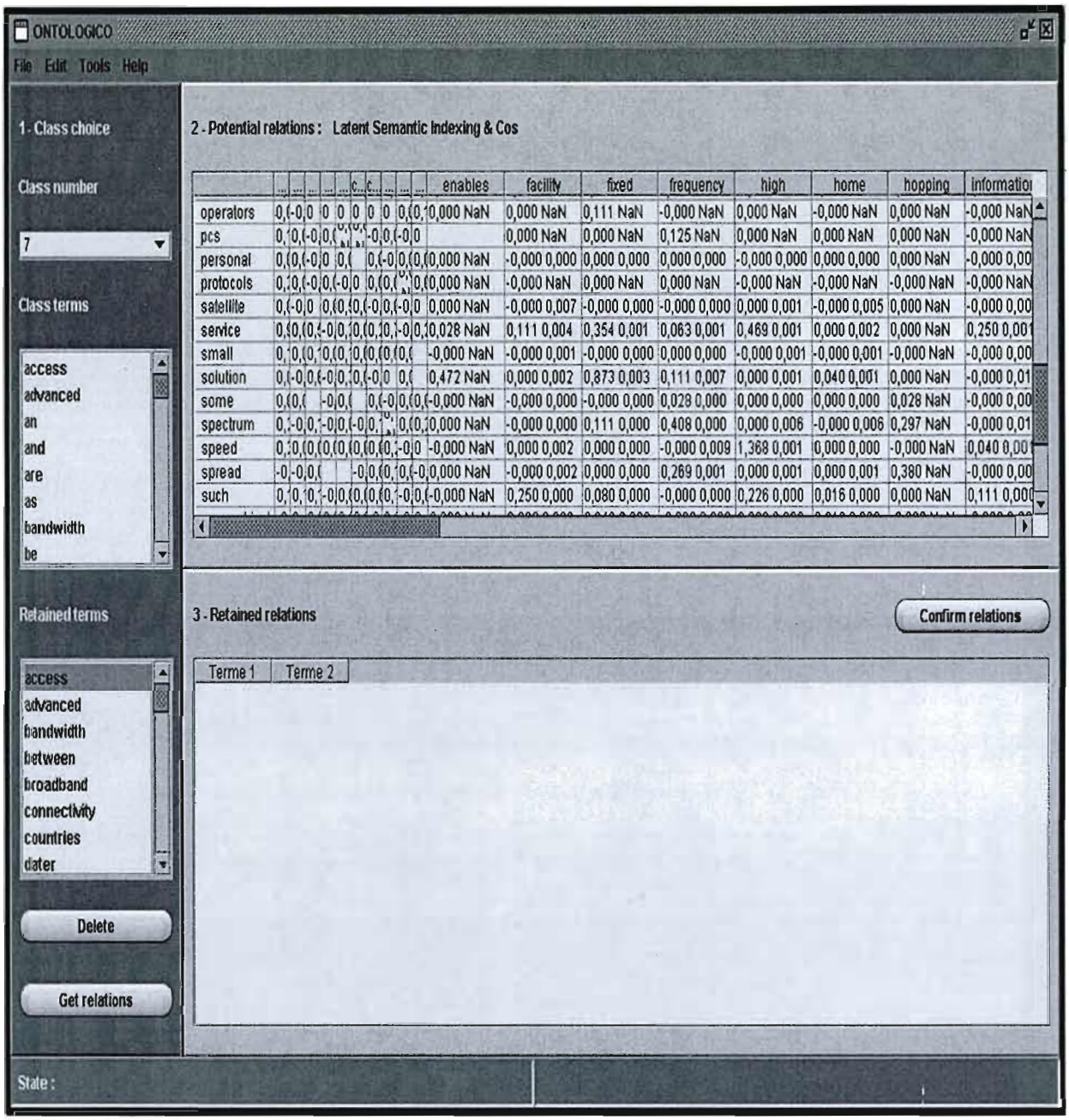
	access	advanced	bandwidth	between	broadband	connectivity	countries	dater	developed	digital
access										
advanced	-0,000 0,000									
bandwidth	0,174 0,000	-0,000 0,000								
between	-0,000 0,000	0,000 0,000	0,000 0,000							
broadband	0,396 0,000	0,000 0,000	0,000 0,000	0,000 0,000						
connectivity	0,500 0,000	0,000 0,000	0,000 0,000	0,040 0,000	-0,000 0,000					
countries	0,016 NaN	0,000 NaN	-0,000 NaN	-0,000 NaN	0,000 NaN	-0,000 NaN				
dater	0,111 NaN	0,028 NaN	0,377 NaN	0,250 NaN	0,365 NaN	0,000 NaN	-0,000 NaN			
developed	-0,000 0,000	0,000 0,000	0,000 0,000	-0,000 0,000	-0,000 0,000	-0,000 0,000	0,000 NaN	0,000 NaN		
digital	0,078 0,000	-0,000 0,000	0,111 0,000	-0,000 0,000	-0,000 0,000	-0,000 0,000	0,016 NaN	-0,000 NaN	-0,000 0,000	
enables	0,063 NaN	0,000 NaN	-0,000 NaN	0,000 NaN	0,028 NaN	-0,000 NaN	-0,000 NaN	0,000 NaN	0,000 NaN	0,000 NaN
facility	0,174 0,006	0,000 0,000	-0,000 0,000	-0,000 0,000	-0,000 0,000	0,000 0,000	0,000 NaN	0,000 NaN	0,000 0,000	0,000 0,000
fixed	0,111 0,000	0,000 0,000	0,063 0,000	-0,000 0,000	1,671 0,000	0,000 0,000	0,000 NaN	0,143 NaN	0,000 0,000	-0,000 0,000

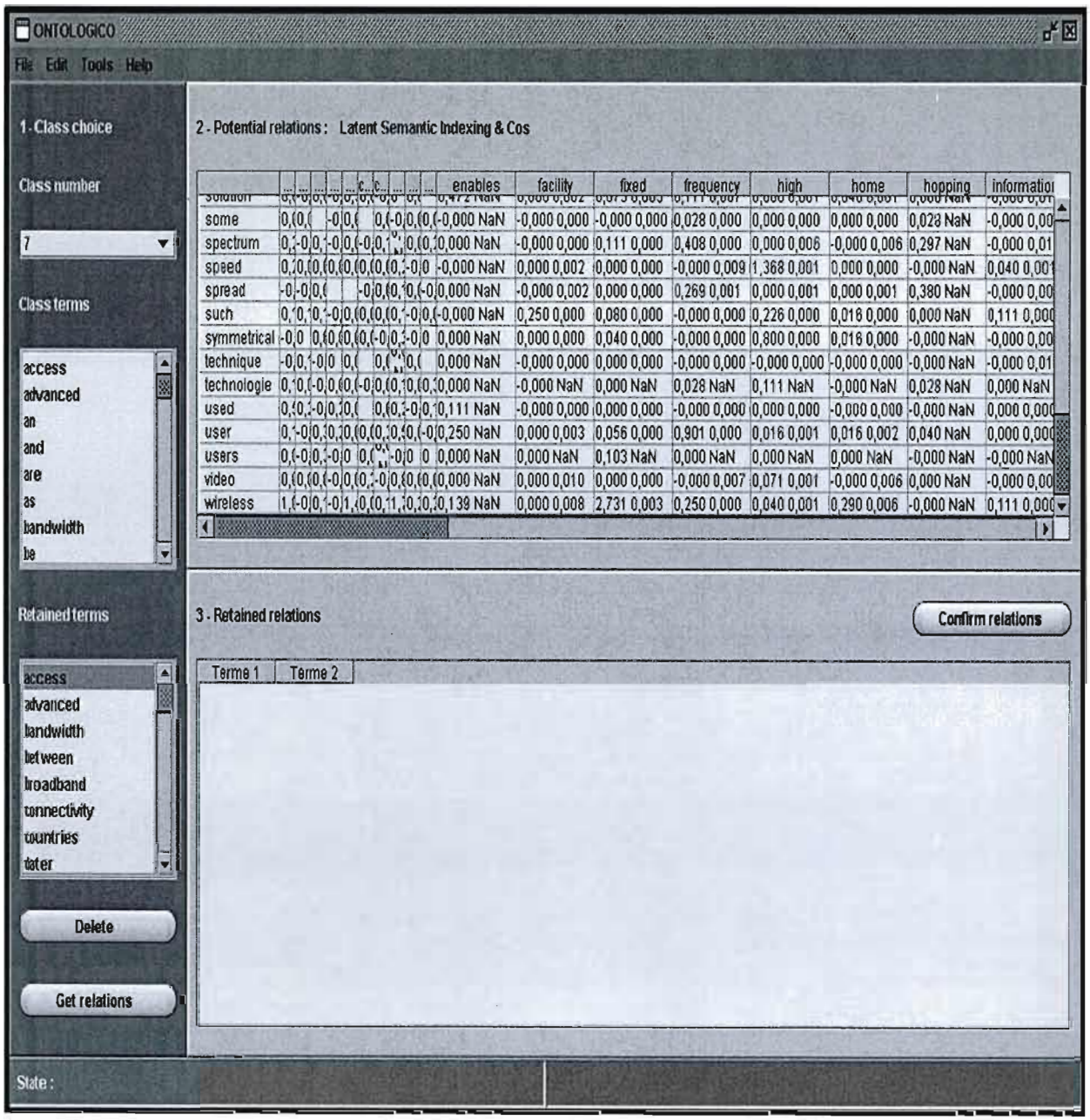

ONTOLOGICO
File Edit Tools Help
1- Class choice
Class number
7
Class terms
access
advanced
an
and
are
as
bandwidth
be
2 - Potential relations : Latent Semantic Indexing & Cos
access advanced bandwidth between broadband connectivity countries dater developed digital
frequency -0,000 0,002 -0,000 0,000 0,222 0,000 -0,000 0,000 -0,000 0,000 -0,000 0,000 0,000 NaN 0,361 NaN 0,028 0,000 0,000 0,000
high 0,262 0,001 0,028 0,000 0,488 0,000 0,040 0,000 0,040 0,000 0,040 0,000 -0,000 NaN 0,828 NaN 0,000 0,000 0,111 0,000
home -0,000 0,003 -0,000 0,000 0,016 0,000 0,000 0,000 0,040 0,000 0,000 0,000 -0,000 NaN -0,000 NaN -0,000 0,000 0,000 0,000
hopping -0,000 NaN -0,000 NaN 0,000 NaN 0,250 NaN -0,000 NaN 0,000 NaN 0,000 NaN 0,111 NaN 0,028 NaN -0,000 NaN
information -0,000 0,005 0,000 0,000 0,000 0,000 0,040 0,000 0,000 0,000 0,040 0,000 -0,000 NaN 0,222 NaN 0,000 0,000 0,000 0,000
large 0,000 0,000 -0,000 0,000 0,016 0,000 -0,000 0,000 0,000 0,000 -0,000 0,000 0,000 NaN 0,111 NaN -0,000 0,000 0,000 0,000
links 0,000 0,000 0,000 0,000 0,016 0,000 -0,000 0,000 0,068 0,000 -0,000 0,000 -0,000 NaN 0,250 NaN 0,000 0,000 -0,000 0,000
most -0,000 0,000 0,000 0,000 0,000 0,000 0,000 0,000 -0,000 0,000 -0,000 0,000 0,222 NaN 0,500 NaN 0,000 0,000 0,333 0,000
needs 0,062 0,000 0,000 0,000 0,300 0,000 -0,000 0,000 0,111 0,000 0,000 0,000 0,000 NaN 0,078 NaN 0,000 0,000 -0,000 0,000
network 0,361 0,007 0,000 0,000 0,111 0,000 -0,000 0,000 0,096 0,000 0,063 0,000 -0,000 NaN 0,000 NaN -0,000 0,000 0,063 0,000
networks 0,062 NaN -0,000 NaN -0,000 NaN -0,000 NaN 0,118 NaN -0,000 NaN -0,000 NaN 0,361 NaN -0,000 NaN 0,174 NaN
operators 0,040 NaN -0,000 NaN -0,000 NaN -0,000 NaN -0,000 NaN -0,000 NaN -0,000 NaN 0,000 NaN 0,111 NaN
pcs 0,111 NaN 0,000 NaN -0,000 NaN 0,000 NaN 0,000 NaN 0,000 NaN -0,000 NaN 0,000 NaN -0,000 NaN -0,000 NaN
Retained terms
access
advanced
bandwidth
between
broadband
connectivity
countries
dater
Delete
Get relations
3 - Retained relations
Confirm relations
Terme 1 Terme 2
State :


ONTOLOGICO
File Edit Tools Help
1- Class choice
Class number
7
Class terms
access
advanced
an
and
are
as
bandwidth
be
2 - Potential relations : Latent Semantic Indexing & Cos
access advanced bandwidth between broadband connectivity countries dater developed digital
personal 0,016 0,000 0,000 0,000 -0,000 0,000 -0,000 0,000 0,000 0,000 0,000 0,000 0,016 NaN -0,000 NaN 0,000 0,000 0,016 0,000
protocols 0,266 NaN 0,000 NaN -0,000 NaN 0,000 NaN -0,000 NaN -0,000 NaN 0,016 NaN 0,000 NaN 0,000 NaN 0,016 NaN
satellite 0,000 0,006 -0,000 0,000 -0,000 0,000 0,000 0,000 0,556 0,000 0,000 0,000 -0,000 NaN 0,000 NaN -0,000 0,000 -0,000 0,000
service 0,936 0,002 0,028 0,000 0,566 0,000 -0,000 0,000 0,282 0,000 0,062 0,000 0,222 NaN 0,797 NaN -0,000 0,000 0,250 0,000
small 0,125 0,001 0,000 0,000 0,125 0,000 0,000 0,000 0,153 0,000 0,000 0,000 0,000 NaN 0,062 NaN 0,000 0,000 0,000 0,000
solution 0,056 0,006 -0,000 0,000 0,611 0,000 -0,000 0,000 0,361 0,000 0,000 0,000 -0,000 NaN -0,000 NaN 0,000 0,000 0,000 0,000
some 0,016 0,000 0,000 0,000 0,000 0,000 -0,000 0,000 0,000 0,000 0,000 0,000 0,016 NaN -0,000 NaN 0,000 0,000 0,016 0,000
spectrum 0,250 0,000 -0,000 0,000 0,111 0,000 -0,000 0,000 0,000 0,000 -0,000 0,000 0,111 NaN 0,111 NaN 0,056 0,000 0,222 0,000
speed 0,222 0,000 0,028 0,000 0,016 0,000 0,040 0,000 0,000 0,000 0,040 0,000 0,000 NaN 0,257 NaN -0,000 0,000 -0,000 0,000
spread -0,000 0,001 -0,000 0,000 0,000 0,000 0,000 0,000 0,000 0,000 -0,000 0,000 0,000 NaN 0,111 NaN 0,028 0,000 -0,000 0,000
such 0,127 0,000 0,139 0,000 0,111 0,000 -0,000 0,000 0,068 0,000 0,063 0,000 0,016 NaN 0,195 NaN -0,000 0,000 0,016 0,000
symmetric... -0,000 0,000 -0,000 0,000 0,016 0,000 0,000 0,000 0,040 0,000 0,000 0,000 -0,000 NaN 0,214 NaN -0,000 0,000 -0,000 0,000
technique -0,000 0,010 0,111 0,000 -0,000 0,000 -0,000 0,000 0,000 0,000 0,000 0,000 0,000 NaN 0,000 NaN 0,000 0,000 0,000 0,000
Retained terms
access
advanced
bandwidth
between
broadband
connectivity
countries
dater
Delete
Get relations
3 - Retained relations
Confirm relations
Terme 1
Terme 2
State :


ONTOLOGICO
File Edit Tools Help
1- Class choice
Class number
7
Class terms
access
advanced
an
and
are
as
bandwidth
be
2 - Potential relations : Latent Semantic Indexing & Cos
access advanced bandwidth between broadband connectivity countries dater developed digital
some 0,016 0,000 0,000 0,000 0,000 0,000 -0,000 0,000 0,000 0,000 0,016 NaN -0,000 NaN 0,000 0,000 0,016 0,000
spectrum 0,250 0,000 -0,000 0,000 0,111 0,000 -0,000 0,000 0,000 0,000 -0,000 0,000 0,111 NaN 0,111 NaN 0,056 0,000 0,222 0,000
speed 0,222 0,000 0,028 0,000 0,016 0,000 0,040 0,000 0,000 0,000 0,040 0,000 0,000 NaN 0,257 NaN -0,000 0,000 -0,000 0,00
spread -0,000 0,001 -0,000 0,000 0,000 0,000 0,000 0,000 0,000 0,000 -0,000 0,000 0,000 NaN 0,111 NaN 0,028 0,000 -0,000 0,00
such 0,127 0,000 0,139 0,000 0,111 0,000 -0,000 0,000 0,068 0,000 0,063 0,000 0,016 NaN 0,195 NaN -0,000 0,000 0,016 0,000
symmetrical -0,000 0,000 -0,000 0,000 0,016 0,000 0,000 0,000 0,040 0,000 0,000 0,000 -0,000 NaN 0,214 NaN -0,000 0,000 -0,000 0,00
technique -0,000 0,010 0,111 0,000 -0,000 0,000 -0,000 0,000 0,000 0,000 0,000 0,000 0,000 NaN 0,000 NaN 0,000 0,000 0,000 0,000
technologie 0,111 NaN 0,000 NaN -0,000 NaN 0,000 NaN 0,028 NaN -0,000 NaN 0,000 NaN 0,111 NaN 0,000 NaN 0,361 NaN
used 0,516 0,000 0,250 0,000 -0,000 0,000 0,250 0,000 0,000 0,000 0,000 0,000 0,016 NaN 0,250 NaN -0,000 0,000 0,127 0,000
user 0,151 0,003 -0,000 0,000 0,377 0,000 0,250 0,000 0,056 0,000 0,000 0,000 0,250 NaN 0,500 NaN 0,000 0,000 -0,000 0,00
users 0,000 NaN -0,000 NaN 0,313 NaN -0,000 NaN -0,000 NaN 0,000 NaN 0,000 NaN -0,000 NaN -0,000 NaN -0,000 NaN
video 0,062 0,009 0,000 0,000 0,078 0,000 -0,000 0,000 0,000 0,000 0,250 0,000 -0,000 NaN 0,813 NaN 0,000 0,000 0,062 0,000
wireless 1,047 0,007 -0,000 0,000 0,111 0,000 -0,000 0,000 1,421 0,000 0,000 0,000 0,111 NaN 1,774 NaN 0,250 0,000 0,236 0,000
Retained terms
3 - Retained relations
Confirm relations
access
advanced
bandwidth
between
broadband
connectivity
countries
dater
Terme 1 Terme 2
Delete
Get relations
State :


ONTOLOGICO
File Edit Tools Help
1- Class choice
Class number
7
Class terms
access
advanced
an
and
are
as
bandwidth
be
2 - Potential relations : Latent Semantic Indexing & Cos
enables facility fixed frequency high home hopping
facility 0,063 NaN
fixed 0,028 NaN -0,000 0,000
frequency 0,000 NaN 0,000 0,003 0,000 0,000
high -0,000 NaN -0,000 0,002 0,040 0,000 -0,000 0,001
home -0,000 NaN -0,000 0,004 0,040 0,000 -0,000 0,000 0,016 0,003
hopping 0,000 NaN 0,000 NaN -0,000 NaN 0,630 NaN 0,250 NaN -0,000 NaN
information -0,000 NaN 0,000 0,001 0,000 0,000 -0,000 0,006 0,373 0,002 -0,000 0,001 -0,000 NaN
large -0,000 NaN -0,000 0,000 0,000 0,000 0,000 0,000 0,016 0,000 0,016 0,000 -0,000 NaN -0,000 0,00
links 0,000 NaN -0,000 0,000 0,366 0,000 -0,000 0,000 0,016 0,000 0,016 0,000 0,000 NaN -0,000 0,00
most 0,000 NaN -0,000 0,000 -0,000 0,000 -0,000 0,000 -0,000 0,000 0,000 0,000 0,000 NaN -0,000 0,00
needs 0,063 NaN 0,063 0,000 0,174 0,000 0,000 0,000 0,031 0,000 -0,000 0,000 -0,000 NaN 0,000 0,00
network -0,000 NaN 0,222 0,010 0,193 0,000 0,000 0,000 -0,000 0,002 0,000 0,006 -0,000 NaN 0,000 0,00
networks 0,000 NaN 0,000 NaN 0,403 NaN 0,000 NaN -0,000 NaN 0,000 NaN 0,000 NaN 0,361 NaN
Retained terms
3 - Retained relations
Confirm relations
Terme 1
Terme 2
access
advanced
bandwidth
between
broadband
connectivity
countries
data
Delete
Get relations
State :


ONTOLOGICO
File Edit Tools Help
1 - Class choice
Class number
7
Class terms
access
advanced
an
and
are
as
bandwidth
be
Retained terms
access
advanced
bandwidth
between
broadband
connectivity
countries
dater
Delete
Get relations
2 - Potential relations : Latent Semantic Indexing & Cos
enables
facility
fixed
frequency
high
home
hopping
operators
pcs
personal
protocols
satellite
service
small
solution
some
spectrum
speed
spread
such
3 - Retained relations
Confirm relations
Terme 1
Terme 2
State :


ONTOLOGICO
File Edit Tools Help
1 - Class choice
Class number
7
Class terms
access
advanced
an
and
are
as
bandwidth
be
2 - Potential relations : Latent Semantic Indexing & Cos
enables facility fixed frequency high home hopping
some 0,000 NaN -0,000 0,000 -0,000 0,000 0,028 0,000 0,000 0,000 0,000 0,000 0,028 NaN -0,000 0,00
spectrum 0,000 NaN -0,000 0,000 0,111 0,000 0,408 0,000 0,000 0,006 -0,000 0,006 0,297 NaN -0,000 0,01
speed -0,000 NaN 0,000 0,002 0,000 0,000 -0,000 0,009 1,368 0,001 0,000 0,000 -0,000 NaN 0,040 0,001
spread 0,000 NaN -0,000 0,002 0,000 0,000 0,269 0,001 0,000 0,001 0,000 0,001 0,380 NaN -0,000 0,00
such -0,000 NaN 0,250 0,000 0,080 0,000 -0,000 0,000 0,226 0,000 0,018 0,000 0,000 NaN 0,111 0,000
symmetrical 0,000 NaN 0,000 0,000 0,040 0,000 -0,000 0,000 0,800 0,000 0,016 0,000 -0,000 NaN -0,000 0,00
technique 0,000 NaN -0,000 0,000 0,000 0,000 -0,000 0,000 -0,000 0,000 -0,000 0,000 -0,000 NaN -0,000 0,01
technologie 0,000 NaN -0,000 NaN 0,000 NaN 0,028 NaN 0,111 NaN -0,000 NaN 0,028 NaN 0,000 NaN
used 0,111 NaN -0,000 0,000 0,000 0,000 -0,000 0,000 0,000 0,000 -0,000 0,000 -0,000 NaN 0,000 0,000
user 0,250 NaN 0,000 0,003 0,056 0,000 0,901 0,000 0,016 0,001 0,016 0,002 0,040 NaN 0,000 0,000
users 0,000 NaN 0,000 NaN 0,103 NaN 0,000 NaN 0,000 NaN 0,000 NaN -0,000 NaN -0,000 NaN
video 0,000 NaN -0,000 0,010 0,000 0,000 -0,000 0,007 0,071 0,001 -0,000 0,006 0,000 NaN -0,000 0,00
wireless 0,139 NaN 0,000 0,008 2,731 0,003 0,250 0,000 0,040 0,001 0,290 0,006 -0,000 NaN 0,111 0,000
Retained terms
3 - Retained relations
Confirm relations
Terme 1
Terme 2
access
advanced
bandwidth
between
broadband
connectivity
countries
later
Delete
Get relations
State :

ANNEXE 5

L'ALGORITHME DVS

1. Algorithme SVD de Fierro - séquence principale des appels

L'algorithme de SVD de Fierro est une séquence d'algorithmes détaillées comme suit :

SVD de Fierro sur $(A^TA)_{n\times n}$ pour les k plus grandes valeurs propres :

Retourne $U_k D_k V_k^T = \bar{A}_k \cong (A^TA)_{n\times n}$

1. **BLCR de Lanczos** sur $(A^TA)_{n\times n}$ avec $2k$ itérations.
   Retourne $Q_{n\times 2k}^{(L)}$ et $T_{2k\times 2k} = tri\text{-}diag(\{\alpha_1, \alpha_2, \ldots, \alpha_{2k}\}, \{\beta_1, \beta_2, \ldots, \beta_{2k-1}\})$

   (La procédure BLCR est appelée sur la matrice A. Cette procédure s'exécute implicitement sur A^TA, mais le produit n'est jamais formé explicitement).

2. **Calcul des paires de Ritz** sur $T_{2k\times 2k}$ et $Q_{n\times 2k}^{(L)}$ pour les vecteurs de Ritz
   Retourne les r valeurs de Ritz $\{\theta_1, \ldots, \theta_r\}$ qui ont convergé et les vecteurs de Ritz correspondant $\{y_1, \ldots, y_r\}_{n\times r}$.

   a) **Tri-diagonalisation de Householder** sur $T_{2k\times 2k}$
   Retourne $T'_{2k\times 2k} = (Q_{2k\times 2k})^T T_{2k\times 2k} Q^T_{2k\times 2k}$ et $Q_{2k\times 2k}$

   b) **Rotation de Wilkinson** sur $D_{22} = T'_{2k\times 2k}$ au départ
   Diagonalise $T'_{2k\times 2k}$ en $D_{2k\times 2k}$.
   Les transformations sont accumulées dans $Q_{2k\times 2k}^{(H\text{-}W)}$.
   Retourne $D_{2k\times 2k}$ et $Q_{2k\times 2k}$.

   c) **Déterminer les *r* valeurs de Ritz convergées**
   $\{\theta_1, \ldots, \theta_r\} \leftarrow$ r plus grands éléments de la diagonale de $D_{2k\times 2k}$ qui ont convergé selon le critère de convergence de Fierro

   d) **Calculer les *r* vecteurs de Ritz correspondants** à partir des vecteurs propres de $T_{2k\times 2k}$ et des vecteurs de Lanczos $Q_{n\times 2k}^{(L)}$

$V_{n\times r} \leftarrow \{y_1, \ldots, y_r\}_{n\times r} = Q_{n\times 2k}^{(L)}\ X_{2k\times r}$ où $X_{2k\times r} \leftarrow r$ colonnes de $Q_{2k\times 2k}^{(H\text{-}W)}$ correspondant aux r valeurs de Ritz déterminées en c)

3. **BLCR de Lanczos** sur $(A_r)_{n\times n} \leftarrow (I_{n\times n} - V_{n\times r}\ (V_{n\times r})^T)\ (A^TA)_{n\times n}\ (I_{n\times n} - V_{n\times r}\ (V_{n\times r})^T)$ avec $2k\text{-}r$ itérations.

   Retourne $Q_{n\times(2k-r)}^{(L)}$ et $T_{(2k-r)\times(2k-r)} = tri\text{-}diag(\{\alpha_1, \alpha_2, \ldots, \alpha_{(2k-r)}\}, \{\beta_1, \beta_2, \ldots, \beta_{(2k-r)-1}\})$

   (La procédure BLCR doit être appelée sur $A_{m\times n}\ (I_{n\times n} - V_{n\times r}\ (V_{n\times r})^T)$ de façon à s'exécuter implicitement sur le produit $[I_{n\times n} - V_{n\times r}\ (V_{n\times r})^T]^T\ (A^TA)_{n\times n}\ [I_{n\times n} - V_{n\times r}\ (V_{n\times r})^T]$, où $[I_{n\times n} - V_{n\times r}\ (V_{n\times r})^T]^T = [I_{n\times n} - V_{n\times r}\ (V_{n\times r})^T]$ car $V_{n\times r}(V_{n\times r})^T$ est symétrique.)

4. **Calcul des paires de Ritz** sur $T_{(2k-r)\times(2k-r)}$ et $Q_{n\times(2k-r)}^{(L)}$ pour les vecteurs de Ritz

   Retourne $\{\theta_1, \ldots, \theta_k\} \leftarrow k$ plus grands éléments de la diagonale de $D_{(2k-r)\times(2k-r)}$ et $V_{n\times k} \leftarrow \{y_1, \ldots, y_k\}_{n\times k} = Q_{n\times(2k-r)}^{(L)}\ X_{(2k-r)\times k}$ où $X_{(2k-r)\times k} \leftarrow k$ colonnes correspondantes de $Q_{(2k-r)\times(2k-r)}^{(H\text{-}W)}$.

   Les r paires de Ritz convergées à la première itération de l'algorithme SVD de Fierro sont conservées. À chaque itération subséquente, on rajoute les paires des valeurs de Ritz nouvellement convergées, jusqu'à ce qu'on obtienne k paires de Ritz dont les valeurs ont convergé.

5. **BLCR de Lanczos** sur $(A_k)_{n\times n} \leftarrow (I_{n\times n} - V_{n\times k}\ (V_{n\times k})^T)\ (A^TA)_{n\times n}\ (I_{n\times n} - V_{n\times k}\ (V_{n\times k})^T)$ avec k itérations.

   Retourne $Q_{n\times k}^{(L)}$ et $T_{k\times k} = tri\text{-}diag(\{\alpha_1, \alpha_2, \ldots, \alpha_k\}, \{\beta_1, \beta_2, \ldots, \beta_{k-1}\})$

6. **Calcul des paires de Ritz** sur $T_{k\times k}$ et $Q_{n\times k}^{(L)}$ pour les vecteurs de Ritz

   Retourne $\{\theta_1, \ldots, \theta_k\} \leftarrow k$ éléments de la diagonale de $D_{k\times k}$ et $V_{n\times k} \leftarrow \{y_1, \ldots, y_k\}_{n\times k} = Q_{n\times k}^{(L)}\ Q_{k\times k}^{(H\text{-}W)}$.

   À cette étape, on compare toute paire de Ritz nouvellement convergée avec les paires déjà conservées. On retient les paires qui correspondent aux plus grandes valeurs de Ritz. Les étapes 5 et 6 sont exécutées jusqu'à ce qu'aucune valeur de

Ritz nouvellement convergée ne soit plus grande que les valeurs conservées à date, pendant deux itérations consécutives.

7. **QR 'thin'** sur $A_{m\times n}\ V_{n\times k}$

   Retourne les matrices Q_k et R_k tronquées suivantes : $Q_{m\times m} \rightarrow Q_{m\times k}$ et $R_{m\times k} \rightarrow R_{k\times k}$

8. **SVD de Golub-Kahan** sur $R_{k\times k}$

   Retourne $U_R = U_{k\times k}$, $D_R = D_{k\times k}$ et $V_R = V_{k\times k}$

9. **Approximations finales**

   a) $U'_k = Q_{m\times k}\ U_{k\times k} = U'_{m\times k}$

   b) $D'_k = D_{k\times k}$

   c) $V'_k = V_{n\times k}\ V_{k\times k}$

2. Algorithme BLCR ("Basic Lanczos with Complete Reorthogonalisation")

L'algorithme BLCR (Fierro and al, 2005) (Golub and al, 1996) factorise partiellement la matrice $A^T A$ en $Q_{n\times p}\ T_{p\times p}\ (R_{n\times p})^T$ où $T_{p\times p}$ est une matrice tri-diagonale formée des éléments $\{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ sur sa diagonale principale et $\{\beta_1, \beta_2, \ldots, \beta_{p-1}\}$ sur les diagonales secondaires. La matrice $Q_{n\times p}$ est formée des vecteurs de Lanczos $\{q_1, q_2, \ldots, q_p\}$ et la matrice $R_{n\times p}$ est formée des vecteurs $\{r_1, r_2, \ldots, r_p\}$.

1. Initialisations :

   a) $r_0 \leftarrow q_1 = (1, 0, \ldots, 0)$

   b) $\beta_0 \leftarrow 1$

   c) $q_0 \leftarrow 0$

2. Pour $i = 1$ à n :

   a) $q_i \leftarrow r_{i-1} / \beta_{i-1}$

   b) $w \leftarrow A^T (A\ q_i)$

   c) $r \leftarrow w - \beta_{i-1}\ q_{i-1}$

   d) $\alpha_i \leftarrow q_i^T\ r$

   e) $r_i \leftarrow w - \alpha_i\ q_i$

   f) $r_i \leftarrow r_i - Q_p\ (Q_p^T\ r_i)$ réorthogonalisation où $Q_p = \{q_1, q_2, \ldots, q_p\}$

g) $\beta_i \leftarrow \| r_i \|$

3. Algorithme SVD de Golub-Kahan :

Cet algorithme (Golub and al, 1996) factorise $A_{m\times n}$ $(m > n)$ en $(U_{m\times m})^T A_{m\times n} V_{n\times n} = D_{m\times n} + E$.

1. Calculer la **bi-diagonale de Householder** $B = U^T A V$ dans :

$$\begin{pmatrix} B_{n\times n} \\ 0_{m-n\times n} \end{pmatrix} \leftarrow \begin{pmatrix} U_1 & \cdots & U_n \end{pmatrix}^T_{m\times m} A_{m\times n} \begin{pmatrix} V_1 & \cdots & V_{n-2} \end{pmatrix}_{n\times n}$$

2. Jusqu'à ce que $q = n$:

   a) Si $|b_{i,i+1}| \leq \varepsilon (|b_{ii}| + |b_{i+1,i+1}|)$, alors $b_{i,i+1} \leftarrow 0$, $\forall\, i = 1$ à n-1. $(10^{-4} \leq \varepsilon \leq 10^{-3})$

   (Un $\varepsilon < 10^{-4}$ risque de faire apparaître des zéros dans la superdiagonale de B_{22}).

   b) Trouver $q_{\max}$ et $p_{\min}$ tel que

$$B = \begin{pmatrix} B_{11} & 0 & 0 \\ 0 & B_{22} & 0 \\ 0 & 0 & B_{33} \end{pmatrix} \begin{matrix} p \\ n-p-q \\ q \end{matrix}$$
$$\begin{matrix} p & n-p-q & q \end{matrix}$$

   B_{33} est diagonale et B_{22} n'a aucun zéro sur sa superdiagonale.

   c) Si $q < n$:

      i) S'il y a des éléments à zéro dans la diagonale de B_{22}, mettre leur élément de la superdiagonale de la même rangée à zéro, i.e. si $b_{i,i} = 0$, alors $b_{i,i+1} \leftarrow 0$, $\forall\, i = 1$ à n_{22}-1.

      ii) Autrement :

         (1) $T \leftarrow B_{22}^T B_{22}$

         (2) $\mu \leftarrow$ valeur propre de $T\,[n_{22}\text{-}1 : n_{22}\,;\, n_{22}\text{-}1 : n_{22}]$ la plus proche de $T\,[n_{22}\,;\, n_{22}]$:

            (a) $T_{22} \leftarrow T\,[n_{22}\text{-}1 : n_{22}\,;\, n_{22}\text{-}1 : n_{22}]^T\, T\,[n_{22}\text{-}1 : n_{22}\,;\, n_{22}\text{-}1 : n_{22}]$

(b) $d = (t_{11} - t_{22}) / 2$

(c) $\lambda_2^2 = t_{22} - \frac{t_{21}^2}{\left(d + sign(d)\sqrt{d^2 + t_{21}^2}\right)}$

(d) $\lambda_1^2 = t_{11} + t_{22} - \lambda_2^2$

(e) $\mu \leftarrow (\lambda_1 \text{ ou } \lambda_2)$ selon $\min_{i=1,2}\left(abs\left(T_{n22,n22} - \sqrt{\lambda_i^2}\right)\right)$

(3) $x = t_{11} - \mu$

(4) $z = t_{12}$

(5) Pour $k = 1$ à $n_{22} - 1$:

(a) $[c, s] =$ **Givens**(x, z) (voir Wilkinson, étape 9)

(b) $B_{22} = B_{22}\, G(k, k+1, \theta)$ (voir Wilkinson, étape 9)

(c) $x = b_{kk}$

(d) $z = b_{k+1,k}$

(e) $[c, s] =$ **Givens**(x, z) (voir Wilkinson, étape 9)

(f) $B_{22} = G(k, k+1, \theta)^T B_{22}$ (voir Wilkinson, étape 9)

(g) Si $k < n_{22} - 1$:

(i) $x = b_{k,k+1}$

(ii) $z = b_{k,k+2}$

(6) $B \leftarrow (I_p, U, I_{q+m-n})^T B\, (I_p, V, I_q)$

3. Accumuler les facteurs de Householder :

a) $U \leftarrow \text{house}U \cdot U$

b) $V \leftarrow \text{house}V \cdot V$

4. Algorithme de factorisation QR 'thin'

La factorisation QR 'thin' est une factorisation $Q_{m\times m}\ R_{m\times n}$ d'une matrice $m \times n$ ($m \geq n$) où on retient seulement les n premières rangées de R (les autres rangées sont nulles puisque R est une matrice triangulaire supérieure) et les n premières colonnes de Q (Doornik and al, 2002). Il s'agit donc de factoriser la matrice $A_{m\times n}\hat{V}_{k\times k}$ par $Q'_{m\times k}\ R'_{k\times k}$ où $A_{m\times n}$ est la matrice termes-documents initiale. L'algorithme Householder QR avec pivotage en colonne est utilisé pour effectuer cette factorisation (Golub and al, 1996). Cet algorithme calcule le rang($A_{m\times n}$) = r et factorise A en $A\Pi$= QR (Π= pivot(*)).

1. Pour $j = 1$ à n:

   $c(j) = A(1{:}m, j)^{\mathrm{T}}\, A(1{:}m, j)$

2. $r = 0$
3. $\tau = \max\{c(1),\ldots, c(n)\}$
4. $k = \min_{k=1:n;c(k)=\tau}(k)$
5. Tant que $\tau > 0$:
   a) $r = r + 1$
   b) pivot(r) = k
   c) $A(1{:}m, r) \leftrightarrow A(1{:}m, k)$
   d) $c(r) \leftrightarrow c(k)$
   e) $[v,\beta]$ = **House**($A(r{:}m, r)$)
   f) $A(r{:}m, r{:}n) = (I_{m-r+1} - \beta\, v\, v^{\mathrm{T}})\, A(r{:}m, r{:}n)$
   g) $A(r{+}1{:}m, r) \leftarrow v(2{:}m{-}r{+}1)$
   h) Pour $i = r + 1$ à n :

      $c(i) = c(i) - A(r,i)^2$

   i) Si $r < n$:

      i) $\tau = \max\{c(r{+}1),\ldots, c(n)\}$

      ii) $k = \min_{k=r+1:n;c(k)=\tau}(k)$

      Autrement :

      $\tau = 0$

L'étape g) $A(r+1{:}m,\ r) \leftarrow v(2{:}m\text{-}r+1)$ sert à accumuler les vecteurs de Householder afin de pouvoir récupérer les produits des matrices de Householder $H_1 \cdot H_2 \cdot \ldots \cdot H_n = Q$. Ces vecteurs sont commodément stokés dans le triangle inférieur de la matrice résultante R. Si la matrice Q est explicitement désirée, on peut remplacer l'étape g) par la suivante, après avoir initialisé la matrice $Q = I_n$:

g) $\quad Q \leftarrow Q \cdot [(I_{r\text{-}1}), (I_{m\text{-}r+1} - \beta\, v\, v^{\mathrm{T}})]$

où $I_{r\text{-}1}$ est le complément supérieur et gauche de $A_{m\text{-}r+1}$ pour obtenir une matrice carrée de dimension $m \times m$, i.e.

$$\begin{bmatrix} I_{r-1} & 0 \\ 0 & A_{m-r+1} \end{bmatrix}_{m\times m} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1_{r-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{11} & \cdots & a_{1,m-r+1} \\ 0 & 0 & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & a_{m-r+1,1} & \cdots & a_{\substack{m-r+1,\\ m-r+1}} \end{bmatrix}_{m\times m}$$

5. Calcul du vecteur de Householder (fonction House)

La fonction ***House()*** est un algorithme qui calcule le vecteur de Householder $[v,\beta]$ à partir d'un vecteur x (Golub and al, 1996):

1. $n = \text{length}(x)$
2. $\beta = x(2{:}n)^{\mathrm{T}}\, x(2{:}n)$
3. $v = \begin{pmatrix} 1 \\ x(2:n) \end{pmatrix}$
4. Si $\sigma = 0$:

   $\beta = 0$

   Autrement :

   a) $\quad \mu = \sqrt{x(1)^2 + \sigma}$

   b) $\quad$ Si $x(1) \le 0$:

   $v(1) = x(1) - \mu$

   Autrement :

   $v(1) = -\sigma / (x(1) + \mu)$

c) $\beta = 2\, v(1)^2 / (\sigma + v(1)^2)$

d) $v = v / v(1)$

6. Calcul des paires de Ritz

Le critère de convergence des pairs de Ritz (Fierro and al, 2005) est :

$$|\beta_p| \cdot |x_{j,p}| < TOL \cdot \hat{\sigma}_1 \sqrt{\theta_j} \sqrt{k} \qquad \text{où}$$

β_p est le dernier élément de la superdiagonale de la matrice tri-diagonale T_p générée par l'algorithme BLCR;

$x_{j,p}$ est le dernier élément du $j^{\text{ième}}$ vecteur propre de la matrice tri-diagonale T_p;

(Pour obtenir les $x_{j,p}$, il faut appliquer un algorithme direct de SVD à la matrice T_p et calculer ainsi les vecteurs propres de gauches de T_p. Nous utilisons l'algorithme de Golub-Kahan à cette fin).

$\hat{\sigma}_1$ est la racine carrée de la plus grande valeur de Ritz convergée à date;

θ_j est la $j^{\text{ième}}$ valeur de Ritz pour laquelle on évalue la convergence;

k est le nombre de valeurs propres recherchées;

TOL est le seuil de tolérance fonction de l'erreur machine, généralement 10^{-4} ou 10^{-5}.

L'algorithme peut s'articuler sur une décomposition de Schur (Golub and al, 1996) où

$$Z^T(Q_1{}^T A Q_1)Z = \text{diag}(\theta_1, \ldots, \theta_r)$$

donne les valeurs de Ritz. Les vecteurs de Ritz correspondants sont alors obtenus par

$$Q_1 Z = [y_1, \ldots, y_r]$$

Il s'agit de trouver les valeurs propres et les vecteurs propres de la matrice symétrique tri-diagonale T_p obtenue dans le processus de factorisation de Lanczos. Cette matrice est une hermitienne dense de dimensions réduites à $p << m$. On peut donc utiliser n'importe quel algorithme SVD direct de base pour la résoudre. Ici on a utilisé l'algorithme SVD de Golub-Kahan (Golub and al, 1996).

Les valeurs propres de la matrice tri-diagonale T_p ($p \times p$) obtenue à partir de $A^T A$ sont des approximations du carré de p des valeurs propres de A. Ces valeurs propres de T_p se nomment les valeurs de Ritz θ_j (j =1 à p) de la matrice $A^T A$. Les vecteurs propres de T_p, dénotés x_j (j =1 à p; $p \times 1$) sont en relation directe avec les vecteurs de Ritz y_j ($n \times 1$) = $Q_p\, x_j$ (($n \times p$) · ($p \times 1$)), j =1 à p, où $Q_p = \{q_1, q_2, \ldots, q_p\}$ sont les vecteurs de Lanczos générés par le processus de tri-diagonalisation de $A^T A$.

β_p et $x_{j,\,p}$ (le dernier élément des vecteurs x_j) sont typiquement utilisés pour déterminer la convergence des valeurs de Ritz θ_j vers des approximations des valeurs propres de $A^T A$. Les vecteurs de Ritz y_j sont typiquement utilisés pour redémarrer la factorisation de Lanczos.

7. Bi-diagonale de Householder

Cet algorithme (Golub and al, 1996) remplace $A_{m \times n}$ ($m > n$) par $B_{m \times n} = (U_{m \times m})^T\, A_{m \times n}\, V_{n \times n}$.

Pour j = 1 à n :

1. $[v, \beta]$ = **house**$(A(j{:}m, j))$
2. $A(j{:}m, j{:}n) = (I_{m-j+1} - \beta v v^T)\, A(j{:}m, j{:}n)$
3. $A(j{+}1{:}m, j) = v(2{:}m{-}j{+}1)$
4. Si $j \leq n$-2 :
   a) $[v, \beta]$ = **house**$(A(j, j{+}1{:}n)^T)$
   b) $A(j{:}m, j{+}1{:}n) = A(j{:}m, j{+}1{:}n)\, (I_{n-j} - \beta v v^T)$
   c) $A(j\ , j{+}2{:}n) = v(2{:}n{-}j)^T$

L'étape 3 sert à stoker les facteurs de U dans la partie inférieure gauche de la matrice B. Elle peut être remplacée par l'étape suivante pour former U explicitement, après avoir initialisé U à I_m :

3. $U \leftarrow U \cdot [(I_{j-1}), (I_{m-j+1} - \beta\, v\, v^T)]$

L'étape 4.c) sert à stoker les facteurs de V dans la partie supérieure droite de la matrice B. Elle peut être remplacée par l'étape suivante pour former V explicitement, après avoir initialisé V à I_n:

c) $V \leftarrow V \cdot [(I_j), (I_{n-j} - \beta v v^T)]$

8. Tri-diagonale de Householder

Cet algorithme (Golub and al, 1996) remplace une matrice symétrique $A_{n\times n}$ par $T = Q^T A Q$.

Pour $k = 1$ à n-2 :

1. $[v, \beta] = \mathbf{House}(A(k+1{:}n, k))$
2. $p \leftarrow \beta A(k+1{:}n, k+1{:}n)\, v$
3. $w \leftarrow p - (\beta p^T v/2)\, v$
4. $A(k+1, k) \leftarrow \|A(k+1{:}n, k)\|_2$
5. $A(k, k+1) \leftarrow A(k+1, k)$
6. $A(k+1{:}n, k+1{:}n) \leftarrow A(k+1{:}n, k+1{:}n) - vw^T - wv^T$
7. $A(k+2{:}n, k) \leftarrow$ zéros[31]
8. $A(k, k+2{:}n) \leftarrow$ zéros[5]

La matrice Q peut être explicitement formée en ajoutant l'étape suivante, après avoir initialisé Q à I_n :

9. $Q \leftarrow Q \cdot [(I_j), (I_{n-j} - \beta v v^T)]$

9. Étape QR symétrique implicite avec rotation de Wilkinson

À partir d'une matrice tri-diagonale T, cet algorithme remplace la matrice T par $Z^T T Z$ où $Z = G_1 \cdot G_2 \cdots G_{n-1}$ est un produit de rotations de Givens.

1. $d = (t_{n-1,n-1} - t_{n,n}) / 2$
2. $\mu = t_{n,n} - \dfrac{t_{n,n-1}^2}{\left(d + sign(d)\sqrt{d^2 + t_{n,n-1}^2}\right)}$

[31] Les étapes 7 et 8 ont été ajoutées pour mettre à zéro les éléments hors de la diagonale et de la première sous-diagonale.

3. $x = t_{11} - \mu$

4. $z = t_{21}$

5. Pour $k = 1$ à n-1 :

   a) $[c, s] =$ **Givens**(x, z)

   b) $T = G_k^{\mathrm{T}} T G_k$ où

$$G_k = G(k, k+1, \theta) = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \iddots & \vdots \\ 0 & \cdots & c & s & \cdots & 0 \\ 0 & \cdots & -s & c & \cdots & 0 \\ \vdots & \iddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ k \\ k+1 \\ \\ \\ \end{matrix}$$
$$\begin{matrix} & & k & k+1 & & \end{matrix}$$

(voir note plus bas)

   c) Si $k < n$-1 :

      i) $x = t_{k+1,\,k}$

      ii) $z = t_{k+2,\,k}$

Puisque T est une matrice tri-diagonale, il n'est pas nécessaire d'effectuer les produits de Givens sur toute la matrice. Avec

$G_1 = G_{2\times 2}$, $T_1 = T\,[k{:}k{+}1;\ k{:}k{+}1]$,

$T_2 = T\,[k{-}1;\ k{:}k{+}1]$, $T_3 = T\,[k{+}2;\ k{:}k{+}1]$,

$T_4 = T\,[k{:}k{+}1;k{-}1]$ et $T_5 = T\,[k{:}k{+}1;\ k{+}2]$,

$$T = \begin{bmatrix} \ddots & \ddots & \vdots & \iddots & \iddots \\ \ddots & \ddots & [T_2] & \iddots & \iddots \\ \cdots & [T_4] & [T_1] & [T_5] & \cdots \\ \iddots & \iddots & [T_3] & \ddots & \ddots \\ \iddots & \iddots & \vdots & \ddots & \ddots \end{bmatrix} \begin{matrix} \\ k-1 \\ k:k+1 \\ k+2 \\ \\ \end{matrix}$$

Il suffit d'effectuer les opérations suivantes :

$$T_1 = G_1^T T_1 G_1$$
$$T_2 = T_2 G_1$$
$$T_3 = T_3 G_1$$
$$T_4 = G_1^T T_4 = T_2^T$$
$$T_5 = G_1^T T_5 = T_3^T$$

Les autres éléments de la matrice T sont inchangés.

10. Calcul d'une rotation de Givens (fonction Givens)

Cet algorithme calcule c = cosinus(θ) et s = sinus(θ) à partir du vecteur $(a \;\; b)^T$ tel que

$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$. Cette fonction retourne $[c,s]$ = **Givens**$(a\ b)$. (Golub and al, 1996)

1. Si $b = 0$:
   a) $c = 1$
   b) $s = 0$
2. Autrement :
   a) Si $|b| > |a|$
      i) $s = 1/\sqrt{1+(a/b)^2}$
      ii) $c = (-a/b)/\sqrt{1+(a/b)^2}$
   b) Autrement
      i) $c = 1/\sqrt{1+(b/a)^2}$
      ii) $s = (-b/a)/\sqrt{1+(b/a)^2}$

RÉFÉRENCES

Abecker, A., Bernardi, A., Hinkelmann, K., Kuhn, O. and Sintek, M. (1998) *Toward a Technology for Organizational Memories*. Ieee Intelligent Systems & Their Applications 13(3): 40-48.

Achaba, H. (2003) *Système De Diffusion Documentaire Basé Sur Des Ontologies (Compétences Des Utilisateurs Et Connaissances Du Domaine)*. Université du Québec à Montréal.

Agrawal, R., Imielinski, T. and Swami, A. (1993) *Mining Association Rules between Sets of Items in Large Databases*. |SIGMOD Record 22(2): 207-16.

Aguirre, E., Ansa, O., Hovy, E. and Martinez, D. (2000) *Enriching Very Large Ontologies Using the WWW*. Proceedings of the Workshop on Ontology Construction of the European Conference of AI(ECAI-00).

Alfonseca, E. and Manandhar, S. (2002) *Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures*. Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web. 13th International Conference, EKAW 2002. Lecture Notes in Artificial Intelligence Vol.2473: 1-7|xi+402.

Alfonseca, E. and Manandhar, S. (2002) *An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery*. Proceedings of the 1st International Conference on General WordNet. Mysore, India.

Alfonseca, E. and Rodríguez, P. (2002) *Automatically Generating Hypermedia Documents Depending on User Goals*. Workshop on Document Compression and Synthesis in Adaptive Hypermedia Systems, AH-2002. Málaga, Spain.

Ambroziak, J. R. (1997). *Conceptual Assisted Web Browsing*. Sun Technical Report 61.

Arpírez, J., Gómez-Pérez, A., Lozano, A. and Pinto, S. (1998) *An Ontology-Based Www Broker to Select Ontologies*. Proceedings of the ECAI-98 Workshop on Applications of Ontologies and PSMs, Brighton. England: 16-24.

Arpirez, J. C., Corcho, O., Fernandez-Lopez, M. and Gomez-Perez, A. (2001) *Webode: A Scalable Workbench for Ontological Engineering*. Proceedings of the First International Conference on Knowledge Capture: 6-13.

Aussenac-Gilles, N., Biebow, B. and Szulman, S. (2000) *Revisiting Ontology Design: A Method Based on Corpus Analysis*. Knowledge Engineering and Knowledge Management Methods, Models, and Tools. 12th International Conference, EKAW 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1937)|Knowledge Engineering and Knowledge Management Methods, Models, and Tools. 12th International Conference, EKAW 2000. Proceedings (Lecture Notes in Artificial

Intelligence Vol.1937): 172-88|xiii+456.

Bachimont, B. (1999). *L'intelligence Artificielle Comme Écriture Dynamique : De La Raison Graphique À La Raison Computationnelle*. Paris, Grasset.

Bachimont, B. (2000) *Engagement Sémantique Et Engagement Ontologique: Conception Et Réalisation D'ontologies En Ingénierie Des Connaissances*. Ingénierie des Connaissances : Evolutions récentes et nouveaux défis. Eyrolles.

Bachimont, B., Isaac, A. and Troncy, R. (2002) *Semantic Commitment for Designing Ontologies: A Proposal*. Knowledge Engineering and Knowledge Management, Proceedings 2473: 114-121.

Bechhofer, S., Horrocks, I., Goble, C. and Stevens, R. (2001) *Oiled: A Reasonable Ontology Editor for the Semantic Web*. Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence. Vienna 2174: 396--408.

Beguin, A., Jouis, C. and Mustafa, W. (1997) *Évaluation D'outils D'aide À La Construction De Terminologie Et De Relations Sémantiques Entre Termes À Partir De Corpus*. Premières Journées Scientifiques et Techniques (JST) du réseau Francophone de l'ingénierie de langue de l'AUPELF-UREF. Avignon: 419-425.

Benhadid, I., Meunier, J. G., Hamidi, S., Remaki, Z. and Nyongwa, M. (1998) *Étude Expérimentale Comparative Des Méthodes Statistiques Pour La Classification Des Données Textuelles*. Actes de JADT-98, Nice, France.

Benjamins, V. R., Fensel, D., Decker, S. and Gomez Perez, A. (1999) *Building Ontologies for the Internet: A Mid Term Report*. International Journal of Human-Computer Studies 51: 687-712.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) *The Semantic Web - a New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities*. Scientific American 284(5): 34-+.

Berry, M. W., Dumais, S. T. and Obrien, G. W. (1995) *Using Linear Algebra for Intelligent Information Retrieval*. Siam Review 37(4): 573-595.

Bessé, B. D. (1990) *La Définition Terminologique*. Paris, Larousse.

Biebow, B. and Szulman, S. (1999) *Terminae: A Linguistic-Based Tool for the Building of a Domain Ontology*. Knowledge Acquisition, Modeling and Management. 11th European Workshop, EKAW'99. Proceedings|Knowledge Acquisition, Modeling and Management. 11th European Workshop, EKAW'99. Proceedings: 49-66|xi+404.

Biemann, C. (2005) *Ontology Learning from Text: A Survey of Methods*. LDV Forum Vol.20, No. 2: 75-93.

Binet, J., Dierickx, J. and Funck-Brentano, J. (1987) *Le Français, Langue Des Sciences Et Des Techniques*. Luxembourg.

Biskri, I. and Delisle, S. (1999) *Un Modèle Hybride Pour Le Textual Data Mining : Un Mariage De Raison Entre Le Numérique Et Le Linguistique*. TALN 99, France: p. 55-64.

Biskri, I. and Meunier, J. G. (2002) *Satim : Système D'analyse Et De Traitement De L'information Multidimensionnelle*. Proceedings of JADT 2002, St-Malo, France, p. 185-196.

Bisson, G., Nedellec, C. and Cañamero, D. (2000) *Designing Clustering Methods for Ontology Building*. Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence, ECAI'00, Berlin, Germany, p. 20-25.

Blazquez, M., Fernandez, M., Garcia-Pinar, J. M. and Gomez-Perez, A. (1998) *Building Ontologies at the Knowledge Level Using the Ontology Design Environment.* Proceedings of the Banff Workshop on Knowledge Acquisition for Knowledge-based Systems.

Borst, P., Akkermans, H. and Top, J. (1997) *Engineering Ontologies*. International Journal of Human-Computer Studies 46(2-3): 365-406.

Bouaud, J., Bachimont, B., Charlet, J. and Zweigenbaum, P. (1994) *Acquisition and Structuring of an Ontology within Conceptual Graphs*. Proceedings of the International Conference on Conceptual Structures (ICCS'94): p.1-25.

Bourigault, D. (1994) *Lexter, Un Logiciel D'extraction De Terminologie, Application À L'acquisition Des Connaissances À Partir De Textes*. Ecole des Hautes Etudes en Sciences sociales, Paris

Bourigault, D. (*2002*) *Upery : Un Outil D'analyse Distributionnelle Étendue Pour La Construction D'ontologies À Partir De Corpus*. Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy.

Bourigault, D. and Aussenac-Gilles, N. (2003) *Construction D'ontologies À Partir De Textes.* Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2003), Batz-sur-Mer, T2: pp. 27-50.

Bozsak, E., Ehrig, M., Handschuh, S., Hotho, A., Maedche, A., Motik, B., Oberle, D., Schmitz, C., Staab, S., Stojanovic, L., Stojanovic, N., Studer, R., Stumme, G., Sure, Y., Tane, J., Volz, R. and Zacharias, V. (2002) *Kaon - Towards a Large Scale Semantic Web*. E-Commerce and Web Technologies, Proceedings 2455: 304-313.

Brachman, R. J. and Schmolze, J. G. (1985) *An Overview of the Kl-One Knowledge Representation System*. Cognitive Science 9(2): 171-216.

Brill, E. (1993) *Automatic Grammar Induction and Parsing Free-Text - a Transformation-Based Approach.* 31st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference: 259-265.

Brill, E. (1994) *Some Advances in Transformation-Based Part of Speech Tagging.* Proceedings of the Twelfth National Conference on Artificial Intelligence, Vols 1 and 2: 722-727.

Briscoe, T. and Carroll, J. (1993) *Generalized Probabilistic Lr Parsing of Natural Language (Corpora) with Unification-Based Grammars.* Computational Linguistics|Computational Linguistics 19(1): 25-59.

Brown, P. F., Della Pietra, V., deSouza, P. V., Mercer, R. L. and Lai, J. C. (1992) *Class-Based N-Gram Models of Natural Language.* Computational Linguistics|Computational Linguistics 18(4): 467-79.

Brunet, E. (2002) *Le Lemme Comme on L'aime.* Actes des 6ièmes Journées internationales d'Analyse statistique des Données Textuelles (JADT). Saint-Malo, 13-15 mars 2002, vol. 1, pp. 221-232. Saint-Malo : IRISA/INRIA.

Callan, J. P. (1994) *Passage-Level Evidence in Document Retrieval.* SIGIR '94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval|SIGIR '94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval: 302-10|358.

Carpenter, G. A. and Grossberg, S. (1987) *Art 2: Self-Organization of Stable Category Recognition Codes for Analog Input Patterns.* IEEE First International Conference on Neural Networks|IEEE First International Conference on Neural Networks: 727-35 vol.2|4 vol. (234+820+790+846).

Carpenter, G. A. and Grossberg, S. (1987) *A Massively Parallel Architecture for a Self-Organizing Neural Pattern-Recognition Machine.* Computer Vision Graphics and Image Processing 37(1): 54-115.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H. and Rosen, D. B. (1992) *Fuzzy Artmap - a Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps.* Ieee Transactions on Neural Networks 3(5): 698-713.

Carpenter, G. A., Grossberg, S. and Reynolds, J. H. (1991) *Artmap - Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network.* Neural Networks 4(5): 565-588.

Carpenter, G. A., Grossberg, S. and Rosen, D. B. (1991) *Fuzzy Art - Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System.* Neural Networks 4(6): 759-771.

Ceusters, W., Martens, P., Dhaen, C. and Terzic, B. (2001) *Linkfactory: An Advanced Formal Ontology Management System*. Proceedings of Interactive Tools for Knowledge Capture, KCAP-2001, Victoria, October 20.

Chabbat, B., Pinon, J. M. and Ou-Halima, M. (1995) *Hypertexte Sémantique Pour L'aide À La Décision*. Ingénierie des systèmes d'informations, Volume 3.

Chomsky, N. (1957) *Syntactic Structures*. The Hague, Mouton & co.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. MIT-Press.

Church, K., Gale, W., Hanks, P. and Hindle, D. (1989) *Parsing, Word Associations and Typical Predicate-Argument Relations*. International Workshop on Parsing Technologies|International Workshop on Parsing Technologies: 389-98|vii+467.

Clark, A. (1997) *Being There : Putting Brain Body and World Together Again*. Cambridge UP.

Daille, B. (1994) *Approche Mixte Pour L'extraction De Terminologie: Statistique Lexicale Et Filtres Linguistiques*. Thèse d'Informatique. Université de Paris VII.

Darwiche, A. and Marquis, P. (2002) *A Knowledge Compilation Map*. Journal of Artificial Intelligence Research 17: 229-264.

David, S. a. P., P. (1990) *Termino Version 1.0*. Rapport de recherche du Centre d'Analyse de Textes par Ordinateurs, Université du Québec à Montréal.

De Chalendar, G. and Grau, B. (2000) *Svetlan' or How to Classify Words Using Their Context*. Knowledge Engineering and Knowledge Management Methods, Models, and Tools. 12th International Conference, EKAW 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1937)|Knowledge Engineering and Knowledge Management Methods, Models, and Tools. 12th International Conference, EKAW 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1937): 203-16|xiii+456.

Decker, S., Erdmann, M., Fensel, D. and Studer, R. (1999) *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information*. Database Semantics. Semantic Issues in Multimedia Systems. IFIP TC2/WG2.6 Eighth Working Conference on Database Semantics (DS-8)|Database Semantics. Semantic Issues in Multimedia Systems. IFIP TC2/WG2.6 Eighth Working Conference on Database Semantics (DS-8): 351-69|xi+456.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990) *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science 41(6): 391-407.

Desclés, J.-P. (1990) *Langages Applicatifs, Langues Naturelles Et Cognition*. Hermès, Paris.

Dijk, T. V. (1977) *Text and Context*. London: Longman.

Domingue, J. (1998) *Tadzebao and Webonto: Discussing, Browsing and Editing Ontologies on the Web*. Proceedings of the Eleventh Knowledge Acquisition Workshop, KAW98, Banff.

Doornik, J. A. and O'Brien, R. J. (2002) *Numerically Stable Cointegration Analysis*. Computational Statistics & Data Analysis 41(1): 185-193.

Dowding, J., Moore, R., Andry, F. and Moran, D. (1994) *Interleaving Syntax and Semantics in an Efficient Bottom-up Parser*. 32nd Annual Meeting of the Association for Computational Linguistics: 110-116.

Duc, T. H. (2003) *Développer Un Système De Diffusion Active Et De Recherche Intelligente*. Rapport de stage. Projet GDST (Gestion et Diffusion du Savoir en Télécommunication). Dirigé par Bernard Lefebvre.

Dunning, T. (1993) *Accurate Methods for the Statistics of Surprise and Coincidence*. Computational Linguistics|Computational Linguistics 19(1): 61-74.

Engels, R. (2001) *Corporum-Ontoextract. Ontology Extraction.* Deliverable 6 Ontoknowledge. http://www.ontonowledge.org/del.shtml

Eriksson, H., Fergerson, R. W., Shahar, Y. and Musen, M. A. (1999) *Automated Generation of Ontology Editors*. Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management (KAW99), Banff, Alberta, Canada, October 16-21.

Eriksson, H., Puerta, A. R. and Musen, M. A. (1994) *Generation of Knowledge-Acquisition Tools from Domain Ontologies*. International Journal of Human-Computer Studies 41(3): 425-453.

Faatz, A., Kamps, T. and Steinmetz, R. (2000) *Background Knowledge, Indexing and Matching Interdependencies of Document Management and Ontology-Maintenance*. ECAI Workshop on Ontology Learning.

Faatz, A., Seeberg, C. and Steinmetz, R. (2002) *Statistical Profiles of Words for Ontology Enrichment*. Soft Methods in Probability, Statistics and Data Analysis: 295-301.

Farquhar, A., Fikes, R. and Rice, J. (1997) *The Ontolingua Server: A Tool for Collaborative Ontology Construction*. International Journal of Human-Computer Studies 46(6): 707-727.

Faure, D. and Nedellec, C. (1999) *Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System Asium*. Knowledge Acquisition, Modeling and Management. 11th European Workshop, EKAW'99. Proceedings|Knowledge Acquisition, Modeling and Management. 11th European

Workshop, EKAW'99. Proceedings: 329-34|xi+404.

Faure, D. a. P., T. (2000) *First Experiments of Using Semantic Knowledge Learned by Asium for Information Extraction Task Using Intex.* Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany

Fensel, D. (2000) *Problem-Solving Methods - Understanding, Description, Development, and Reuse - Introduction.* Problem-Solving Methods: Understanding, Description, Development, and Reuse 1791: 1-+.

Fensel, D. (2001) *Ontologies, a Silver Bullet for Knowledge Management & Electronic Commerce*. Springer-Verlag publishers.

Fernandez, M., Gomez-Perez, A. and Juristo, N. (1997) *Methontology: From Ontological Art Towards Ontological Engineering*. Proceedings of the Spring Symposium Series on Ontological Engineering (AAAI'97).

Ferre, S. and Ridoux, O. (2004) *Introduction to Logical Information Systems*. Information Processing & Management 40(3): 383-419.

Fierro, R. D. and Jiang, E. P. (2005) *Lanczos and the Riemannian Svd in Information Retrieval Applications*. Numerical Linear Algebra with Applications 12(4): 355-372.

Firth, J. R. (1957) *A Synopsis of Linguistic Theory*. J.R. Firth et al. Studies in Linguistic Analysis. Special volume of the Philological Society. Oxford: Blackwell.

Fodor, J. A. and Pylyshyn, Z. W. (1988) *Connectionism and Cognitive Architecture - a Critical Analysis*. Cognition 28(1-2): 3-71.

Frakes, W. B. and Baeza-Yates, R. (1992) *Information Retrieval : Data Structures and Algorithms*. Englewood Cliffs : Prentice-Hall.

Gadamer, H. G. (1976) *Vérité Et Méthode*. Édition du Seuil, Paris.

Gal, A., Modica, G. and Jamil, H. (2004) *Ontobuilder: Fully Automatic Extraction and Consolidation of Ontologies from Web Sources*. 20th International Conference on Data Engineering, Proceedings: 853-853.

Ganter, B. and Wille, R. (1999) *Formal Concept Analysis: Mathematical Foundations*. . Springer, Berlin-Heidelberg.

Gargouri, Y., Lefebvre, B. and Meunier, J. G. (2003) *Maintenance Des Ontologies À Partir D'analyses Textuelles.* ACFAS'2003, Rimouski, Québec, 21 Mai 2003.

Gargouri, Y., Lefebvre, B. and Meunier, J. G. (2003) *Ontology Maintenance Using Textual Analysis*. 7th World Multiconference on Systemics, Cybernetics and Informatics, Vol

I, Proceedings: 248-253.

Gargouri, Y., Lefebvre, B. and Meunier, J. G. (2004) *Ontologico : Vers Un Outil D'assistance Au Développement Itératif Des Ontologies*. Journées d'études sur Terminologie, Ontologie, et Représentation des connaissances (TERMINO'2004). Lyon, France.

Gargouri, Y., Lefebvre, B. and Meunier, J. G. (2005) *Domain and Competences Ontologies and Their Maintenance for an Intelligent Dissemination of Documents*. MICAI 2005: Advances in Artificial Intelligence. 4th Mexican International Conference on Artificial Intelligence. Proceedings (Lecture Notes in Artificial Intelligence Vol.3789): 90-7|xxvi+1198.

Gelbukh, A., Sidorov, G. and Guzmán-Arenas, A. (1999) *Text Categorization Using a Hierarchical Topic Dictionary*. Proc. Text Mining workshop at 16th International Joint Conference on Artificial Intelligence (IJCAI'99), Stockholm, Sweden, July 31 - August 6, pp. 34-35.

Godin, R., Missaoui, R. and April, A. (1993) *Experimental Comparison of Navigation in a Galois Lattice with Conventional Information-Retrieval Methods*. International Journal of Man-Machine Studies 38(5): 747-767.

Golub, G. H. and Van Loan, C. F. (1996) *Matrix Computations*. Johns Hopkins University Press, ISBN 0-8018-5414-8.

Gomez-Perez, A. (1999) *Evaluation of Taxonomic Knowledge in Ontologies and Knowledge Bases*. Proceedings of the 12th Workhop on Knowledge Acquisition, Modeling and Management (KAW'99).

Gomez-Perez, A., Fernandez, M. and De Vicente, A. J. (1996) *Towards a Method to Conceptualize Domain Ontologies*. In ECAI-96 Workshop on Ontological Engineering, Budapest.

Grefenstette, G. (1992) *Use of Syntactic Context to Produce Term Association Lists for Text Retrieval*. SIGIR Forum|SIGIR Forum spec. issue.: 89-97.

Grossberg, S. (1988) *Neural Network and Natural Intelligence*. Cambridge: MIT Press.

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

Gruninger, M. and Fox, M. S. (1995) *Methodology for the Design and Evaluation of Ontologies* IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, August 19-20th.

Guarino, N. (1995) *Formal Ontology, Conceptual Analysis and Knowledge Representation*. International Journal of Human-Computer Studies 43(5-6): 625-640.

Guarino, N., Masolo, C. and Vetere, G. (1999) *Ontoseek: Content-Based Access to the Web*. Ieee Intelligent Systems & Their Applications 14(3): 70-80.

Guarino, N. and Welty, C. (2000) *Identity, Unity and Individuality: Towards a Formal Toolkit for Ontological Analysis*. ECAI 2000. 14th European Conference on Artificial Intelligence. including Prestigious Applications of Intelligent Systems (PAIS-2000). Proceedings (Frontiers in Artificial Intelligence and Applications Vol.54)|ECAI 2000. 14th European Conference on Artificial Intelligence. including Prestigious Applications of Intelligent Systems (PAIS-2000). Proceedings (Frontiers in Artificial Intelligence and Applications Vol.54): 219-23|xvi+778.

Gupta, K. M., Aha, D. W., Marsh, E. and Maney, T. (2002) *An Architecture for Engineering Sublanguage Wordnets*. Proceedings of the First International Conference On Global WordNet (pp. 207-215). Mysore, India: Central Institute of Indian Languages.

Hahn, U. and Marko, K. G. (2001) *Joint Knowledge Capture for Grammars and Ontologies*. Proceedings of the First International Conference on Knowledge Capture: 68-75|x+209.

Hahn, U. and Schnattinger, K. (1998) *Towards Text Knowledge Engineering*. Proceedings Fifteenth National Conference on Artificial Intelligence (AAAI-98). Tenth Conference on Innovative Applications of Artificial Intelligence|Proceedings Fifteenth National Conference on Artificial Intelligence (AAAI-98). Tenth Conference on Innovative Applications of Artificial Intelligence: 524-31|xxiv+1218.

Hahn, U. and Schulz, S. (2000) *Towards Very Large Terminological Knowledge Bases: A Case Study from Medicine*. Advances in Artificial Intelligence. 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1822)|Advances in Artificial Intelligence. 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1822): 176-86|xii+450.

Hans, K. T., Apitz, R., Lattner, A. D. and Schlieder, C. (2001) *Knowwork - an Approach to Co-Ordinate Knowledge within Technical Sales, Design and Process Planning Departments*. Proceedings of the 7th International Conference on Concurrent Enterprising, pages 231-239, Bremen, Germany.

Harnad, S. (1990) *The Symbol Grounding Problem*. Physica D 42(1-3): 335-346.

Harris, Z. S. (1968) *Mathematical Structures of Language*. Wiley & Sons, New York, USA.

Hearst, M. A. (1992) *Automatic Acquisition of Hyponyms from Large Text Corpora*. Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France, July 1992.

Hirst, G. and St-Onge, D. (1998) *Lexical Chains as Representations of Context for the*

Detection and Correction of Malapropisms. In Fellbaum, C. (dir. publ.). Wordnet: an electronic lexical database. Cambridge (Mass.): MIT Press, pp. 305-332.

Hogue, S. (2003) *Réalisation D'un Environnement D'exploration Adapté Aux Usagers Pour Une Ontologie Dans Le Domaine Des Télécommunications Sans Fils*. Mémoire de maîtrise en informatique, Université du Québec à Montréal. Dirigé par Bernard Lefebvre.

Hung, H. H. (2003) *Développement D'un Outil Efficace Pour Annoter Des Documents*. Mémoire de fin d'études. Institut de la Francophonie pour l'Informatique d'Hanoï. Dirigé par Bernard Lefebvre de l'Université du Québec à Montréal.

Hwang, C. H. (1999) *Incompletely and Imprecisely Speaking: Using Dynamic Ontologies for Representing and Retrieving Information*. In Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99), Linköping, Sweden.

Ibekwe-San, J. F., Condamines, A. and Cabré, M. T. (2005) *Application-Driven Terminology Engineering* Special issue of Terminology 11-1.

Ikeda, M., Hayashi, Y., Lai, J., Chen, W., Bourdeau, J., Seta, K. and Mizoguchi, R. (1999) *An Ontology More Than a Shared Vocabulary*. AI-ED 99. Workshop on Ontologies for Intelligent Educational Systems, Le Mans, France.

Ikeda, M., Seta, K. and Mizoguchi, R. (1997) *Task Ontology Makes It Easier to Use Authoring Tools*. Ijcai-97 - Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, Vols 1 and 2: 342-347.

Jalabert, F. and Lafourcade, M. (2004) *Nommage De Sens À L'aide Des Vecteurs Conceptuels*. Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle, RFIA'2004, Toulouse, France.

Jin, L., Chen, W., Hayashi, Y., Ikeda, M., Mizoguchi, R., Takaoka, Y. and Ohta, M. (1999) *An Ontology-Aware Authoring Tool - Functional Structure and Guidance Generation*. Artificial Intelligence in Education. Open Learning Environments: New Computational Technologies to Support Learning, Exploration and Collaboration|Artificial Intelligence in Education. Open Learning Environments: New Computational Technologies to Support Learning, Exploration and Collaboration: 85-92|xv+804.

Jinsuk, K. and Myoung Ho, K. (2004) *An Evaluation of Passage-Based Text Categorization*. Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies 23(1): 47-65.

Jones, S. and Paynter, G. W. (2002) *Automatic Extraction of Document Keyphrases for Use in Digital Libraries: Evaluation and Applications*. Journal of the American Society for Information Science and Technology 53(8): 653-677.

Jouis, C. (1995) *Seek, Un Logiciel D'acquisition Des Connaissances Utilisant Un Savoir Linguistique Sans Employer De Connaissances Sur Le Monde Externe*. Actes de JAVA 95. Grenoble.

Jouis, C., Biskri, I., Desclès, J.-P., Le Priol, F., Meunier, J. M., Mustafa, W. and Nault, G. (1997) *Vers L'intégration D'une Approche Sémantique Linguistique Et D'une Approche Numérique Pour Un Outil D'aide À La Construction De Bases Terminologiques*. Actes de la première Journée Scientifique et Technique (JST) du réseau Francophone de l'ingénierie de langue de l'AUPELF-UREF, pp 427-432. Avignon.

Jun-Tae, K. and Moldovan, D. I. (1995) *Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction*. IEEE Transactions on Knowledge and Data Engineering|IEEE Transactions on Knowledge and Data Engineering 7(5): 10.1109/69.469825.

Justeson, J. J. and Katz, M. (1995) *Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text* Language Engineering Vol.1(1), pp.9-27.

Kashyap, V. (1999) *Design and Creation of Ontologies for Environmental Information Retrieval*. Twelfth Workshop on Knowledge Acquisition, Modelling and Management Voyager Inn, Banff, Alberta, Canada.

Kaszkiel, M. and Zobel, J. (2001) *Effective Ranking with Arbitrary Passages*. Journal of the American Society for Information Science and Technology 52(4): 344-364.

KBSI (1994) *The Idef5 Ontology Description Capture Method Overview*. KBSI Report, Texas.

Khan, L. F. and Feng, L. (2002) *Ontology Construction for Information Selection*. 14th Ieee International Conference on Tools with Artificial Intelligence, Proceedings: 122-127.

Kheirbek, A. and Chiaramella, Y. (1995) *Integrating Hypermedia and Information Retrieval with Conceptual Graphs Formalism*. Hypertext - Information Retrieval - Multimedia. Proceedings HIM '95|Hypertext - Information Retrieval - Multimedia. Proceedings HIM '95: 47-60|337.

Kietz, J. U., Maedche, A. and Volz, R. (2000) *A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet*. In Aussenac-Gilles N, Biébow B, Szulman S (eds) EKAW'00 Workshop on Ontologies and Texts. Juan-Les-Pins, France. CEUR Workshop Proceedings 51:4.1-4.14. Amsterdam, The Netherlands.

Lafourcade, M. and Prince, V. (2001) *Synonymies Et Vecteurs Conceptuels*. TALN 2001, Tours, pp. 233-242.

Landauer, T. K. and Dumais, S. T. (1997) *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of*

Knowledge. Psychological Review 104(2): 211-240.

Landauer, T. K., Foltz, P. W. and Laham, D. (1998) *An Introduction to Latent Semantic Analysis*. Discourse Processes 25(2-3): 259-284.

Lebart, L. and Salem, A. (1988) *Analyse Statistique Des Données Textuelles* Paris, Dunod.

Lebart, L. and Salem, A. (1994) *Statistique Textuelle* Paris, Dunod.

Lefebvre, B., Gauthier, G., Tadie, S., Duc, T. H. and Achaba, H. (2005) *Competence Ontology for Domain Knowledge Dissemination and Retrieval*. Applied Artificial Intelligence 19(9-10): 845-859.

Lefebvre, B., Tadié, S., Cherkaoui, O., Gauthier, G., Gerbé, O. and Meunier, J.-G. (2003) *Le Projet Gdst* In Proceeedings of Colloque Médiation et Ingénierie des Connaissances, Marseille, France.

Lelu, A., Halleb, M. and Delprat, B. (1998) *Recherche D'information Et Cartographie Dans Des Corpus Textuels À Partir Des Fréquences De N-Grams*. Proceedings of JADT-98, Nice, France.

Liebowitz, J. (1999) *Knowledge Management Handbook*. Raton, Fla, CRC Press.

Lonsdale, D., Ding, Y., Embley, D. W. and Melby, A. (2002) *Peppering Knowledge Sources with Salt; Boosting Conceptual Content for Ontology Generation*. Proceedings of the AAAI Workshop on Semantic Web Meets Language Resources, Edmonton, Alberta, Canada.

Luger, G. F. (2002) *Artificial Intelligence, Structures and Strategies for Complex Problem Solving*. Fourth Edition, Addison-Wesley, Fourth edition published.

Luke, S., Spector, L., Rager, D. and Hendler, J. (1997) *Ontology-Based Web Agents*. Proceedings of the First International Conference on Autonomous Agents|Proceedings of the First International Conference on Autonomous Agents: 59-66|xvi+549.

Maedche, A. and Staab, S. (2000) *Discovering Conceptual Relations from Text*. ECAI 2000. 14th European Conference on Artificial Intelligence. including Prestigious Applications of Intelligent Systems (PAIS-2000). Proceedings (Frontiers in Artificial Intelligence and Applications Vol.54)|ECAI 2000. 14th European Conference on Artificial Intelligence. including Prestigious Applications of Intelligent Systems (PAIS-2000). Proceedings (Frontiers in Artificial Intelligence and Applications Vol.54): 321-5|xvi+778.

Maedche, A. and Staab, S. (2001) *Ontology Learning for the Semantic Web*. Ieee Intelligent Systems & Their Applications 16(2): 72-79.

Maedche, A. and Staab, S. (2002) *Measuring Similarity between Ontologies*. Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web. 13th International Conference, EKAW 2002. Proceedings (Lecture Notes in Artificial Intelligence Vol.2473): 251-63|xi+402.

Maedche, A. and Volz, R. (2001) *The Text-to-onto Ontology Extraction and Maintenance Environment*. In Proceedings of the ICDM Workshop on integrating data mining and knowledge management, San Jose, California, USA.

Mahesh, K. (1996) *Ontology Development for Machine Translation: Ideology and Methodology*. Technical Report MCCS 96-292, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

Manning, C. D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. MIT Press.

Memmi, D. (2000) *Le Modèle Vectoriel Pour Le Traitement De Documents*. Cahiers Leibniz n°2000-14.

Memmi, D., Meunier, J. G. and Gabi, K. (1998) *Dynamical Knowledge Extraction from Texts by Art Networks*. Proceedings of Neurap. Marseille. pp. 205-210.

Meunier, J. G. (1992) *Le Problème De La Catégorisation Dans La Représentation Des Connaissances*. INTELLICA, 1-2 pp 353.

Meunier, J. G. (1995) *La Lecture Et L'analyse De Texte Assistée Par Ordinateur : La Chaîne D'analyse*. Cahiers de recherche du Laboratoire d'ANalyse Cognitive de l'Information. Vol. 6.

Meunier, J. G. (1996) *La Théorie Cognitive: Son Impact Sur Le Traitement De L'information Textuelle*. In V. Rialle et Fisette, D (Ed.), Penser l'Esprit, Des sciences de la cognition à une philosophie cognitive. (pp. 289-305). Grenoble: Presses de L'Université de Grenoble.

Meunier, J. G. (2002) *La Représentation Et Les Sciences Cognitives*. RSSI, 2002 et cahiers du LANCI 2001.

Meunier, J. G., Biskri, I., Nault, G. and Nyongwa, M. (1997) *Aladin Et Le Traitement Connexionniste De L'analyse Terminologique*. Actes de RIAO-97, Montréal, Canada, pp. 661-664.

Meunier, J. G., Forest, D. and Biskri, I. (2005) *Classification and Categorization in Computer Assisted Reading and Analysis of Texts*. In Lefebvre, C. et Cohen, H. (dir. publ.). 2005. Handbook of categorization in cognitive science. New York: Elsevier, pp. 955-978.

Mikheev, A. and Finch, S. (1997) *Workbench for Finding Structure in Texts*. Proceedings of

ANLP-97 (Washington D.C.). ACL March 1997. pp 8.

Mili, H., Ah-Ki, E., Godin, R. and McHeick, H. (1997) *Another Nail to the Coffin of Faceted Controlled-Vocabulary Component Classification and Retrieval.* Software Engineering Notes|Software Engineering Notes 22(3): 89-98.

Miller, G. A. (1990) *Wordnet: An on-Line Lexical Database.* International Journal of Lexicography, 3(4), 235312. OSGOOD, C.E., SUCI, G.J., AND TANNENBAUM, P.H. 1957. The Measurement of Meaning. University of Illinois Press, Chicago.

Miller, G. A. (1995) *Wordnet - a Lexical Database for English.* Communications of the Acm 38(11): 39-41.

Missikoff, M., Navigli, R. and Velardi, P. (2002) *Integrated Approach to Web Ontology Learning and Engineering.* Computer 35(11): 60-+.

Missikoff, M., Navigli, R. and Velardi, P. (2002) *The Usable Ontology: An Environment for Building and Assessing a Domain Ontology.* The Semantic Web - ISWC 2002. First International Web Conference. Proceedings (Lecture Notes in Computer Science Vol.2342): 39-53|xvi+476.

Mitra, P., Wiederhold, G. and Jannink, J. (1999) *Semi-Automatic Integration of Knowledge Sources.* Proceedings of the Second International Conference on Information Fusion. FUSION '99|Proceedings of the Second International Conference on Information Fusion. FUSION '99: 572-80 vol.1|2 vol.xxvi+1296.

Modica, G., Gal, A. and Jamil, H. M. (2001) *The Use of Machine-Generated Ontologies in Dynamic Information Seeking.* Cooperative Information Systems. 9th International Conference CoopIS 2001. Proceedings (Lecture Notes in Computer Science Vol.2172): 433-47|xi+450.

Moffat, A., Sacks-Davis, R., Wilkinson, R. and Zobel, J. (1994) *Retrieval of Partial Documents.* Second Text REtrieval Conference (TREC-2) (NIST-SP 500-215)|Second Text REtrieval Conference (TREC-2) (NIST-SP 500-215): 181-90|viii+486.

Moldovan, D. I. and Girju, R. C. (2001) *An Interactive Tool for the Rapid Development of Knowledge Bases.* International Journal on Artificial Intelligence Tools (Architectures, Languages, Algorithms) 10(1-2): 65-86.

Morin, E. (1998) *Prométhée Un Outil D'aide a L'acquisition De Relations Sémantiques Entre Termes.* 5ème Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN'98), pp. 172-181, Paris, France.

Morin, E. (1999) *Automatic Acquisition of Semantic Relations between Terms from Technical Corpora.* TKE'99. Terminology and Knowledge Engineering. Proceedings Fifth International Congress on Terminology and Knowledge Engineering|TKE'99.

Terminology and Knowledge Engineering. Proceedings Fifth International Congress on Terminology and Knowledge Engineering: 268-78|x+832.

Murray, T. (1998) *Special Purpose Ontologies and the Representation of Pedagogical Knowledge*. International Conference on the Learning Sciences, 1996 Proceedings of ICLS 96|International Conference on the Learning Sciences, 1996 Proceedings of ICLS 96: 235-42|ix+597.

Nault, G. (2001) *Approche Cognitive De L'hypertextualisation Semi-Automatique. Effets Sur La Conception D'un Système D'assistance Interactive Fondé Sur Un Optimiseur Émergentiste*. Thèse de doctorat. Université du Québec à Montréal.

Newell, A. and Simon, H. A. (1976) *Computer Science as Empirical Enquiry: Symbols and Search*. ACM 19 (1976), pp. 113-126.

Nkambou, R. (1996) *Modélisation Des Connaissances De La Matière Dans Un Système Tutoriel Intelligent : Modèles, Outils Et Applications*. Thèse de doctorat, Université de Montréal, Montréal, Canada.

Nobécourt, J. (2000) *A Method to Build Formal Ontologies from Text*. In: EKAW-2000 Workshop on ontologies and text, Juan-Les-Pins, France.

Noy, N. F., Fergerson, R. W. and Musen, M. A. (2000) *The Knowledge Model of Protege-2000: Combining Interoperability and Flexibility*. Knowledge Engineering and Knowledge Management Methods, Models, and Tools. 12th International Conference, EKAW 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1937)|Knowledge Engineering and Knowledge Management Methods, Models, and Tools. 12th International Conference, EKAW 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1937): 17-32|xiii+456.

Noy, N. F. and McGuinness, D. L. (2001) *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March.

Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W. and Musen, M. A. (2001) *Creating Semantic Web Contents with Protege-2000*. Ieee Intelligent Systems & Their Applications 16(2): 60-71.

Ogden, C. K. and Richards, I. A. (1969) *The Meaning of Meaning*. London 1923, 10th edition 1969.

Orliac, B. (2004) *Automatisation Du Repérage Et De L'encodage Des Collocations En Langue De Spécialité*. Thèse de doctorat présentée à l'Université de Montréal.

Paquette, G. (2002) *Une Taxonomie Intégrée Des Habiletés. Dans : Modélisation Des Connaissances Et Des Compétences*. Un langage graphique pour concevoir et

apprendre. Presses de l'Université du Québec.

Patwardhan, S. and Pedersen, T. (2006) *Using Wordnet-Based Context Vectors to Estimate the Semantic Relatedness of Concepts*. In the Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together. Trento, Italy.

Pavel, T. G. (1975) *Possible Worlds in Literary Semantics*. Journal of Aesthetics and Art Criticism 34(2): 165-176.

Pereira, F. and Schabes, Y. (1992) *Inside-Outside Reestimation from Partially Bracketed Corpora*. 30th Annual Meeting of the Association for Computational Linguistics: 128-135.

Pereira, F. C. (1998) *Modeling Divergent Production: A Multi Domain Approach*. European Conference of Artificial Intelligence, ECAI'98, Brighton, UK.

Pierce, C. S. (1978) *Écrits Sur Le Signe*. Paris, éd. Seuil.

Rastier, F. (1990) *La Triade Sémiotique, Le Trivium Et La Sémantique Linguistique*. Nouveaux Actes sémiotiques (9), pp. 59-68. PULIM, Université de Limoges.

Rastier, F. (1991) *Linguistic Analysis of 'Expert' Texts*. Genie Logiciel & Systemes Experts|Genie Logiciel & Systemes Experts(23): 16-23.

Rastier, F. (1991) *Sémantique Et Recherché Cognitive*. PUF, Paris.

Rastier, F. (1994) *Tropes and Linguistic Semantics*. Langue Francaise(101): 80-101.

Razmerita, L., Angehrn, A. and Maedche, A. (2003) *Ontology-Based User Modeling for Knowledge Management Systems*. User Modeling 2003, Proceedings 2702: 213-217.

Remaki, L. and Meunier, J. G. (2000) *Un Modèle Hmm Pour La Détection Des Mots Composés Dans Un Corpus Textuel*. Actes de JADT-2000, Lausanne, Suisse.

Rigau, G. (1998) *Automatic Acquisition of Lexical Knowledge from Mrds.* . Ph.D. Thesis, Comptur Systems and linguistics Department. Polytechnic University of Catalunya.

Rogers, J., Roberts, A., Solomon, D., van der Haring, E., Wroe, C., Zanstra, P. and Rector, A. (2001) *Galen Ten Years On: Tasks and Supporting Tools*. Medinfo 2001: Proceedings of the 10th World Congress on Medical Informatics, Pts 1 and 2 84: 256-260.

Rosario, B. (2000) *Latent Semantic Indexing: An Overview*. INFOSYS 240.

Roux, C., Proux, D., Rechermann, F. and Julliard, L. (2000) *An Ontology Enrichment Method for a Pragmatic Information Extraction System Gathering Data on Genetic*

Interactions. Position paper in Proceedings of the ECAI2000 Workshop on Ontology Learning (OL2000), Berlin, Germany. August.

Rubin, D. L., Hewett, M., Oliver, D. E., Klein, T. E. and Altman, R. B. (2002) *Automatic Data Acquisition into Ontologies from Pharmacogenetics Relational Data Sources Using Declarative Object Definitions and Xml.* In: *Proceedings of the Pacific Symposium on Biology*, Lihue, HI, (Eds. R.B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale and T.E. Klein).

Sabah, G. (1998) *L'ia Et Le Langage : Représentation Des Connaissances.* . Editions Hermès.

Salton, G. (1968) *Automatic Information Organisation and Retrieval.* McGraw-Hill, New York.

Salton, G. (1989) *Automatic Text Processing.* Addison-Wesley.

Salton, G. and McGill, M. J. (1983) *Introduction to Modern Information Retrieval.* Introduction to modern information retrieval: xv+448.

Saussure, F. D. (1916) *Cours De Linguistique Générale.* Translated by W. Baskin as Course in General Linguistics, Philosophical Library, New York, 1959.

Schultz, C. K. (1969) *H.P. Luhn: Pioneer of Information Science. Selected Works.* London : Macmillan.

Schwab, D., Lafourcade, M., V. et Prince, V. (2002) *Amélioration De La Représentation Sémantique Lexicale Par Les Vecteurs Conceptuels : Le Rôle De L'antonymie.* JATD 2002, vol. 2, pp. 701-712.

Sebastiani, F. (2002) *Machine Learning in Automated Text Categorization.* Acm Computing Surveys 34(1): 1-47.

Shimizu, H., Seta, K., Hayashi, S., Motomatsu, M., Ikeda, M. and Mizoguchi, R. (1999) *A Basic Consideration on Design Patterns for Ontology Building.* Proc. of the 58th National Conference of Information Processing Soc. Of Japan, 3U-9.

Smadja, F. (1993) *Retrieving Collocations from Text : Xtract.* In Computational Linguistics, n° 19(1), pp 143-178.

Smolensky, P. (1988) *On the Proper Treatment of Connectionism.* Behavioral and Brain Sciences 11(1): 1-23.

Srivastava, S., Ladadrid, G. D. and Elvadapu, C. S. (2002) *Document Ontology: A Statistical Approach.* SSGRR'2002, L'Aquila, Italy.

Staab, S., Studer, R., Schnurr, H. P. and Sure, Y. (2001) *Knowledge Processes and*

Ontologies. Ieee Intelligent Systems & Their Applications 16(1): 26-34.

Steve, G. and Gangemi, A. (1996) *Onions Methodology and the Ontological Commitment of Medical Ontology On8.5*. In Proceedings of the 10 th Knowledge Acquisition Workshop - KAW'96, Banff, Canada, November 9-14.

Stojanovic, N., Stojanovic, L. and Volz, R. (2002) *A Reverse Engineering Approach for Migrating Data-Intensive Web Sites to the Semantic Web*. Intelligent Information Processing. IFIP 17th World Computer Congress - TC12 Stream on Intelligent Information Processing: 141-54|xii+316.

Stumme, G., Studer, R. and Sure, Y. (2000) *Towards an Order-Theoretical Foundation for Maintaining and Merging Ontologies*. Verbundtagung Wirtschaftsinformatik 2000 (Meeting on Business Informatics)|Verbundtagung Wirtschaftsinformatik 2000 (Meeting on Business Informatics): S136-49|225.

Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R. and Wenke, D. (2002) *Ontoedit: Collaborative Ontology Engineering for the Semantic Web*. In Proceedings of the International Semantic Web Conference (ISWC 2002), Sardinia, Italia, June 9-12.

Swartout, B., Ramesh, P., Knight, K. and Russ, T. (1997) *Toward Distributed Use of Large-Scale Ontologies*. In Symposium on Ontological Engineering of AAAI, Stanford, California, March.

Tallis, M. and Gil, Y. (1999) *Designing Scripts to Guide Users in Modifying Knowledge-Based Systems*. Proceedings Sixteenth National Conference on Artificial Intelligence (AAI-99). Eleventh Innovative Applications of Artificial Intelligence Conference (IAAI-99)|Proceedings Sixteenth National Conference on Artificial Intelligence (AAI-99). Eleventh Innovative Applications of Artificial Intelligence Conference (IAAI-99): 242-9|xxvi+998.

Thompson, C. A. and Mooney, R. J. (1997) *Semantic Lexicon Acquisition for Learning Parsers* Technical Note. January 1997.

Toussaint, Y., Royaute, J., Muller, C. and Polanco, X. (1997) *Analyse Linguistique Et Infométrique Pour L'acquisition Et La Structuration Des Connaissances*. Actes des deuxièmes rencontres Terminologie et Intelligence Artificielle (TIA'97), pp 27-46. Toulouse.

Uschold, M. (1996) B*uilding Ontologies: Towards an Unified Methodology*. In Proceedings of the 16th conference of the British Computer Society Specialist Group on Expert Systems.

Uschold, M. and King, M. (1995) *Towards a Methodology for Building Ontologies*. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at the International Joint Conference on Artificial Intelligence (IJCAI'1995).