Ce rapport est une partie du rapport d'un stage effectué au sein du laboratoire Dialogue et Intermédiations Intelligentes de la Direction des Interactions Humaines DIH/D2I au centre de Recherche et Développement de France Télécom (FTR&D) à Lannion du 15 avril au 13 septembre 2002.

Introduction

Le terme « ontologie » vient du domaine de la philosophie, où il signifie « explication systématique de l'existence » (de ontos « l'être, ce qui est », et de logos « discours »). Dans cette perspective, une ontologie est indépendante du langage utilisé pour sa description [TAM01]. De nombreuses acceptions du mot « ontologie » existent (vocabulaire technique, référentiel métier, terminologie/thesaurus, système de classes d'une représentation par objet, base de connaissances terminologique, …) [CORBY02].
Bien que la communauté d'I.A. semble s'accorder sur l'utilisation et le sens du mot « ontologie », il n'existe pas de définition formelle unanimement acceptée.

Une ontologie est vue comme un moyen de « décrire de façon explicite la conceptualisation des connaissances représentées dans une base de connaissances » [BENJAM99]. Cette définition est une extension de celle de Gruber (1993), décrivant une ontologie comme « une spécification explicite d'une conceptualisation ».

Une ontologie est une modélisation des connaissances du monde, d'informations extra-linguistiques. Organisée en un réseau de concepts, elle consiste en un ensemble de définitions de catégories de base (objet, relations, propriétés) qui permettent de décrire les objets du domaine que l'on traite, leurs propriétés et les relations qu'ils entretiennent les uns avec les autres [BOUIL98]. On les appelle aussi bases de connaissances (knowledge base). Ce modèle du monde est idéalement indépendant des langues traitées par un système ; les mots décrits dans les lexiques de diverses langues pointeront vers les concepts de l'ontologie qu'ils expriment.

Les ontologies ont pour but de saisir la connaissance dans un domaine, d'une façon générale et de fournir une représentation communément acceptée qui pourra être ré-utilisée et partagée par divers applications et groupes. Les ontologies fournissent un vocabulaire commun dans un domaine et définissent - à différents niveaux de formalisation - la signification de termes et leurs relations [BENJAM99].
« Une ontologie peut prendre une grande variété de formes, mais elle inclura nécessairement une liste de termes et des spécifications sur leurs significations. Cela comprend des définitions et des indications sur la façon dont les concepts sont reliés entre eux, imposant une structure au domaine et contraignant les interprétations possibles des termes. » [JASP99].
Une ontologie est donc un « modèle conceptuel spécifique élaboré dans le domaine de la gestion du savoir. Une ontologie peut représenter des relations complexes entre des objets et inclure les règles et axiomes manquants dans un réseau sémantique. Une ontologie qui décrit le savoir dans un domaine précis est souvent reliée à des systèmes de prospection de données et de gestion des connaissances. » selon [NKOS].

Le développement d'une ontologie dans un système de T.A.L. a pour but d'améliorer la qualité et la généralité d'un système. Elle permet d'obtenir une représentation du texte plus profonde, plus abstraite, voire indépendante de toute langue. D'une part l'ontologie fournit aux différents modules (lexical, syntaxique, sémantique et pragmatique) des connaissances encyclopédiques indispensables pour lever certaines ambiguïtés résiduelles. D'autre part, comme certain système de représentation des connaissances comprennent un mécanisme d'inférence, l'ontologie peut également fournir des informations implicites dans le texte (ce qui permet à un système d'interrogation de donner des réponses « intelligentes » qui tiennent compte des évidences pour un locuteur).
Dans un système de T.A.L., une ontologie est indispensable si l'on veut comprendre le contenu d'un texte et le représenter dans un langage formel. Cette représentation formelle du

sens d'un texte peut être utilisée dans diverses applications pour générer un nouveau texte en langue naturelle.
La conception d'une ontologie pour le T.A.L. exige que soit dépassée la spécificité linguistique de la sémantique lexicale (qui n'est pas linguistiquement neutre) et réduite la généralité des bases de connaissances de l'I.A.. La taille et la complexité d'une ontologie dépendront de l'architecture du système en question et de l'existence d'autres sources de connaissances : syntaxiques, sémantiques, rhétoriques (cohérence textuelle) et pragmatiques (situation de communication).

Les principaux buts aujourd'hui, dans le domaine de conception des ontologies, sont de faire des ontologies partageables en développant des formalismes et des outils communs, de développer le contenu des ontologies (conception d'ontologie), et de comparer, de rassembler, de traduire et de constituer diverses ontologies. Le travail récent en conception d'ontologie a produit une gamme de projets divers, des ontologies représentant la connaissance du monde générale, aux ontologies de domaines spécifiques en passant par les ontologies de système de représentation de connaissances qui englobent les «cadres ontologiques». La communauté de l'« ingénierie ontologique » s'accorde à dire qu'il serait bénéfique de pouvoir intégrer des ontologies de façon à partager et réutiliser les connaissances de chacun [NOY97].

(…)

Ainsi, après un état de L'Art des ontologies linguistiques, quelques pistes sur les théories d'intégration seront données.

A- Ontologies:

Il s'agit de présenter ici les différentes formalisations utilisées pour la modélisation de la langue afin d'approcher une description fine de la structure d'une ontologie, d'un point de vue théorique et applicatif. Une présentation des ontologies existantes focalisée plus particulièrement sur les ontologies de haut-niveau permettra une connaissance de base des concepts nécessaire aux techniques d'intégration.

1 Les représentations du langage

[HEIJ97] met en place une classification d'ontologie sur la base de la quantité et le type de structure de conceptualisation, de laquelle il définit des ontologies d'information telles que les schémas de bases de données, des ontologies de modélisation de connaissances et des ontologies terminologiques telles que les lexiques. C'est ce dernier type que l'on peut considérer comme la base des représentations du langage. Une ontologie est semblable à un dictionnaire ou à un glossaire, mais avec plus de détails et une structure qui permet à des ordinateurs de traiter son contenu.

La langue peut être représentée et modélisée de différentes façons. Nous verrons tout d'abord quels sont les grands types de représentation du langage, puis les formes de représentation les plus courantes, et enfin les types de langages utilisés.

1.1. Typologie selon les types de connaissances modélisées
[MCGUI01] [TAM01]

Les ontologies sont classifiées ici sur la base de leur force d'expression, c'est à dire sur la base de l'information que l'ontologie doit exprimer.

1.1.1 Lexiques contrôlés

Le lexique est la notion la plus simple possible d'ontologie, qui est une liste finie de termes. Le lexique est un ensemble de sens lexicaux associés à des traits syntaxiques, morphologiques et sémantiques.

La théorie principale ayant trait au lexique est celle de Pustejovski. La théorie du lexique génératif en 1998 [GARD01] naît à la suite du problème engendré par le principe de lexique énumératif, où des sens distincts correspondent à des unités lexicales (Uls) distinctes, et où l'on se contente d'énumérer les Uls. En effet, sous cette vision, on ne rend pas compte de la différence entre ambiguïté (ambiguïté contrastive) et généralité (ambiguïté complémentaire).

Le principe du lexique génératif est que lors de la construction du sens de la phrase, un ensemble de sens lexicaux noyaux est utilisé pour générer un ensemble plus large de sens lexicaux.

Le but de Pustejovky est d'expliquer l'interprétation des mots en contexte, de dériver d'un ensemble fini de ressources un ensemble infini de sens lexicaux, et enfin d'expliquer la polymorphie (un mot, plusieurs sens) par exemple, de prédire les relations systématiques entre sens lexicaux.

Pour cela il faut, grouper les mots en classes sémantiques (puisque la catégorie sémantique d'un mot détermine son comportement syntaxique et la dénotation de ses éléments) et définir les relations lexicales entre mots.

Levin en 1993 propose d'utiliser les ressemblances et différences syntaxiques pour déterminer les classes sémantiques (exemple : transitif causal/intransitif ; forme transitive et intransitive sont reliées par le trait sémantique de « cause »). Alors que les mécanismes génératifs utilisés par Pustejovski sont, le liage sélectif, la coercion de type et la co-composition par exemple.

Le génératif de Pustejovski possède 4 niveaux de représentation :
1. La structure argumentale : Dans laquelle on peut distinguer 4 types d'arguments pour un élément lexical :
   - Argument vrai : un argument nécessairement réalisé au plan syntaxique.
   - Argument par défaut : un argument sémantique qui n'est pas nécessairement réalisé syntaxiquement
   - Argument ombre : un élément sémantique incorporé dans le sens de l'élément lexical et qui ne peut être réalisé que par une sur-spécification sémantique
   - Modificateur : élément sémantique modifiant le sens mais qui n'est lié à la représentation sémantique d'aucun élément lexical. Les modificateurs sont associés à des classes sémantiques et non à des mots.
2. La structure événementielle : dans laquelle un événement est traité comme étant composé de plusieurs sous-événements. Une structure événementielle comprend trois éventualités dont deux (les sous-éventualités) sont incluses dans la troisième :
   - Une relation temporelle (précède, coïncide, coïncide partiellement) entre les deux sous-éventualités
   - Une relation de prominence indiquant la tête de l'évènement (*e1* est la tête de l'évènement *e2*)
3. La structure de qualia : Cette structure spécifie 4 aspects du sens d'un mot (ou qualia) :
   - Constitutif : la relation entre l'objet et ses composantes (matière, poids, parties)
   - Formelle : le relation qui distingue l'objet d'un domaine plus large (orientation, taille, forme, dimension, couleur, position)
   - Télic : la fonction de l'objet
   - Agentif : les facteurs impliqués dans la création de l'objet (créateurs, chaîne causale)

Il est à noter que chaque catégorie syntaxique peut être associée à une structure de qualia ; toutes les Uls n'ont pas nécessairement une valeur pour chacun des attributs dans le qualia. De plus, le qualia contient une liste de propriétés qui sont connues de l'objet décrit et il est aussi la base sémantique qui permet de rendre compte de la polysémie.
4. La structure d'héritage lexical.

Les mécanismes génératifs reliant ces niveaux et permettant une interprétation compositionelle sont :
La coercion de types : un foncteur force un changement de type sémantique pour un argument.
Le liage sélectif : un foncteur opère sur une partie d'un constituant.
La co-composition : plusieurs éléments fonctionnent comme foncteur sémantique.
Une même entrée lexicale regroupe plusieurs sens – ces "méta-entrées" sont appelées des paradigmes conceptuels lexicaux (Lexical Conceptual Paradigm, LCP).

1.1.2 Glossaires

Ce sont des listes de termes avec leurs significations. Les significations sont le plus souvent exprimées par des énoncés en langue naturelle qui sont principalement destinés à des agents humains.

1.1.3 Thesauri

Ils ajoutent aux glossaires la sémantique ressortant des définitions des relations entre les termes (comme la relation de synonymie). Généralement, ils ne fournissent pas la structure hiérarchique explicite, bien que celle-ci puisse être déduite des spécifications de termes plus larges ou plus proches.

1.1.4 Hiérarchies Is-a informelles

Cette catégorie inclut la plupart des ontologies du web. Ce sont des ontologies où une notion vague de généralisation et de spécialisation est fournie bien que ce ne soit pas une hiérarchie stricte de sous-classe (ex : Yahoo !)

1.1.5 Hiérarchies Is-a formelles

Ce sont des ontologies où les concepts sont organisés selon une hiérarchie de sous-classe stricte. Le concept d'héritage est toujours applicable dans ce type d'ontologie. Cette ontologie peut inclure uniquement des noms de classe.

1.1.6 Instances formelles

Les ontologies incluant des relations d'instances formelles sont une extension naturelle des ontologies appliquant une structure de hiérarchie stricte.

1.1.7 Cadres

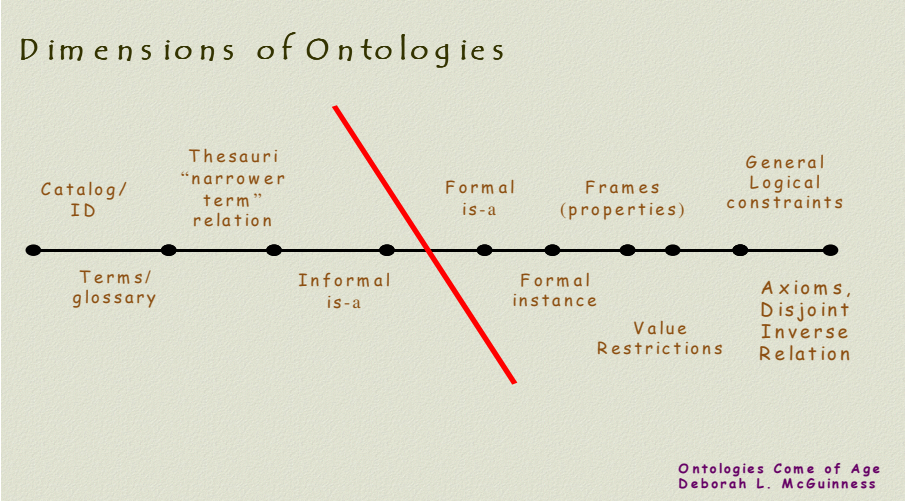
(Description de propriétés de concept) Ce sont des ontologies dont les concepts sont décrits en terme de propriétés caractéristiques. Le fait d'inclure des propriétés dans la description du concept devient intéressante dans la mesure où l'on peut appliquer le principe d'héritage sur ces propriétés.

1.1.8 Restriction de valeur

Ces ontologies permettent d'appliquer des restrictions aux valeurs associées aux propriétés (ex : nombre maximum de noms pour décrire le concept).

1.1.9 Contraintes de logique générale

Ces ontologies sont celles qui ont la plus grande force d'expression. Par exemple, ces ontologies peuvent être basées sur des équations mathématiques qui utilisent des valeurs d'autres propriétés ou les propriétés peuvent être exprimées comme des énoncés logiques. Ce type d'ontologie est en général écrit dans un langage d'ontologie très expressif, tel que Ontolingua par exemple.



[PATIL]

1.2 Les formalismes de représentation
[BOUIL98]

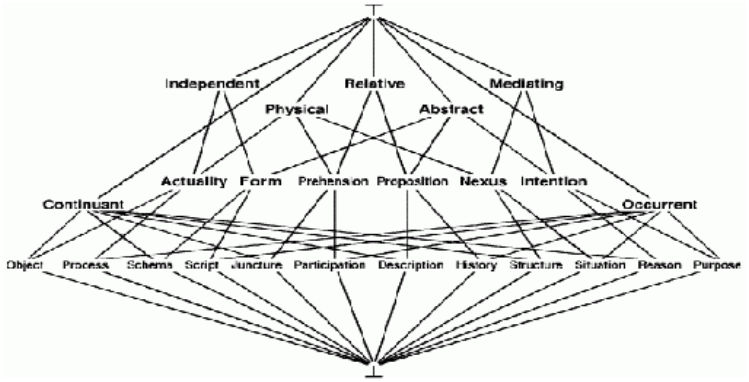
De nombreux formalismes ont été développés pour représenter les connaissances, de la logique des prédicats jusqu'aux langages sophistiqués basés sur des structures de données appelées schémas.

1.2.1 Les formalismes logiques :

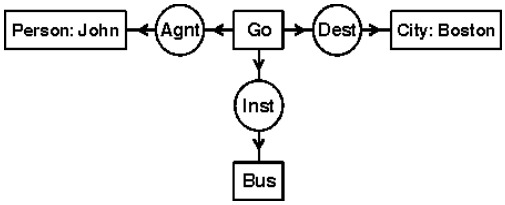
Dans une représentation logique, la base de connaissances consiste en un ensemble d'axiomes décrivant une situation, un état de choses, sur lesquels des règles d'inférence opèrent et fournissent de nouvelles formules que l'on peut considérer comme valides. Celles-ci constituent alors de nouveaux états de choses dans la base.
Le langage de programmation Prolog est fondé directement sur ces principes.

1.2.2 Les réseaux sémantiques :

Un réseau sémantique est un modèle de représentation du contenu sémantique des concepts sous forme de graphe. Un graphe est formé de nœuds, représentant les concepts, reliés par des arcs décrivant les relations entre eux.
En I.A., Quillian fut le premier à développer de tels réseaux en tant que modèles de la mémoire associative humaine. Actuellement, la théorie des graphes conceptuels de [SOWA84], représentant les relations sémantiques, constitue le formalisme le plus répandu pour conceptualiser les ontologies.
Les concepts y sont organisés en un treillis de types, reliés par une relation d'ordre correspondant au lien *sorte-de*.
Ontologie « Top-Level » de Sowa, treillis :

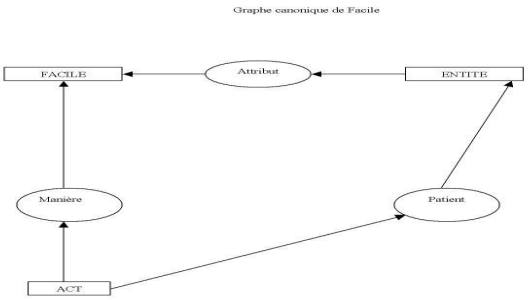


Graphe conceptuel à quatre concepts de la phrase: John is going to Boston by bus



La théorie repose sur la notion de graphe canonique permettant de définir de nouveaux concepts et de représenter des propositions.

1.2.3 Les graphes canoniques :



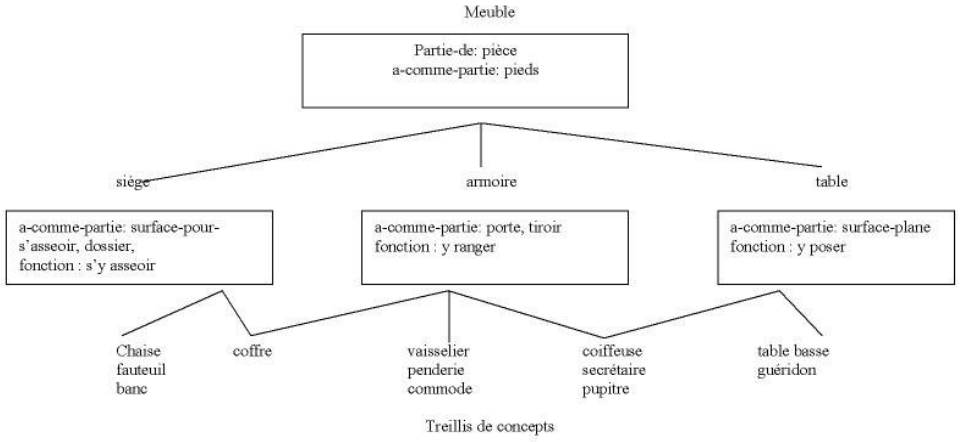
Ils représentent des rôles associés à chaque concept.
La théorie permet de définir de nouveaux types de concepts à l'aide des graphes canoniques. De même, on peut les combiner pour représenter des propositions.

1.2.4 Les primitives :

Les primitives ont été utilisées dans la théorie de la dépendance conceptuelle de Schank en 1972. Elle propose un modèle de représentation conceptuelle fondé sur l'organisation de la mémoire et insiste sur l'universalité des primitives qui reflètent la pensée plutôt que la langue. Étant universelles, les primitives sont définies d'une façon très générale et en petit nombre, ce qui pose problème pour la définition de domaines spécifiques.
Schank distingue six classes de concept, dont deux pour le contexte ; objets physiques, actions, attributs d'objets, attributs d'actions, lieu et temps. De même, il définit onze actions de base permettant de représenter toutes les actions possibles et six rôles conceptuels décrivant les relations entre entités et actions.

1.2.5 Les schémas :

Un schéma est une structure de données qui regroupe un ensemble d'informations concernant un concept particulier. Les concepts sont généralement organisés en treillis suivant la relation sorte-de. A chaque nœud est associé un schéma qui décrit les propriétés du concept dont héritent les concept descendants.



Treillis de concepts

La notion de schéma dont on trouve l'origine chez Minsky en 1974, a été reprise en I.A. pour développer des formalismes de représentation des connaissances, par exemple les langages KRL (1977) et KL-ONE (1985). Ces formalismes intègrent souvent des stratégies d'inférence spécifiques, comme le raisonnement par défaut et l'héritage automatique des propriétés en suivant les arcs d'une hiérarchie de concepts. (ex : ONTOS).

Définition d'une « frame » donnée par Minsky :
« Une frame est une structure de données représentant une situation stéréotypée, comme se trouver dans un certain type de salon ou se rendre à un goûter d'anniversaire d'un enfant. Divers types d'informations sont associés à chaque frame. Certaines d'entre elles concernent l'utilisation de cette frame. D'autres portent sur ce que l'on s'attend à ce qu'il arrive par la suite. D'autres encore portent sur ce qu'il faut faire si ces attentes ne sont pas confirmées. »

1.2.6 Les scripts :

La notion de scripts ou scénarios a été introduite par Schank et Abel en 1977, sur le modèle des frames pour le traitement du langage naturel. Les frames servent alors à représenter des séquences d'actions stéréotypées appelées scénarios/scripts.
Un scénario consiste en un ensemble d'actions élémentaires, ou de références à d'autres scénarios, ordonnées selon leur déroulement dans le temps.
Les scripts décrivent la chronologie et décompose les procédures. On y décrit les objets et les acteurs, les conditions initiales et les résultats, des scènes de la vie courante.
Le traitement d'un texte consiste à reconnaître un événement décrit en accédant au scénario adéquat et à l'interpréter, c'est à dire à en extraire des prédictions (ou résultats). Reste le problème des situations inattendues ainsi que des situations faisant intervenir plusieurs scénarios simultanément.

[HEIJ97] définit d'ailleurs sa typologie sur le sujet de la conceptualisation et sur le type de représentation utilisé dans la conception d'ontologie. Ses « ontologies de représentation » regroupent les primitives de représentation utilisées dans la formalisation des connaissances

dans les paradigmes de représentation de connaissances. (ex : Frame-ontolgy dans Ontolingua Server).

1.3 La formalisation et les langages

Il existe divers formalismes pour les ontologies. Certains sont définis en « extension » (par Gruber 93), d'autres en « intension » (Guarino 98).

Une ontologie formalisée en extension possède une définition formelle avec une sémantique déclarative (vocabulaire, thésaurus, etc.). Elle est constituée d'un ensemble de « lexons », expressions élémentaires construites d'un élément de contexte, de terme et de rôle.

Une ontologie formalisée en intension possède une définition à travers des « mondes possibles » (thésaurus de domaine vus comme une synthèse organisée des arrangements de termes possibles linguistiquement).

Définie d'une façon plus ou moins formelle, l'ontologie n'acceptera pas les mêmes langages quant à son implémentation.

L'ontologie doit être implémentée dans un langage. Quatre sortes d'ontologies sont distinguées en fonction du type de langage utilisé [Usch96].

1.3.1 Les ontologies hautement informelles

Exprimées en langage naturel (type Guarino).

1.3.2 Les ontologies semi-informelles

Exprimées dans un langage naturel structuré et limité, c'est à dire que des patrons ont été mis en œuvre.

1.3.3 Les ontologies semi-formelles

Exprimées dans un langage défini artificiellement et formellement.

1.3.4 Les ontologies rigoureusement formelles

Exprimées dans un langage contenant une sémantique formelle, des théorèmes et des preuves de propriétés telles que la robustesse et l'exhaustivité. La plupart des ontologies est aujourd'hui implémentée en langage formel (Ontolingua, CycL, Loom, Flogic).

1.3.5 Les langages UNL et OWL

UNL **Universal Networking Language**	**OWL** **Ontology Web Language**
Création en 1996	Création en 1997
	Dernières spécif. Mai 2002
Fondé sur XML (Extensible Markup Language)	Fondé sur RDF (Resource Description Framework) Et RDFs (RDF Schema) Fondé sur et donc lien possible avec DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer)
Utilisé pour traduction web.	Utilisé pour Web sémantique, raisonnement sur ontologie
Remarque: Interlingua anglo-sémantique	Remarque: RDF pas assez expressif pour W3c, et trop expressif pour Informaticiens.
Composé de: Lexique= UW Syntaxe= attribut, relation Sémantique= KB C.A.D.: - UW= concepts - Attributs booléens= nombre, modalité,… - Relations sémantiques= agent, bénéficiare, -moyens,… - UNL KB= ensemble d'entrées de KB, définition de relations binaires possibles - language server= enconversion/déconversion - Dico= correspondance () mots et UW	Composé de: Docs Web, référencés par URI Composants non- logiques Classes Propriétés "individuals" axiomes= liens () classes et propriétés
Document UNL= liste de relations entre concepts	Ontology OWL= séquence d'axiomes et de faits + références à d'autres ontologies
Construction d'ontologies sans raisonnement dessus	
Voir: http://www.unl.ias.unu.edu/publications/UNL-beyond%20MT.html	Voir: http://www.w3.org/2001/sw/WebOnt/ http://lists.w3.org/Archives/Public/www-archive/2002May/att-0021/01-_owl.html

2 Description d'une ontologie

Les ontologies fournissent le vocabulaire commun d'un domaine et définissent, de façon plus ou moins formelle, le sens des termes et les relations entre ces derniers.
Le propos général d'une ontologie est de catégoriser un même monde. Pourtant les ontologies peuvent être très différentes aussi bien au niveau de leur « top-level » qu'au niveau du traitement de leurs composants de base telles que les choses, les procès, les relations…

2.1 Ses composants

Une ontologie débute par une taxinomie, un arrangement structuré d'informations en classes qui catégorisent un sujet et ses dépendants.
La taxinomie est la partie centrale de la plupart des ontologies. Cependant, l'organisation taxinomique peut varier d'une façon très importante elle aussi.
Selon (Gruber 1993), la formalisation d'une ontologie se met en place grâce à 5 types de composants (« Modelling primitives »). Nous en présenterons un en plus ci-dessous, de Sowa, le rôle.

2.1.1 Les classes/ les concepts

Une classe ou un concept, représente un type d'objet dans l'univers.
Les classes sont habituellement organisées en taxinomies auxquelles on applique des mécanismes d'héritage. Tous les concepts peuvent être organisés en une large taxinomie. Il peut aussi y avoir un grand nombre de hiérarchisations plus petites, ou bien pas de taxinomie explicite du tout.
Les concepts sont utilisés dans leur sens large. Ils peuvent être abstraits ou concrets, élémentaires (électrons) ou composés (atome), réels ou fictifs.

Il arrive que les définitions des ontologies aient été diluées, en ce sens que les taxinomies sont considérées comme des ontologies complètes (Studer et al. 1998).
En résumé, un concept peut être tout ce qui peut être évoqué et, partant de là, peut consister en la description d'une tâche, d'une fonction, d'une action, d'une stratégie ou d'un processus de raisonnement, etc.
Selon [PINTO00] Il est souvent fait référence aux concepts en tant qu'union de classes et d'instances, alors que chacun des constituants de l'ontologie est considéré comme un fragment de connaissance (knowledge piece).

2.1.2 Les relations

Les relations représentent un type d'interaction entre les notions d'un domaine. Elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles. Des exemples de relations binaires sont sous-classe-de, connecté-à.
Certaines relations binaires entre des objets sont considérées comme des rôles comme le définit Borgida [BORG96] (extrait de [TAM01]).

2.1.3 Les rôles

Selon Sowa, « un rôle caractérise une entité par quelque rôle qu'elle joue dans sa relation à une autre entité. Le type « Humain », par exemple, est un type de phénomène qui dépend de la

forme interne de l'entité ; mais la même entité peut être caractérisée par des rôles du type, Mère, Employé ou Piéton. ».

2.1.4 Les fonctions

Les fonctions sont aussi des cas particuliers de relations dans lesquelles le $n^{ième}$ élément de la relation est défini à partir des n-premiers. Comme exemple de fonctions binaires il y a la fonction mère-de ou carré-de, comme fonction ternaire, le prix d'une voiture usagée sur lequel on peut se baser pour calculer le prix d'une voiture d'occasion en fonction de son modèle, de sa date de construction et de son kilométrage.

2.1.5 Les axiomes

Les axiomes sont utiles à la structuration de phrases qui sont toujours vraies. Ils permettent de contraindre les valeurs de classes ou d'instances.

2.1.6 Les instances

Les instances sont utilisées pour représenter des éléments dans un domaine.

2.2 Types d'ontologies

Les types d'ontologies [GOM99] mises au point sont très diverses. Cette section n'a pas l'ambition de fournir une typologie exhaustive des ontologies telle que celles de van Heijst (*et al.* 1997) et Mizoguchi (*et al.* 1995). Elle présente néanmoins les types d'ontologies les plus couramment utilisés. D'une façon générale, on identifie les catégories suivantes: les ontologies de haut-niveau et les ontologies spécialisées, constituées principalement des ontologies de domaine, d'application et de tâches.

Les ontologies présentes dans la littérature sont classifiées suivant différentes dimensions que nous avons tenté d'homogénéiser.

2.2.1 Les ontologies de haut-niveau

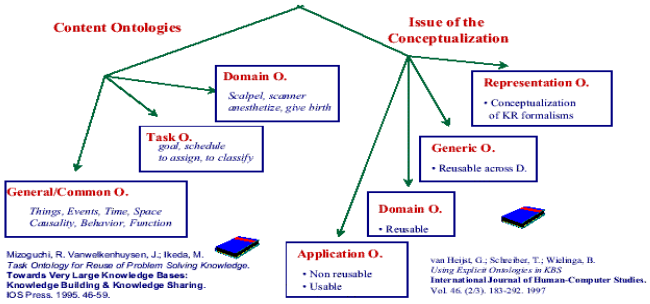


Figure 1: Types of ontologies according to Mizoguchi and colleagues [82] and van Heijst and colleagues citevanHeiist:97a

Ce type d'ontologie décrit des concepts très généraux ou des connaissances de sens commun telles que l'espace, le temps, l'événement, l'action…, qui sont indépendantes d'un problème ou d'un domaine particulier.

Qu'elles soient appelées « top level ontologies » [VANZYL99] [GUAR98], « ontologies de sens commun/ général » [MIZ95] ou encore « meta-ontologies » [MIZ95] ou « ontologies génériques ou noyaux d'ontologies » [HEIJ97], ces ontologies de haut-niveau (Top-Level

ontology, ou Upper Level ontology) fournissent des notions générales auxquelles tous les termes des ontologies existantes doivent être reliés. Elles sont réutilisables d'un domaine à l'autre (définies en relation de « partie-de » et ses propriétés)..
Une ontologie de haut niveau est généralement conçue afin de réduire les incohérences des termes définis plus bas dans la hiérarchie.
Elles incluent du vocabulaire en lien avec les choses, les événements, le temps, l'espace, la cause, le comportement, etc. (ex : Cyc)

Il n'existe pas pour le moment d'ontologies de haut-niveau unifiées. (ex. d'ontologie de haut niveau : Sowa's boolean lattice, PANGLOSS, Penman Upper Level, Cyc, Mikrokosmos, proposition de Guarino, Mereology - Borst 1997-).

2.2.2 Les ontologies spécialisées

Ce sont des ontologies qui « spécialisent » un sous-ensemble d'ontologies génériques en un domaine ou un sous-domaine. Elles peuvent être de domaine, d'application, techniques, ou partagées selon [VANZYL99], mais cela dépend des théories. Les trois principales sont :

Les ontologies de domaine :

Les ontologies de domaine sont spécialisées pour un certain type d'artefact.
Ce sont des ontologies réutilisables au sein d'un domaine donné, mais pas d'un domaine à un autre. Elles fournissent le vocabulaire des concepts d'un domaine.
Ce type d'ontologie décrit un vocabulaire en relation avec un domaine générique comme la médecine ou la physique.
Elles se retrouvent dans la typologie de [HEIJ97] grâce à une classification selon le sujet de conceptualisation.

Les ontologies d'application :

Les ontologies d'application sont généralement spécifiques à une application ; Elles contiennent suffisamment de connaissances pour structurer un domaine particulier.

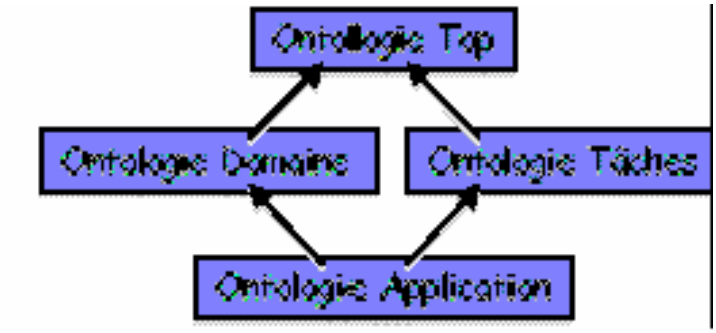
Selon [GUAR98], ce type d'ontologie décrit des concepts qui dépendent à la fois d'un domaine particulier et d'une tâche particulière. Elles seraient souvent des spécialisations à la fois des ontologies de domaine et des ontologies de tâches et correspondraient aux rôles joués par les entités de domaine lorsqu'elles effectuent certaines activités.

Les ontologies de tâche :

Selon [GUAR98], ce type d'ontologie décrit un vocabulaire en relation avec une tâche ou une activité générique comme le diagnostic ou la vente.

Les ontologies de tâche fournissent un lexique systématisé de termes utilisés pour résoudre les problèmes associés à des tâches particulières (dépendantes ou non du domaine). Ces ontologies fournissent un ensemble de termes au moyen desquels on peut décrire au niveau générique comment résoudre un type de problème. Elles incluent des noms génériques (par ex., plan, objectif, contrainte), des verbes génériques (par ex., assigner, classer, sélectionner), des adjectifs génériques (par ex., assigné) et d'autres mots qui relèvent de l'établissement d'échéances.

Les méta-ontologies, les ontologies de domaine et les ontologies d'application saisissent les connaissances statiques indépendamment de la façon dont on résout les problèmes alors que les ontologies de tâches sont axées sur les connaissances visant à résoudre des problèmes.
Tous ces types d'ontologie peuvent être combinés de façon à construire une nouvelle ontologie (C'est alors, une ontologie partagée qui peut être une ontologie ou une combinaison de quelques-unes d'entre-elles).
Si l'on applique le problème du compromis entre l'utilisabilité et la réutilisabilité (Klinker *et al.* 1991) au domaine de l'ontologie, on peut affirmer que plus une ontologie est réutilisable, moins elle est utilisable, et inversement [GOM99].



3 Les ontologies existantes

Description des différents projets

Le traitement du langage naturel est un challenge des plus étudiés par la technologie de programmation. De nombreuses équipes ont essayé de produire des systèmes de TAL capables de lire et de comprendre, tout d'abord des textes d'anglais ordinaire, mais aujourd'hui le multilinguisme touche de nombreux systèmes. Plusieurs projets de différentes époques et équipes sont présentés ci-dessous.
Seules les ontologies linguistiques les plus connues seront citées. Une présentation générale en sera faite, reprenant les fondements théoriques de l'ontologie et son architecture globale avec un détail sur son ontologie supérieure.

3.1 Cyc

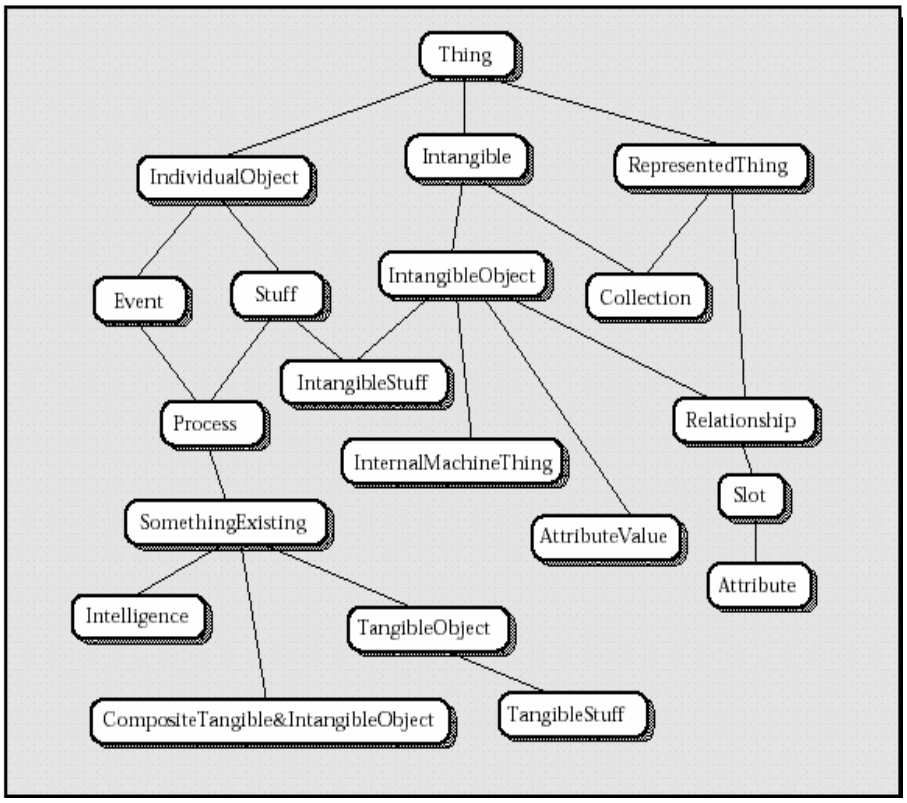
[CYC] *["Cyc" sounds like "psych"]*

Genèse du projet

Cyc® est un logiciel développé par Cycorp, Inc.(Austin, Texas), qui est le fournisseur principal de « sens formalisé ». Ce logiciel est en développement depuis 1984 au Microelectronics and Computer Technology Corporation (MCC) par Lenat et Guha. En 1995, l'entreprise devient autonome.
Tous les produits développés chez Cycorp, sont activés par une importante base de connaissance multi-contexte et un moteur d'inférence performant.
Ce composant permet à Cyc de comprendre et de raisonner sur plusieurs domaines d'application.
Cyc peut trouver des associations entre une question d'un utilisateur et une image dont la légende emploierait un vocabulaire voisin ; il peut combiner des informations de base de données multiples afin de deviner une nouvelle information.
L' « interprète sémantique » de Cyc inclut des principes de la sémantique de Montague.
Le support de la représentation est le langage formel CycL, décrit ci-dessous.

Structure

La base de connaissances est construite sur un noyau de plus d'un million d'assertions (ou règles) entrées manuellement et conçues pour rendre compte d'une partie importante de ce que l'on considère une connaissance consensuelle du monde. Par exemple, Cyc sait que les arbres sont généralement à l'extérieur, que quand quelqu'un est mort il ne peut généralement plus acheter quelque chose, ou qu'un verre rempli d'un liquide doit être transporté à l'endroit.



CYC: Top-Level Categories (adapted from Lenat and Guha [1990]). [NOY97]

En premier lieu est représenté « thing » en opposition avec « InternalMachineThing ». Chaque catégorie de CYC doit être une instance d'un seul de ces ensembles. InternalMachineThing est tout qui est interne au fonctionnement de la plateforme CYC (chaînes de caractères, nombres, etc.). « RepresentedThing » est tout le reste. (Sowa critique cette organisation, disant que CYC ne devrait pas être exclu d'une représentation de son propre fonctionnement).
En second lieu, on trouve « IndividualObject » en opposition à « Collection », qui est une autre partition de « Thing ». Collection inclut toutes les catégories mentionnées dans CYC. En conséquence, Collection ne représente pas la masse et est insaisissable. Sowa soutient que ce serait peu clair si l'on trouvait quelque chose telle qu'une volée d'oiseaux (qui est une collection qui est clairement perceptible).
En troisième position est présenté l'« Intangible » en opposition au « TangibleObject » et au « CompositeTangible&IntangibleObject ». Chaque unité dans CYC est une instance d'une de ces trois catégories. Intangible est tout ce qui n'a aucune masse (ensemble de toutes les personnes, numbre42, etc.), tandis que TangibleObject est tout ce qui a de la masse et de l'énergie (un rocher, le corps d'une personne). CompositeTangible&IntangibleObject est quelque chose qui a une ampleur physique et une ampleur intangible. Par exemple, une personne particulière a un corps (ampleur physique) et un esprit (ampleur intangible). Il est intéressant de noter que « Événement » et sa sous-classe, « Process », sont des sous-classes d'IndividualObject.

Applications

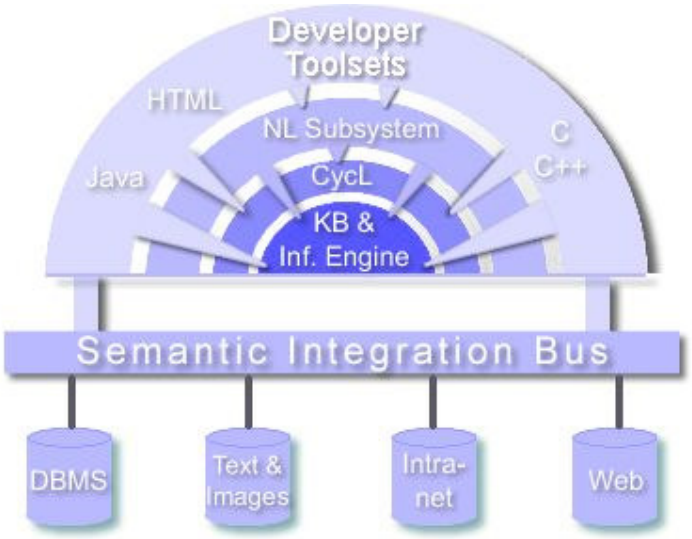
Diverses applications fondées sur cette base de connaissance sont disponibles, principalement CycAnswers et Cyc Knowledge Server.

CycAnswers est une application de gestion de connaissances intégrées et de réponse à des questions. Elle peut manipuler de larges volumes de questions automatiquement et intelligemment.
Cette application répond aux questions par des réponses précises et complètes, grâce à un raisonnement, utilisant des connaissances représentées formellement et provenant de ressources diverses.
CycAnswers surveille l'état de ses connaissances afin de détecter ses incohérence et ses imperfections. Dans cette optique, elle intègre progressivement des bases de données et de connaissances ; elle analyse automatiquement les questions non-répondues et recommande les améliorations (dont les extensions appropriées). De plus elle résiste à une recherche d'information perfectionnée lorsque les questions ne peuvent pas être satisfaites.

CycAnswers est composé de plusieurs composants modulables dont principalement :
- Le CycAnswers Semantic Knowledge Source Integrator ("SKSI"), qui exploite la distinction connaissances/données en enfermant les informations abstraites, générales et continuelles dans une base de connaissances tout en gérant les informations à un niveau individuel et transitoire dans une base de données. Les deux modules peuvent être gérer et optimiser indépendamment.

- Le CycAnswers Integrity Auditor qui inspecte automatiquement l'état de la base de connaissances et de toutes les bases de données qui sont interfacées avec lui pour détecter les incohérences et les lacunes potentielles avant que l'utilisateur final ne soit témoin d'incohérences gênantes ou de réponses incomplètes.

Cyc Knowledge Server



Cyc Knowledge Server est une très large base de connaissances multi-contextuelles et un moteur d'inférence développés par Cycorp. Le but de Cycorp est de construire des fondations aux connaissances de « sens commun » de base (un substrat sémantique des termes, des règles et des relations) pour permettre le développement d'une grande variété de produits et services liés au traitement des connaissances.

Cyc est conçu en vue de fournir une « couche profonde » de compréhension qui peut être utilisée par d'autres programmes afin de les rendre plus flexible.
La technologie de Cyc comprend plusieurs composants essentiels :

1. La base de connaissances Cyc:
La base de connaissance de Cyc est une représentation formalisée d'une vaste quantité de connaissances humaines fondamentales: des faits, des principes de base, et des heuristiques pour le raisonnement au sujet des objets et des événements de la vie quotidienne.
La base de connaissances se compose des termes (qui constituent le vocabulaire de CycL) et d'assertions qui relient ces termes. Ces affirmations incluent à la fois les affirmations et les règles de base.
Cyc n'est pas un système basé sur les frames ; l'équipe de Cyc pense à une base de connaissances à la place, en tant que « mer d'affirmations », où chaque assertion n'est pas plus « au sujet d' » un des termes impliqué qu'un autre. Elle consiste en un ensemble de termes et d'affirmations liées à ces termes. Elle se décompose, par ailleurs, en différentes « microthéories ». Chaque microthéorie rend compte seulement d'un point de vue important d'un domaine de connaissances. Certains domaines peuvent traiter plusieurs microthéories, qui représentent différentes perspectives et affirmations, divers niveaux de granularité et de distinction.
La base de connaissances Cyc est divisée en beaucoup de (actuellement une centaine) microthéories, dont chacune est essentiellement un paquet d'assertions qui partagent un ensemble commun d'hypothèses; certaines microthéories sont concentrées sur un domaine particulier de la connaissance, un niveau particulier de détail, un intervalle particulier de temps, etc. Le mécanisme de la microthéorie permet à Cyc de maintenir des affirmations indépendamment qui sont à première vue contradictoires, et d'améliorer les performances du système de Cyc en concentrant le processus d'inférence.
À l'heure actuelle, la base de connaissances Cyc contient des dizaines de milliers de termes et de nombreuses assertions saisies manuellement « à propos de » ou « impliquant » chaque terme. De nouvelles affirmations sont continuellement ajoutées à la base de connaissances par des humains. Les chiffres mentionnés ci-dessus n'incluent pas les termes non-atomiques comme (#$LiquidFormOf #$Nitrogen), ni le grand nombre d'affirmations ajoutées à la base de connaissances par Cyc lui-même comme produit du processus d'inférence.

2. Le moteur d'inférence Cyc:
Le moteur d'inférence de Cyc effectue des déductions logiques générales (modus ponens, modus tolens, et quantification universelle et existentielle), ainsi que des mécanismes d'inférence bien connus en I.A. (héritage, classification automatique, etc.).
Puisque la base de connaissances Cyc contient des centaines de milliers d'affirmations (des « règles »), beaucoup d'approches généralement adoptées par d'autres moteurs d'inférence (tels que les interpréteurs de commandes interactifs de système expert basés sur les frames, l'unification de RETE, le prolog, etc.) ne suffisent pas pour des bases de connaissances de la taille de celle de Cyc. En conséquence, l'équipe de Cyc a été forcée de développer d'autres techniques.
Cyc inclut également plusieurs modules d'inférences spécifiques pour manipuler quelques classes spécifiques d'inférence. Certains de ces modules s'occupent par exemple du raisonnement d'égalité, du raisonnement temporel, ou du raisonnement mathématique.

CycL: Le langage de représentation Cyc:

CycL, le langage de représentation de Cyc, est un langage de représentation de la connaissance large et flexible. C'est essentiellement un accroissement du calcul de prédicats de premier ordre (FOPC), avec des extensions pour manipuler l'égalité, le raisonnement par défaut, le skolémisation, et quelques fonctions du deuxième ordre (par exemple, on permet la quantification des prédicats dans certaines circonstances, et des affirmations complètes peuvent apparaître comme composants intensionnel d'autres affirmations.) CycL utilise une forme de circonscription, inclut l'hypothèse des noms uniques, et peut se servir de l'hypothèse du monde clos le cas échéant.

Le sous-système de traitement du langage naturel :
Le système de CYC-NL a trois composants: le lexique, le programme d'analyse syntaxique, et « l'interprète sémantique ».
Le composant sémantique de Cyc-NL transforme les analyses syntaxiques en formules en CycL. La sortie du composant sémantique est du CycL " pur ": une phrase analysée peut immédiatement être présentée dans la base de connaissances par exemple, ou une question analysée peut être présentée au générateur de SQL afin de poser une requête de base de données.
Les structures sémantiques sont construites morceau par morceau et combinées en plus grandes structures. Pour chaque règle syntaxique, il y a un procédé sémantique correspondant qui s'applique. La sémantique clausale de Cyc-NL est fondamentalement « verb-driven ». Des verbes sont enregistrés dans le lexique avec des patrons pour leur traduction en CycL.
Le composant sémantique de Cyc-NL se sert des connaissances dans la base de connaissances à pratiquement chaque niveau du procédé d'interprétation. L'utilisation de la connaissance de « sens commun » pour guider le procédé de traduction permet de traiter du problème toujours présent de l'ambiguïté en langage naturel sans avoir à compter seulement sur des techniques statistiques.

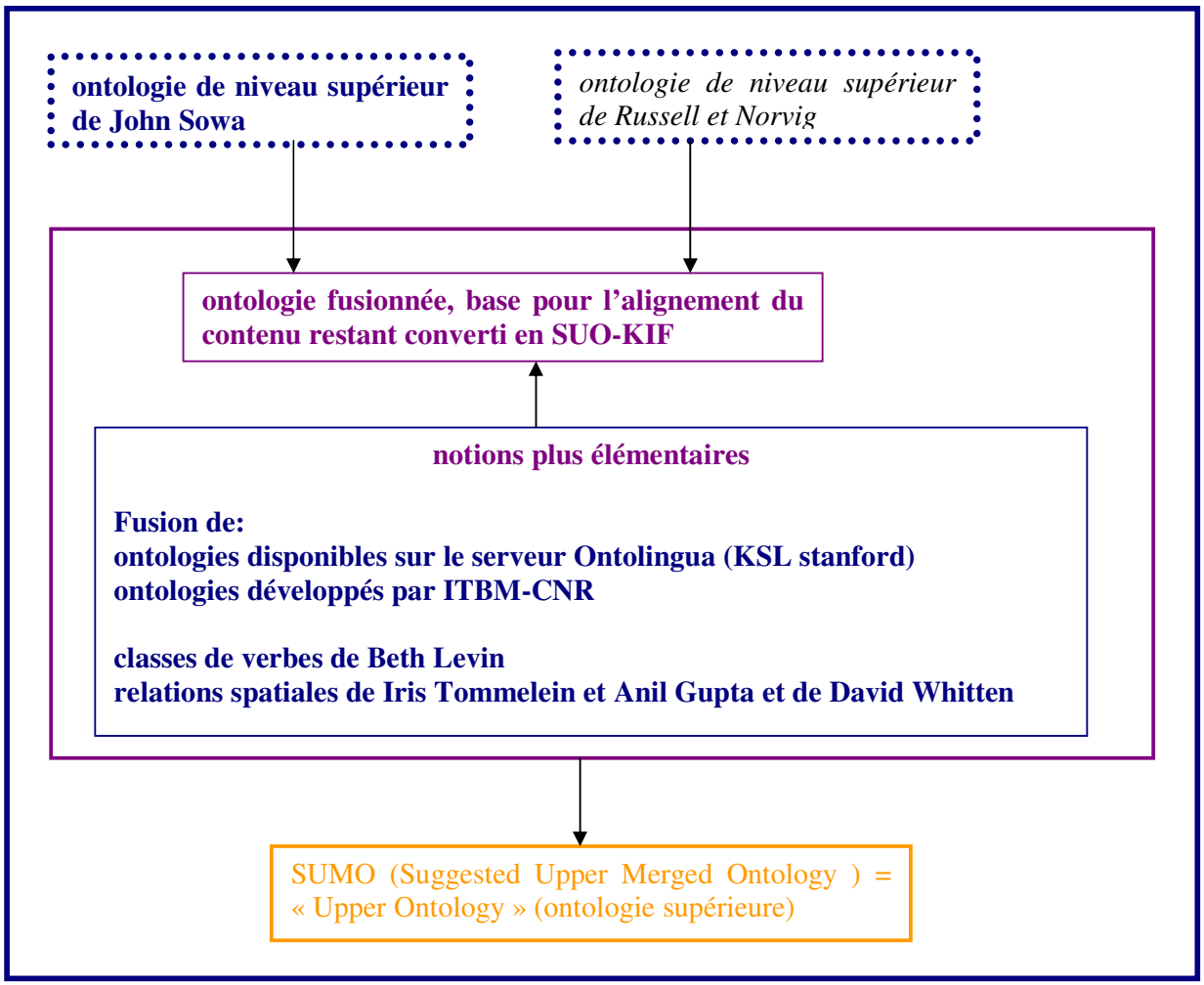
Cycorp développe des interfaces qui permettront à des personnes de faire des affirmations et de questionner CYC en utilisant l'anglais au lieu du CycL. Cycorp travaille également à un composant de génération, qui produira les chaînes de caractères anglaises à partir des formules de CycL.
Les capacités de CYC-NL forment la base pour des applications dans l'amélioration en recherche d'information, et pour les interfaces conviviales à d'autres applications, y compris les applications d'intégration de base de données.
Les futures directions pour CYC-NL incluront:
L'exploration du rôle possible de CYC dans la traduction automatique ;
L'utilisation de CYC-NL pour post-traiter les sorties des systèmes de reconnaissance de la parole.

3.2 SUMO

[SUMO]

Genèse du projet

SUMO (Suggested Upper Merged Ontology) est une « Upper Ontology » (ontologie supérieure) créée par le IEEE Standard Upper Ontology (SUO) Working Group (2001).
L'ontologie a été développée par le groupe de travail du SUO en fusionnant les sources « contenu du candidat » du SUO et en raffinant et en étendant ce contenu sur la base de divers projets d'ingénierie de la connaissance et entrée du groupe de travail de SUO.



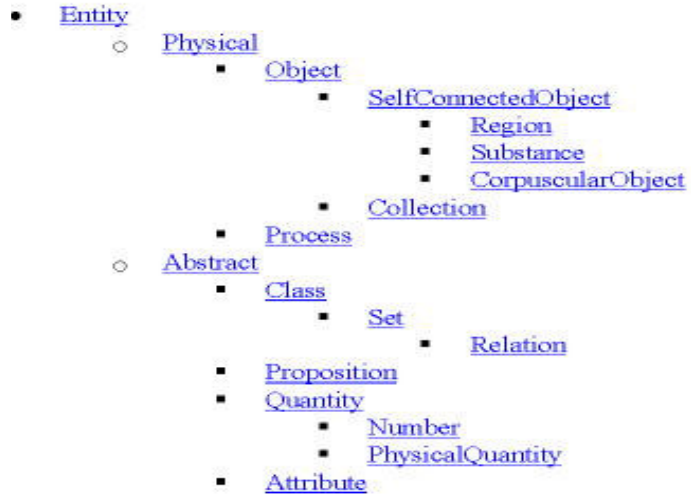
Ces ontologies se divisent en deux classes, à savoir celles définissant des concepts de niveau très élevé (ontologie de niveau supérieur de John Sowa et celle de Russell et Norvig) et celles définissant des notions plus élémentaires. Après que cette séparation a été faite, les deux ontologies de niveau supérieur ont été fondues en une structure conceptuelle simple. Puisque les deux ontologies source sont très compactes et contiennent une quantité significative de contenu superposable, cette fusion n'a pas posé de problèmes pratiques ou théoriques sérieux. Cette ontologie fusionnée a été alors utilisée comme base pour l'alignement du contenu restant converti en SUO-KIF.

SUMO se présente sous forme d'arbre et est formalisé par une version du langage KIF (Knowledge Interchange Format) appelé SUO-KIF
SUMO est sensible à la typographie (les noms des classes et des individus commencent par une lettre majuscule, et les prédicats des noms avec une lettre minuscule).
De plus, il y a des cas d'héritage multiple dans SUMO. C'est-à-dire que certains concepts peuvent avoir plus d'un parent (par exemple : TimeInterval est un enfant de TimeDuration et de TimePosition). En conclusion, chaque concept de SUMO est « hyper-relié » à sa définition dans le browser en ligne d'ontologie. Ces liens permettront à l'utilisateur du browser d'aller directement d'un terme à sa définition formelle complète.

Structure

Ceci n'est qu'une vue d'ensemble des plus importants concepts de SUMO et des dépendances entre ces concepts ([SUMOARB]). Cependant, les fichiers source pour toutes les versions du SUMO peuvent être trouvés à [SUMO01].
La contenu de SUMO peut être divisée par thèmes comme suit:

Principal Distinctions	Basic Binary Relations	Organic
Objects	Artifacts	Temporal Concepts
Processes	Spatial Relations	Mereology
Abstract Entities	Number	Semiotics
Structural Ontology	Measure	



Principal Distinctions:
La liste ci-dessus représente la taxinomie supérieure « Top-level taxonomy » de SUMO, qui incorpore le contenu de beaucoup de sources (John Sowa, Russell et Norvig, et d'ITBM-CNR).
Le noeud racine de SUMO est, comme dans beaucoup d'ontologies, l'Entité, et ce concept englobe immédiatement le concret et l'abstrait. La première inclut tout qui a une position dans l'espace/temps, et la dernière classe inclut tout le reste.
Sous le concept Concret nous avons les concepts disjoints d'Objet (voir ci-dessous) et de Procès (cette distinction représente une controverse ontologique ; En effet, cela signifie que le SUMO assume une orientation en 3D, plutôt qu'une orientation en 4D).
Selon ceux qui adoptent une orientation 3D (ou "les endurantistes, "comme ils s'appellent parfois), il y a une distinction de base et catégorielle entre les objets et des processus. Selon ceux qui adoptent une orientation 4d ("les perdurantistes), " il n'y a aucune distinction telle. L'orientation 3D pose en principe que les objets, à la différence des processus, sont complètement présents à tout moment de leur existence, alors qu'une orientation 4D considère tout comme « une vis sans fin d'espace-temps » (ou une part d'un tel engrenage). Dans la dernière optique, les processus paradigmatiques et les objets sont simplement les extrémités opposées d'un continuum de phénomènes spatio-temporels.

Applications

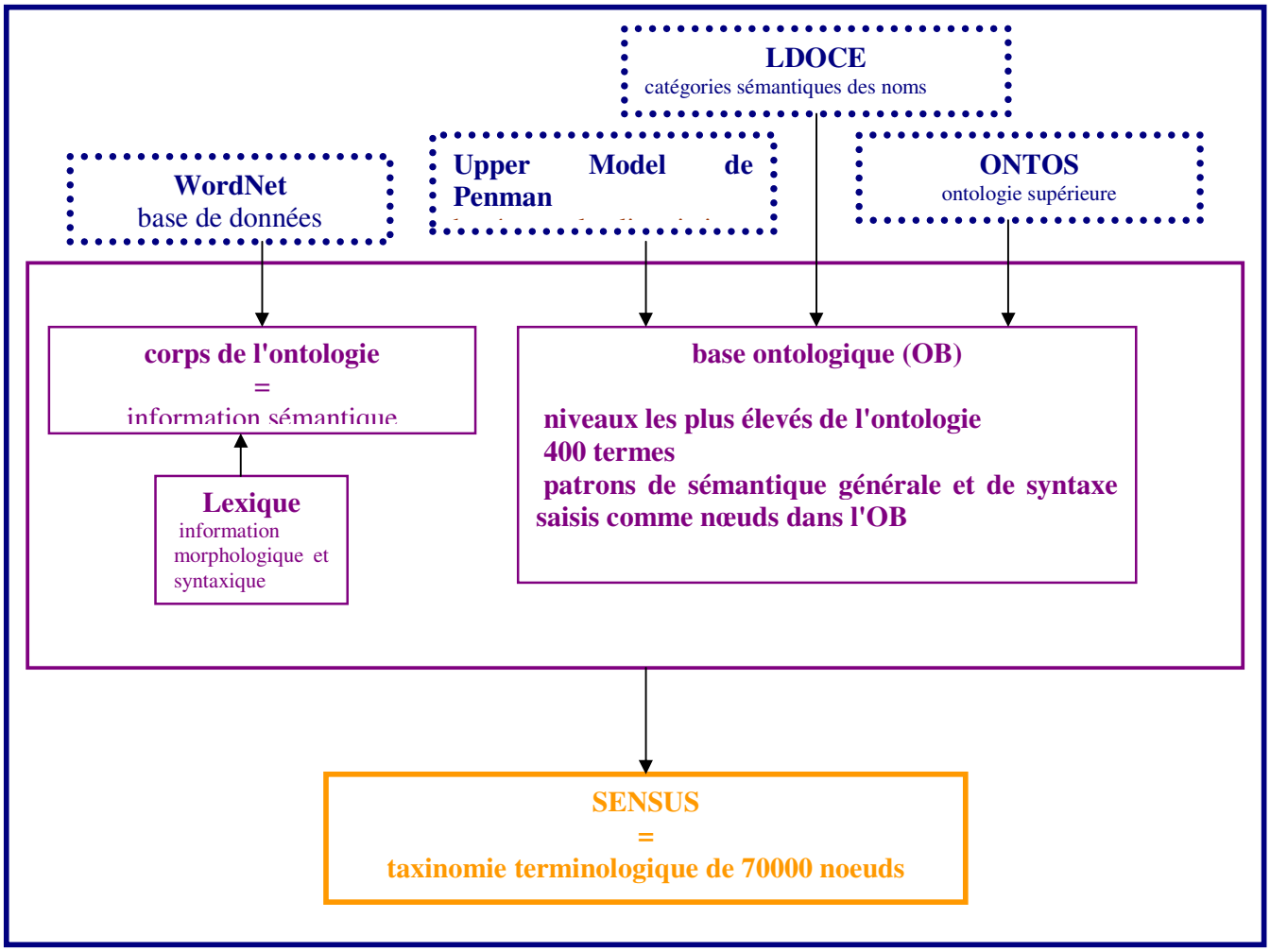
Cette norme va spécifier une ontologie supérieure capable de soutenir des applications telles que l'interopérabilité de données, la recherche de l'information, l'inférencement automatisé, et le traitement de langage naturel.
Les concepts spécifiques à des domaines donnés ne seront pas inclus; cependant, cette norme fournira une structure et un ensemble de concepts généraux sur lesquels des ontologies de domaine (par exemple médical, financier, technique, etc.) pourront être construites.
L'un des champs d'application possible de SUMO est la compréhension de langage naturel. En effet, ces applications de compréhension de langage naturel fonctionne grâce à un système raisonnant basé sur la connaissance qui emploie une ontologie pour désambiguïser des traductions probables d'énoncés de langage naturel.

3.3 SENSUS

Genèse du projet

L'ontologie SENSUS est un projet de l'ISI à l'« University of Southern California » (USC). La construction initiale de SENSUS et de ses algorithmes d'alignement ont été exécutés par Kevin Knight et Steve Luk ; les algorithmes postérieurs et la suite du travail est effectuée par Eduard Hovy et Bruce Jakeway.
SENSUS est une ontologie basée sur le langage naturel qui a pour fonction de fournir une vaste structure conceptuelle aux travaux menés en matière de traduction automatique. Il a été mis au point en rassemblant et en extrayant des données de ressources électroniques telles que ; Penman Upper model, Ontos, WordNet (George Miller, Christiane Fellbaum; Université de Princeton) et des dictionnaires électroniques de langages naturels.
L'ontologie est représentée en Loom, FrameKit, et Prolog.
SENSUS est une taxinomie terminologique de 70000 noeuds, qui sert de cadre dans lequel on peut ajouter des connaissances supplémentaires.

Le premier objectif de SENSUS est la création et l'utilisation de grandes taxinomies de concepts (50000 ou plus) et d'ontologies pour le traitement du langage naturel en combinant les ressources en ligne telles que des dictionnaires et des thesaurus, des méthodes statistiques adaptées au texte, et des interfaces d'acquisition de connaissances humaines traditionnelles. En particulier, créer et organiser une taxinomie de concepts de 70.000 éléments pour une utilisation dans PANGLOSS (traduction automatique), ou PENMAN (génération de phrase) et par la suite dans d'autres systèmes.
Cette recherche aborde le besoin d'acquérir de larges ressources en connaissances sémantiques et lexicologiques, à la fois pour le travail spécifique de PENMAN et pour permettre le partage de la connaissance entre les modules de PANGLOSS et d'autres sites.



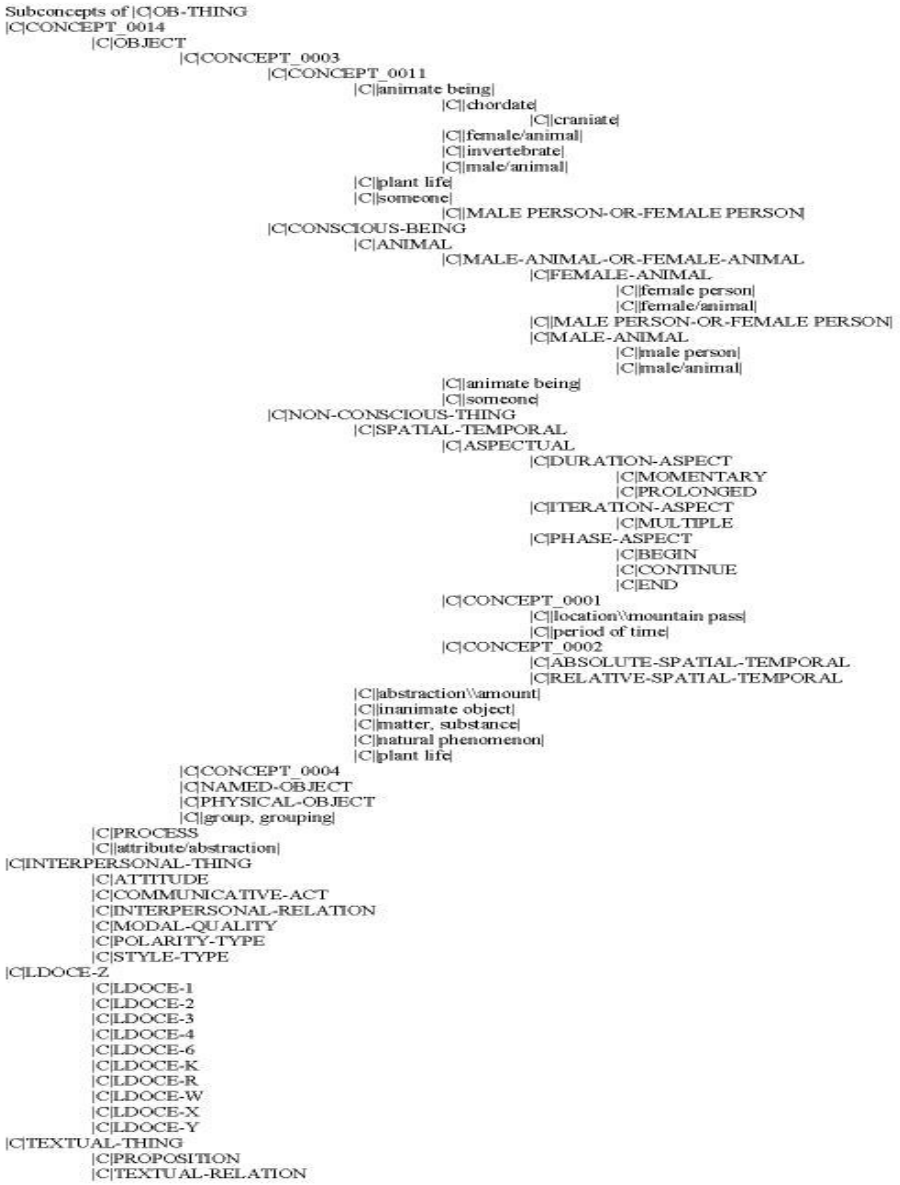
Elle maintient les distinctions présentes dans le Upper Model de Penman de sorte que tous les termes de l'ontologie subordonnée peuvent être correctement générés en anglais ; elle maintient les catégories de LDOCE de sorte qu'ULTRA[1] puisse faire les distinctions nécessaires en analysant les noms ; et elle maintient les distinctions d'ONTOS de sorte que l'analyse sémantique puisse se dérouler correctement.
La source primaire pour le corps de l'ontologie est la base de données sémantique WordNet. Pour construire le corps principal de l'ontologie, le travail a été effectué automatiquement en reliant des concepts de WordNet et des éléments lexicaux anglais en découvrant des paires de sens correspondants. En plus de contenir les symboles pour représenter la signification sémantique, l'ontologie contient des pointeurs de chaque symbole vers les éléments lexicaux appropriés dans divers langages. Le lexique anglais de Penman contient actuellement environ 50000 formes orthographiées (correspondant à approximativement 90000 mots) ; le lexique japonais de Japangloss contient plus de 120000 mots.

Structure
[LOPEZ99]

[1] ULTRA : Nom du projet

La version actuelle a commencé par la fusion du PENMAN Upper model et ONTOS, ainsi que des catégories sémantiques d'un dictionnaire pour produire une base ontologique. WordNet a ensuite été ajouté.
SENSUS peut être considérée comme une extension et une réorganisation de WordNet ; au niveau supérieur, ont été ajoutés des nœuds du Upper Model de Penman, et les branches principales de WordNet ont été réarrangées pour s'adapter. En outre, des nœuds basés sur des travaux avec d'autres ontologies sont venus compléter.
Enfin, des entrées lexicales japonaises et espagnoles ont été apportées.

Organisation des concepts de SENSUS : [SENSARB]



Applications

SENSUS peut être utilisée dans le développement, ou la création d'ontologies. Voici les étapes de construction d'une ontologie à l'aide de SENSUS. [LOPEZ99]

- Tout d'abord, une série de termes est prise comme 'seeds' (racines) * ces termes sont reliés à la main à SENSUS ;
- Ensuite, tous les concepts dans le chemin qui va du concept 'seed' à la racine sont inclus ;
- Puis, les termes qui pourraient être pertinents pour le domaine et qui ne sont pas encore apparus sont ajoutés ;
- Et enfin, pour les nœuds ayant de nombreux chemins les traversant, le sous-arbre complet du nœud est quelque fois ajouté (Etape manuelle).

Cette méthodologie casse le modèle actuel car elle se base sur un ajout de termes à une ontologie existante, SENSUS, qui est ensuite enlevée.
La stratégie d'identification des concepts est la suivante. Elle est dite « bottom-up ». Les concepts les plus spécifiques de l'application sont recherchés. Ensuite, la méthode « recherche et enlèvement » est appliquée pour entrer plus de concepts abstraits.
SENSUS étant fortement orientée traitement du langage naturel, l'ontologie a été utilisée tout d'abord en traduction automatique, mais aussi en outil de résumé de texte automatique multilingue (SUMMARIST).
Des outils ont été développés par rapport à Sensus comme par exemple Ontoseek ou Ontosaurus, un visualisateur avec lequel SENSUS peut être parcouru.

3.4 ONTOS

Genèse du projet

ONTOS( ) a été conçue directement dans l'optique de T.A.L.. Tout d'abord développée pour le système, de traduction automatique Dionysus (CMU, « Carnegie Mellon University »), elle a servi ensuite à la création d'ontologie telles que Mikrokosmos et PANGLOSS [BOUIL98].
Les auteurs de ce système de modélisation conceptuelle utilisant un formalise à schémas, sont principalement Sergei Nirenburg, Ira Monarch, Todd Kaufmann et Lynn Carlson. Ce travail a débuté en 1988 et a eu cours jusque 1990 [ONTOS02].
Dans le projet de Dionysus à CMU le langage de représentation du sens des textes s'appelle Tamerlan (c'est un langage formel pour représenter le sens des textes en langue naturelle). Selon [BOUIL98], le formalisme utilisé pour représenter l'ontologie est basé sur le langage de représentation des connaissances FrameKit qui permet la définition d'un ensemble de schémas.

Structure

ONTOS contient essentiellement des concepts généraux, interlinguistiques, qui sont reliés par des liens ontologiques et forment un réseau. La modélisation d'un domaine complémentaire et indispensable à toute application du système.

La modélisation du monde est prise en charge par le système ONTOS, qui se compose a) d'un langage de contrainte, b) d'une ontologie, ou d'un ensemble de concepts généraux, c) d'un ensemble de modèle de domaines et d) une interface intelligente d'acquisition de la connaissance.
Les dispositifs de base du langage de contrainte d'ONTOS sont comme suit. Un modèle du monde est une collection de frames. Une frame est un ensemble fixé de fentes, interprété comme un concept ontologique (événement volontaire-olfactif, entité géopolitique). Une fente représente une propriété ontologique (la température, causé-par) et se compose d'un ensemble prédéfini de facettes. Une facette est un ensemble fixé de « fillers » (symbole, nombre,… de remplissage).

L'ontologie contient trois types de concepts : les objets, les événements et les propriétés. Celles-ci constituent avec les liens is-a, les relations les plus importantes. Les concepts sont organisés en treillis et caractérisés par un grand nombre de propriétés dont héritent par défaut les concepts descendants.

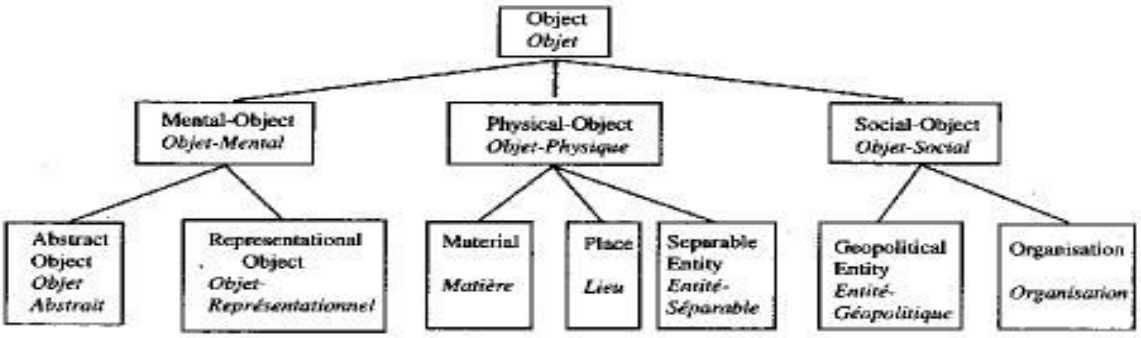


Figure (25) : Sous-réseau d'objet

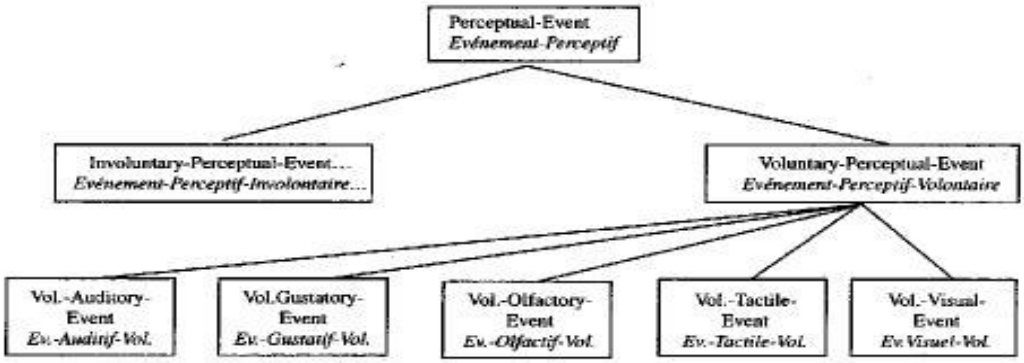


Figure (26) : Sous-réseau d'événements perceptifs

[BOUIL98]

Applications

Tout d'abord utilisée pour la traduction automatique, l'ontologie ONTOS, linguistiquement neutre, est réutilisée dans divers projets multilingues tels que SMEARR (mappage de sémantique linguistique en sémantique informatique ; 1988-91 Victor Raskin), Pangloss, ou Mikrokosmos.

3.5 Generalized Upper Model

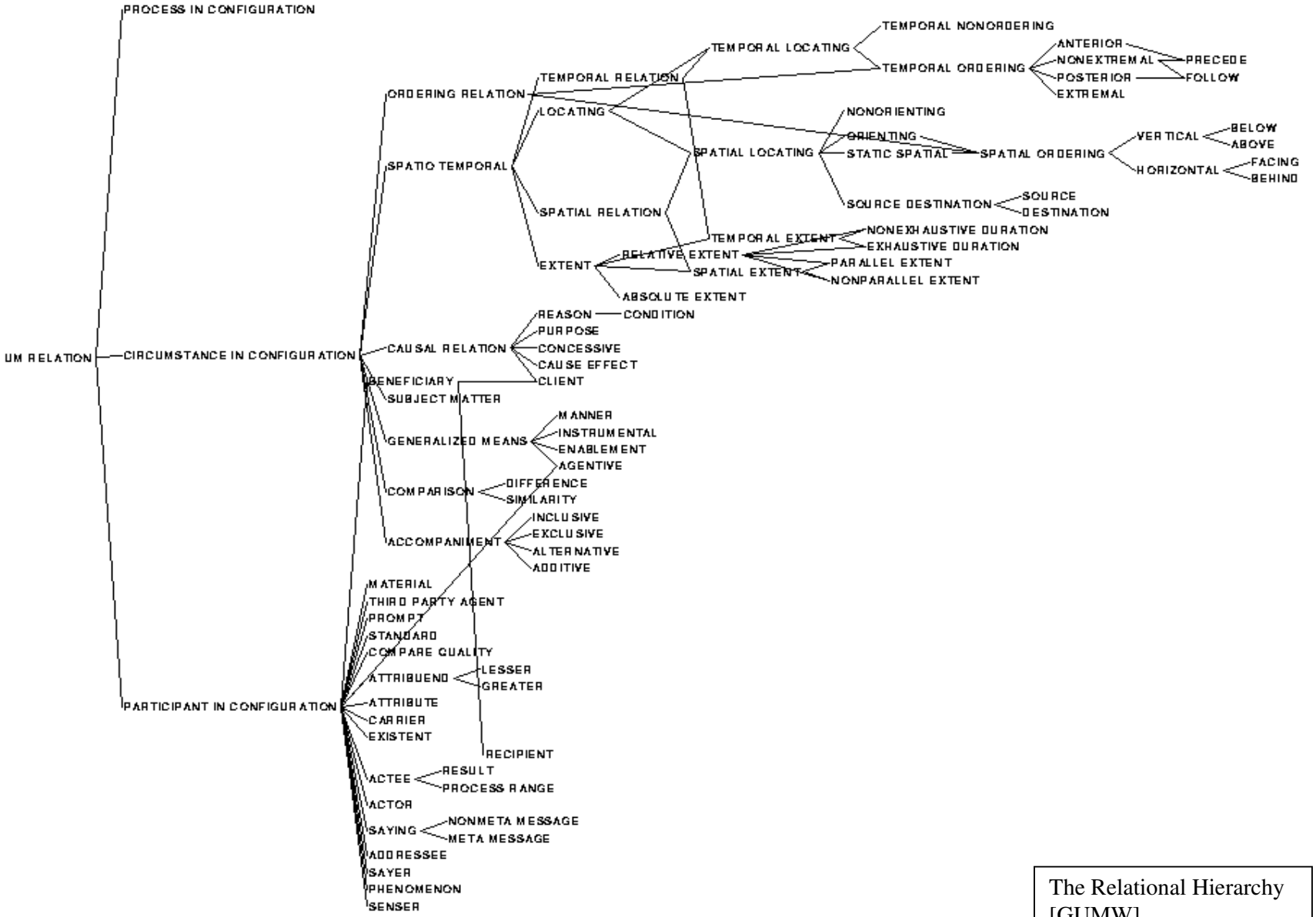
Genèse du projet

Le Generalized Upper model (GUM) a été développé par Bateman, Magnini, et Rinaldi vers 1994.

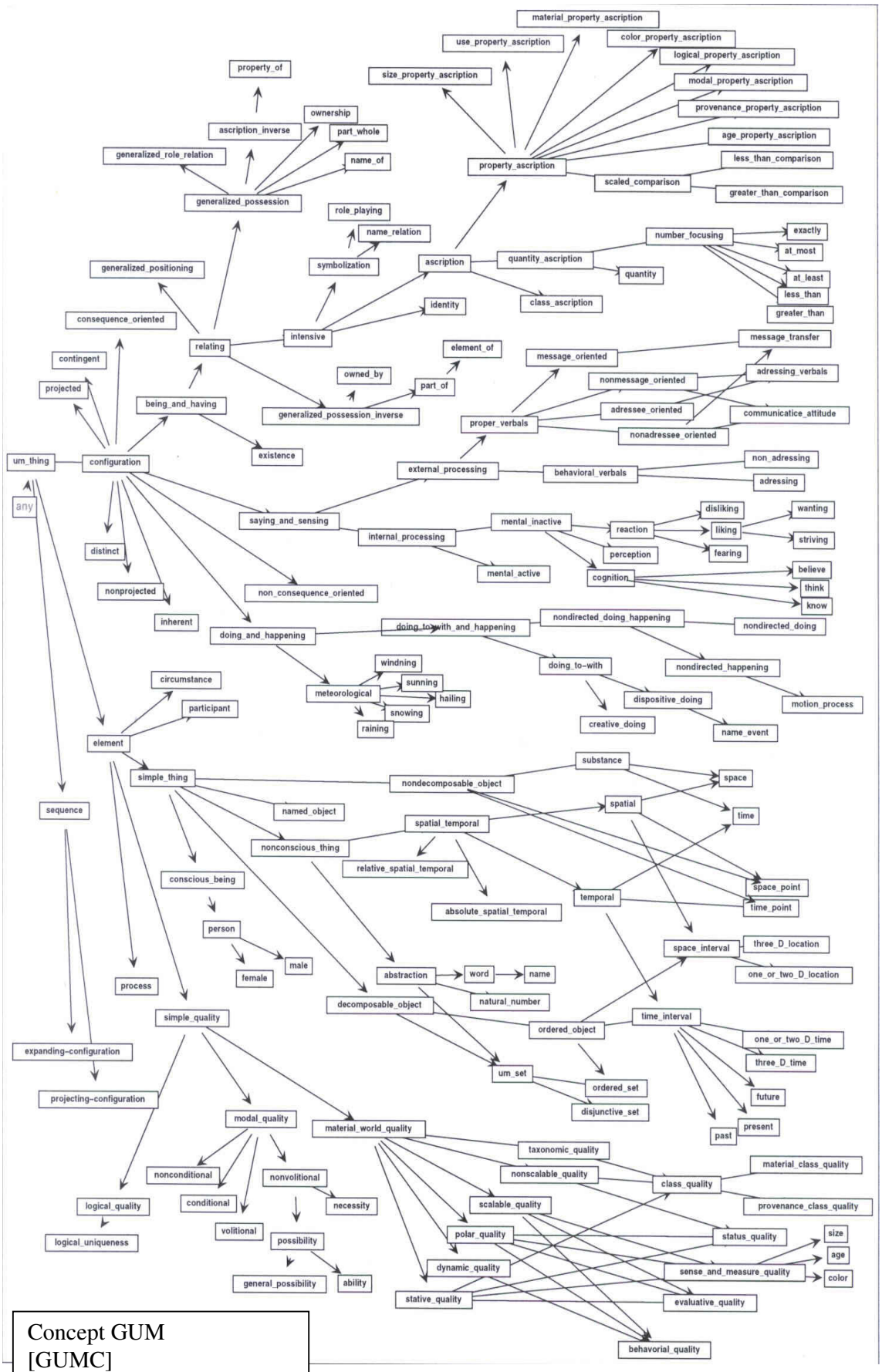
Pour le GENERALIZED UPPER MODEL, qui a étendu une ontologie assez large déjà existante (PANGLOSS) au travail sur différentes langues, le processus de conception était le suivant.
D'abord, il y avait le UPPER MODEL de PENMAN pour l'anglais ; ensuite, l'allemand a été ajouté, et le MERGED UPPER MODEL a été créé.
Le GUM est l'extension du MERGED UPPER MODEL afin de couvrir les trois langues suivantes: (1) anglais, (2) allemand, et (3) italien. Pour chaque sous-hiérarchie du MERGED UPPER MODEL, l'ensemble des comportements linguistiques pertinents en italien a été identifié. Ces comportements ont été alors comparés à l'anglais. Si les comportements italiens et anglais-allemands étaient compatibles, aucune modification n'était nécessaire ; autrement, une modification était proposée, et le modèle anglais-allemand était réévalué.
L'ontologie GUM est un modèle d'organisation générale de concepts défini dans le langage de représentation des connaissances LOOM.

Structure

GUM est une ontologie linguistique générale, indépendante de tout domaine et de tout type de tâche. Afin de pouvoir la transférer dans différentes langues, il a été prévu que l'ontologie Gum n'inclue que les notions linguistiques principales et leur organisation dans toutes les langues ; elle omet ainsi les détails qui différencient les langues. Cette philosophie a permis d'utiliser Gum pour créer des ontologies pour des langues spécifiques, telles que l'anglais, l'allemand, l'espagnol et l'italien en rajoutant les traits sémantiques propres à chaque langue.



The Relational Hierarchy [GUMW]



Concept GUM
[GUMC]

Le GUM est une ontologie linguistiquement motivée de tâche générale et de domaine-indépendant qui permet le traitement du langage naturel sophistiqué en plusieurs langues. Son niveau d'abstraction est entre la connaissance lexicologique et la connaissance conceptuelle. Il prétend simplifier l'interface entre la connaissance de domaine spécifique et les ressources linguistiques générales. Le modèle propose une organisation de domaine et consiste seulement en une taxinomie (avec la prétention que les outils de traitement du langage naturel qui l'utiliseront encoderont l'information axiomatique dans le code de traitement du langage naturel lui-même).
Il existe une hiérarchie étendue des concepts (environ 250) aussi bien qu'une hiérarchie séparée des relations (um-relation). Au faîte de la hiérarchie de concept (voir schéma) se trouve le concept « umthing » qui représente les phénomènes ou la situation les plus généraux. Il est subdivisé en trois sous-types principaux : (1) Configuration, « une configuration d'éléments participant tous à une certaine activité ou état d'affaires » ; (2) Element, un terme conceptuel seul ; et (3) Sequence, « une situation complexe où les diverses activités ou configurations sont reliées par une quelconque relation pour former une séquence ».
Des relations sont alors employées pour relier des éléments dans des configurations et des séquences. La plupart des relations sont entre un processus et ses participants, la manière, et ainsi de suite. Ainsi, la catégorie um-relation est subdivisée en processus dans la configuration, en circonstance dans la configuration, et en participant dans la configuration. La relation causale, par exemple, serait alors l'une des circonstances dans des relations de configuration, et l'attribut serait l'un des participants dans des relations de configuration.

Applications

La généricité du système permet une grande réutilisabilité.
Puisque, les motivations pour les concepts du GUM sont tirées d'une évidence linguistique, le modèle lui-même convient très bien aux applications de traitement de langage naturel.
Les utilisations prévues du modèle supérieur généralisé sont donc :
- une proposition pour l'organisation de modèle de domaine ;
- un niveau d'organisation pour l'interfaçage de modèles de domaine et de composants de langue naturelle.

L'utilisation la plus étendue du GUM et de ses prédécesseurs a été jusqu'ici dans le contexte de la génération de textes ; tout d'abord monolingue pour l'anglais, et maintenant de plus en plus souvent multilingue. Le GUM est actuellement utilisé dans l'environnement de génération multilingue du KOMET-PENMAN développé au GMD/IPSI (KOMET et KPML). Actuellement, le système génère du texte cohérent en anglais, allemand et néerlandais. Des extensions à d'autres langues sont en cours de développement.

3.6 WordNet/EuroWordNet

(◄)

Genèse du projet

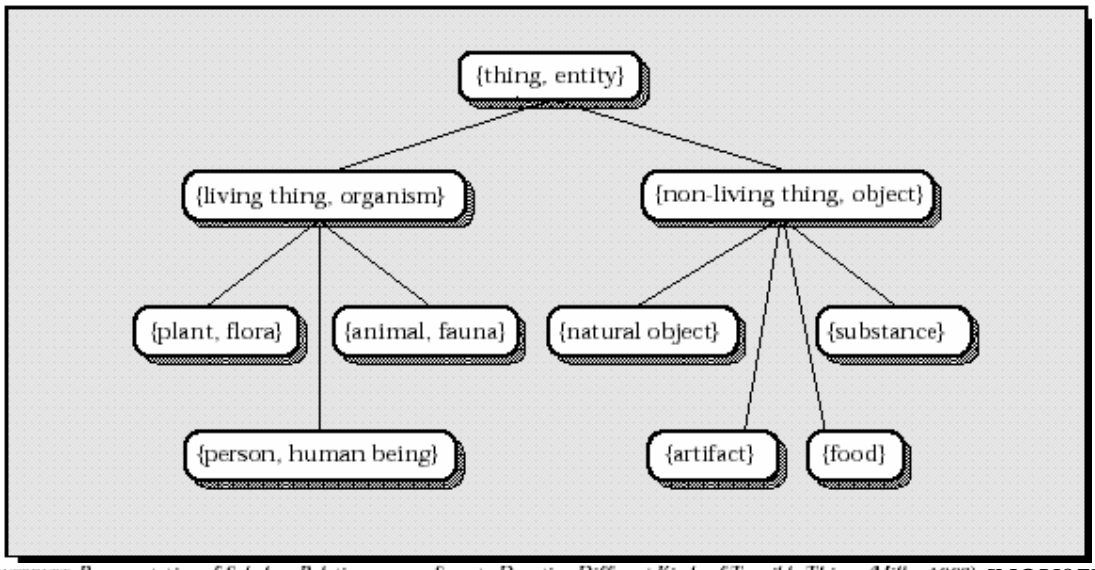
L'une des ontologies lexicologiques les mieux développées est WordNet [MILL90]. WordNet est une base de données lexicale pour l'anglais fondée sur des principes psycholinguistiques. Ce système de référence lexicologique en ligne est construit manuellement.

Sa naissance remonte à 1985. Le projet est amorcé par les mots du corpus Brown. Elle ouvre sur une vision « relationnelle », qui s'oppose à la vision componentielle, atomisante issue de Katz et Fodor.
En 1989, il devient indispensable d'ajouter des définitions puisque passée une certaine taille, les relations ne suffisent plus. En juin 1991, la version 1.0 paraît, et en 2001, on en est à la version 1.7. Celles-ci sont gratuites pour les différentes architectures.
Les objets lexicologiques dans WordNet sont organisés sémantiquement (avec la distinction de base entre les noms, les verbes, les adjectifs, et les adverbes). Ses informations sont ventilées en unités appelées « synsets » en anglais, qui sont des jeux de synonymes interchangeables dans un contexte particulier utilisés pour représenter différents sens. Si un mot a plus d'un sens, il apparaîtra dans plus d'un synset.
Selon [GARD02] la version 1.5 (1997) comporte 168 000 mots et 91 600 synsets. Sa couverture est supérieure à celle du *Petit Robert* [HABE01].

Les spécifications du projet EuroWordNet ont été terminées été 1999. La conception, de la base de données, des relations définies, de l'ontologie supérieure et de l'Index Inter-Langue sont maintenant gelées. Néanmoins, beaucoup d'autres Instituts et groupes de recherche développent des WordNets semblables dans d'autres langues (européennes et non-européennes) utilisant le cahier des charges d'EuroWordNet. Si compatibles, ces WordNets peuvent être ajoutés à la base de données et, par l'intermédiaire de l'Index, reliés à n'importe quel autre WordNet.
EuroWordNet est donc un ensemble de réseaux monolingues inspirés de WordNet et reliés entre eux. Il est crée dans une optique multilingue. Plusieurs objectifs sont liés à ce projet, tels que la construction de réseaux monolingues (qui sont des ontologies linguistiques qui effectuent des inférences) ; la recherche d'information translinguistique (grâce à l'augmentation des synonymes) et une série de traits enregistrés pour chaque langue (les 20 à 50000 mots les plus fréquents d'une langue ; des domaines hiérarchisés ; des liens entre parties du discours (Xpos-synonymy, par exemple) ; une ontologie générale (Top Ontology, Base Concepts) ; l'ajout d'étiquettes aux relations et de nouvelles relations.
Il semblerait que des WordNets soient actuellement développés pour le suédois, le norvégien, le danois, le grec, le portugais, le basque, le catalan, le roumain, le lithuanien, le russe, le bulgare, le slovène. [EUROW]

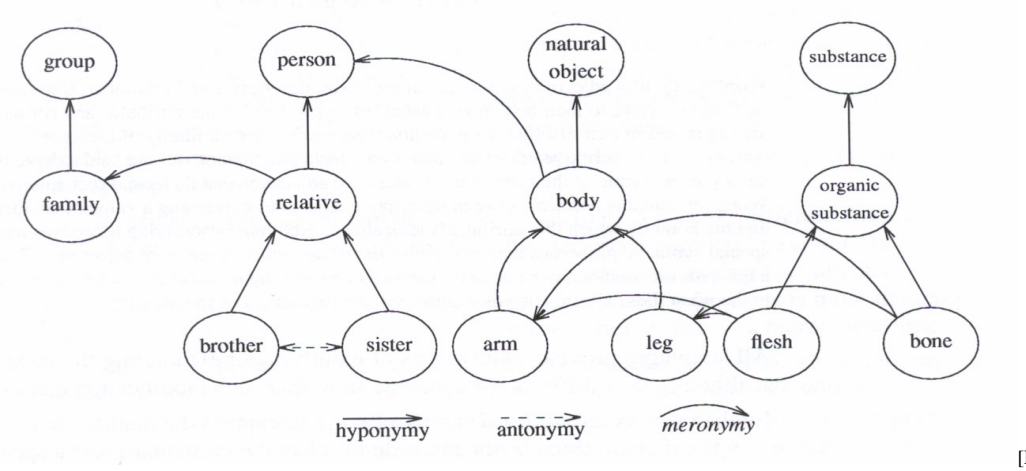
Structure

WordNet contient une série de paires (w, m) où w est une série de caractères ASCII et m un élément d'un ensemble de sens, ou synset. Les synsets sont accompagnés dans leur plus grand nombre de glossaires explicatifs, et ils sont organisés en réseau sur la base de relations sémantiques, au nombre desquelles : l'antonymie, l'hyponymie, la métonymie, l'implication.
Les synsets sont organisés en hiérarchie de sous- et super-classe (désignée sous le nom hypéronymie - hyponymie). Une partie de la hiérarchie de WordNet, celle de « tangible things » est présentée dans le schéma ci-dessous. Pour chaque concept (synset), il y a un pointeur vers les noms représentant ses parties. Par exemple, les parties du concept oiseau sont bec et ailes.



WORDNET: Representation of Subclass Relation among Synsets Denoting Different Kinds of Tangible Things (Miller 1990). [NOY97]

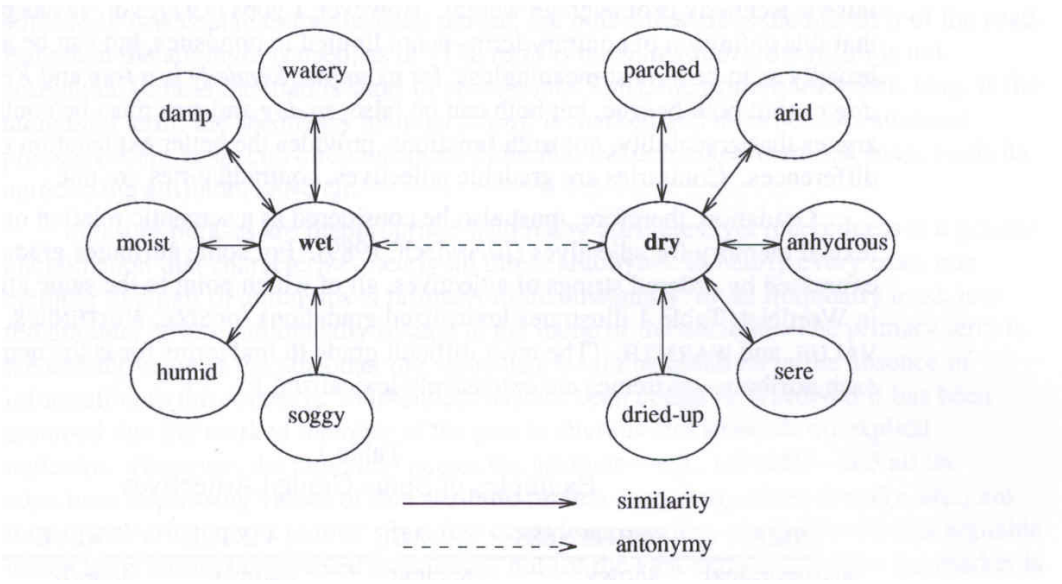
Il y a une tolérance dans l'implémentation de WordNet pour d'autres types de pointeurs (par exemple, du nom au verbe pour représenter des fonctions ou à l'adjectif pour représenter des propriétés), mais ces types de pointeurs n'ont pas encore été implémentés.
WordNet se compose de quatre réseaux qui représentent les catégories syntaxiques principales : noms, verbes, adjectifs et adverbes.
WordNet est une taxinomie ; cette ontologie n'a pas de concepts ou d'axiomes structurés. WordNet utilise une seule taxinomie pour les synsets de noms mais il utilise une organisation différente de synsets pour les verbes et les adjectifs.



[MILL90]

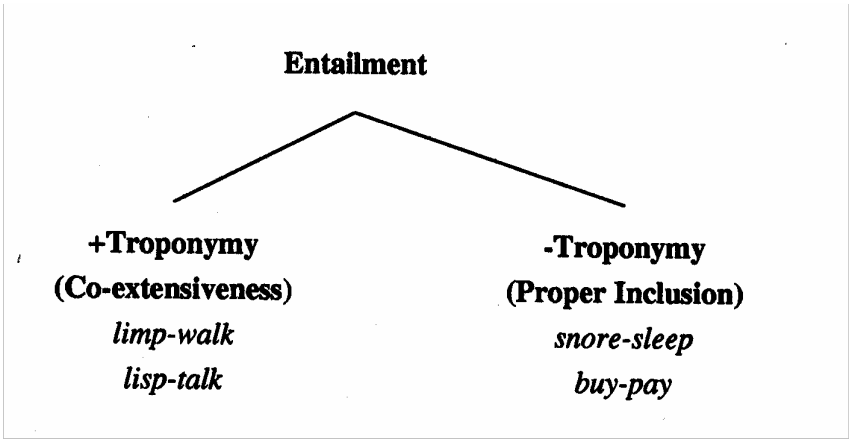
network representative lexical concept.

Les adjectifs descriptifs sont organisés en amas (clusters) bipolaires basés sur l'antonymie. Par exemple, un cluster bipolaire est produit par « sec » et « humide » avec des synonymes de chacun des adjectifs sur la face correspondante du cluster. Des adjectifs apparentés, tels qu'en anglais « fraternal » dans « fraternal twins », sont organisés seulement en synsets avec des pointeurs vers les noms correspondants.



bipolar adjective structure [MILL90]

Les verbes dans WordNet sont divisés en 15 clusters selon leur sens, avec une relation de substitution en tant que relation primaire entre les verbes dans un amas. La plupart de ces amas correspondent à des domaines sémantiques : verbes de soin ou de fonctions corporels, verbes de changement, de connaissance, de transmission, de concurrence, etc. Les verbes, comme suffire, appartenir, ou ressembler qui n'appartiennent à aucun des domaines sémantiques et ne se rapportent pas à des états, forment un fichier séparé.



two kinds of entailment with temporal inclusion

Chaque EuroWordNet est composé comme suit :
- L'index inter-lingue, qui consiste en une liste d'enregistrements sous la forme de "synsets" (ensembles/réseaux sémantiques, principalement issus de WordNet5.1 ou créés manuellement).

- Ontologie supérieure : une ontologie de 63 classes sémantiques de base reposant sur des distinctions fondamentales.
La construction d'une ontologie générale (Top-ontology, Top Concepts – TC-), est vue plutôt comme un ensemble de traits sémantiques classificatoires. Il n'existe pas d'héritage multiple au sein de cette ontologie, mais un concept de base peut renvoyer à plusieurs concepts généraux ou TC.

Trois types d'entités sont disponibles: [HABE01].

1er ordre

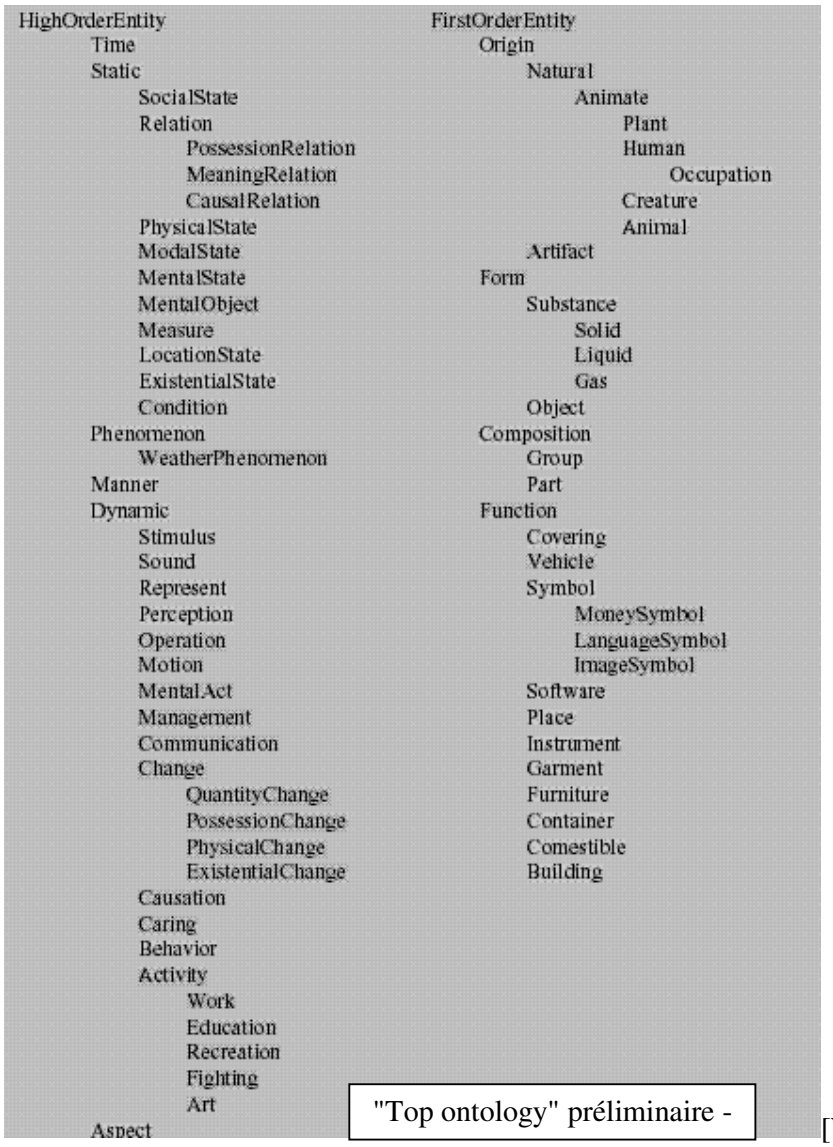
concrètes, perceptibles, situées dans le temps et dans l'espace : *table*. Qualia (Pustejovsky,) : origine (naturelle / artéfactuelle) ; forme ; composition (méronymie) ; fonction.

2nd ordre

situations dynamiques ou statiques, situables dans le temps : *se produire, continuer*. Type de situation (dynamique bornée ou non / statique : propriété / relation) ; composants (cause ; condition ; but...).

3ème ordre

propositions existant indépendamment du temps et de l'espace, vraies ou fausses plutôt que réelles : *idée, pensée, information*.



"Top ontology" préliminaire - [VOSS97]

- Ontologie de domaine : une ontologie de domaines sujets assignés aux enregistrements de l'index inter-lingue

- Une sélection d'enregistrements de l'index inter-lingue, concepts de base, qui jouent un rôle majeur dans les différents WordNets.
Les concepts de base « Base Concepts » (BC) sont les concepts les plus abstraits (vagues et polysémiques) groupés en 79 classes sémantiques. Une trentaine de synsets sont pris en compte au départ (24 synsets nominaux et 6 synsets verbaux), (intersection des choix par langue et lien à WordNet), puis 871 (694 noms, 177 verbes, concepts proposés au moins par deux sites), pour arriver enfin à 1024.

- WordNet1.5 (91591 synsets; 168217 sens; 126520 mots d'entrée) au format EuroWordNet.

Ex : Organisation notionnelle (Péchoin 92) et *vendre*

1. la société
    1. le rapport à l'autre / les comportements / la trahison
    2. les activités économiques
        1. le commerce et les biens
            1. cession
            2. commerce
            3. marchandise
        2. l'économie / cherté
2. le monde / le temps / avance
3. l'homme / le projet et son résultat / projet

1. trahir 597.10 [la société / le rapport à l'autre / les comportements / la trahison] **Trahir** ;
2. 823.10 [la société / les activités économiques / le commerce et les biens / cession] **Céder** ;
3. 827.24 [la société / les activités économiques / le commerce et les biens / commerce] **Acheter 835**, vendre, revendre ;
4. 828.21 [la société / les activités économiques / le commerce et les biens / marchandise] Distribuer, vendre **827** ;
5. *vendre au-dessus des cours* 832.6 [la société / les activités économiques / l'économie / cherté] **Faire payer cher**, vendre au-dessus des cours. - Voler ou rançonner le client ;
6. 180.9 [le monde / le temps / avance] Devancer, **précéder** *vendre la peau de l'ours avant de l'avoir tué* ;
7. 534.16 [l'homme / le projet et son résultat / projet] **Imaginer 404**, s'imaginer, **rêver** *vendre la peau de l'ours avant de l'avoir tué*.

Applications

Malgré l'intérêt de WordNet, certaines limites lui sont reconnues. En particulier l'absence, d'informations sur les propriétés syntaxiques de mots, de liens entre parties du discours (noms, verbes et adjectifs sont des continents séparés). De plus WordNet ne contient pas de savoir « encyclopédique » (à l'inverse du projet d'extraction de connaissance MindNet de Microsoft Research Team), ni d'indications de domaine. Sa densité est variable et le grain de distinction de sens est parfois trop fin (break : 63 sens).

EuroWordNet enrichit les relations disponibles dans WordNet sur le plan de la conjonction/disjonction, du factitif (*tuer* / *mourir*) et non factitif (*chercher* / *trouver*). De même, on peut faire des distinctions au coup par coup ou converses systématiques, des négations pour traiter des exceptions, utiliser des relations sémantiques trans-parties du

discours, intégrer des « cas sémantiques » pour les verbes (agent, patient, instrument, lieu, source, but, résultat, manière) et les noms (Involved_agent).
L'alignement trans-linguistique est possible grâce à l'Index Inter-Langue (ILI).

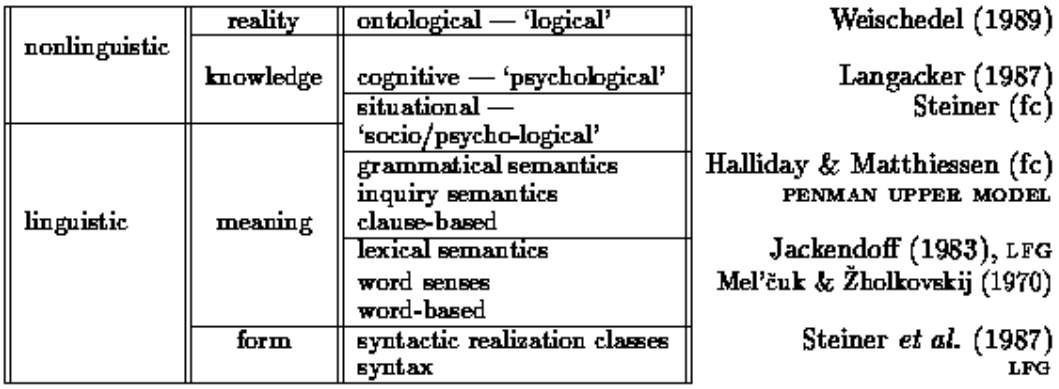
Cependant, pour EuroWordNet, la constitution de corpus suffisamment vastes pour représenter correctement tous les sens des mots polysémiques est hors de portée et certains problèmes subsistent : [HABE01]
- Les relations sémantiques ont une transitivité réduite : {*maison*} a-pour-partie {*porte*} a-pour-partie {*poignée*} mais, pas {*maison*} a-pour-partie {*poignée*}.
- Lier les hiérarchies est une commodité, voire une nécessité (cf. l'ajout d'une racine postiche pour les hiérarchies séparées de noms dans WordNet et les liens multiples dans EuroWordNet). Mais il n'y a pas d'« Être suprême ». - Les réseaux sémantiques naturels sont des buissons et des taillis peu reliés.
- Les sens ne sont pas toujours discrets

A l'avenir, il faut s'accorder sur un dictionnaire sémantique pour la cardinalité et la structuration des catégories (par rapport aux étiquettes morpho-syntaxiques et aux types de constituants).
Il faut aussi veiller à l'évolution diachronique en effectuant des remaniements par élagage et des ajouts incessants (le répertoire d'étiquettes morpho-syntaxiques varie très peu, en comparaison). Parallèlement, il existe des décalages en synchronie (des domaines « densifient » des zones du réseau.).
Et enfin, il faut probabiliser les sens grâce à une sensibilité plus grande au domaine et à la période.

Le principal problème reste que pour le français, les ressources adéquates sont quasi-inexistantes, n'étant ni électroniques, ni de large couverture ni de surcroît, gratuites.

Pour récapituler, Sensus (Swartout et al. 1997), le Generalized Upper model (Bateman et al. 1995), et WordNet (Miller 1990), représentent le mieux les ontologies linguistiques.
GUM est une ontologie linguistique indépendante du domaine et de tâche générale (general task).
WorNet est une base de données lexicales organisée en synsets interchangeables dans des contextes particuliers. SENSUS est une ontologie basée sur le langage naturel, et développée par la fusion et l'extraction d'informations de ressources électroniques.
Toutes trois ont en commun d'être réutilisées et fusionnées de manière à pallier leurs faiblesses propres. Une ontologie générale est difficilement utilisable directement, mais combinée propose de grands avantages.
Cyc est une ontologie qui semble être très développée, mais reste un travail privé qui ne propose donc pas des ressources gratuites et disponibles comme les travaux universitaires.

Il existe des ontologies très diverses. Ceci est dû aussi bien aux choix de conception au niveau des formalismes et des théories sous-jacentes (orientation plus ou moins formelles, plus ou moins logique..) qu'aux applications prévues pour ces base de connaissances.



nonlinguistic	reality	ontological — 'logical'	Weischedel (1989)
	knowledge	cognitive — 'psychological'	Langacker (1987)
		situational — 'socio/psycho-logical'	Steiner (fc)
linguistic	meaning	grammatical semantics inquiry semantics clause-based	Halliday & Matthiessen (fc) PENMAN UPPER MODEL
		lexical semantics word senses word-based	Jackendoff (1983), LFG Mel'čuk & Žholkovskij (1970)
	form	syntactic realization classes syntax	Steiner *et al.* (1987) LFG

[BAT90]

Certaines ontologies ont aujourd'hui une très large couverture. C'est le cas dans les ontologies de TAL, comme GUM, SENSUS, WORDNET (et EUROWORDNET) ou CYC, qui pour la plupart ne s'arrête plus à un traitement monolingue mais ont pour but un traitement multilingue.

Les ontologies linguistiques sont en développement constant, bien que l'heure ne soit plus vraiment à la conception ex-nihilo, mais plus à la réutilisation, la fusion d'ontologies pré-existantes.
Une étude des ontologies existantes et de leur types d'architecture est donc importante, afin de pouvoir identifier leurs similarités et leurs différences dans leur façon de traiter les aspects de la représentation des connaissances de base.
Une comparaison objective permet une meilleure utilisation des modèles existants, que ce soit pour l'utilisation d'une ontologie dans une application spécifique, pour la conception d'une ontologie ou sa réutilisation.
Il est important de savoir quelle ontologie correspond à une application, si l'ontologie est générale ou de domaine, comment elle peut être intégrée à une autre ontologie plus générale ou comment une ontologie plus spécifique pourrait y être reliée.
Dans cette optique, une ontologie de référence a été construite par [PINTO00], une ontologie de domaine sur les ontologies (les pages jaunes des ontologies). Elle a pour but d'aider les utilisateurs à trouver l'ontologie appropriée à l'application désirée.
La recherche dans le domaine de l'intégration de sources d'informations hétérogènes, ou plus simplement, le partage du savoir, a pour but de reconnaître et de combiner les connaissances pertinentes de façon à fournir une compréhension plus riche du domaine.

L'intégration est particulièrement intéressante lorsqu'elle permet de réconcilier les différences entre les diverses sources de connaissances tout en maintenant leur autonomie.

B- Bases de l'intégration

L'intégration est une opération nécessaire afin de concevoir une ontologie de large couverture pour le traitement de la langue naturelle à coût réduit. Que ce soit de l'intégration de données ou de l'intégration d'autres ontologies dans leur ensemble.
Le plus souvent, la fusion de données se fait grâce à une ontologie commune, ici nous verrons la fusion de deux ontologies.

L'intégration est grosso-modo une conception d'ontologie particulière. Historiquement, on distingue trois courants majeurs. Ces 3 grands courants d'idées qui en découlent représentent un continuum progressif et ne s'opposent pas [SMI00].

La première génération de projets, qui commence dans les années 60, visait la création de bases de données utilisables pour diverses applications.
Cette première génération représente une période où tout ce qui était nécessaire à l'environnement de l'application (base de données par ex.) devait être construit ex-nihilo.
L'intégration dans les systèmes de première génération est une conséquence logique de l'utilisation d'un système indépendant avec une seule base de données pour le plus d'objectifs possibles.

La seconde génération apparaît dans les années 80, grâce au développement des technologies du LAN (local area network). Les projets de seconde génération intégraient des informations et des données de sources et de systèmes variés grâce à l'architecture d'information des entreprises, afin de surmonter les barrières des projets de première génération. C'est dans cette génération que se développent les projets fédérateurs, normalisateurs.

Depuis la fin des années 80, la troisième génération paraît, poussée par la recherche informatique et le développement de projets.
Les projets de troisième génération relient entre elles des sources d'informations disparates mais accessibles afin de fournir l'apparence d'une intégration. Cette génération utilise des techniques allant de la structuration d'informations à l'extraction.
En d'autres mots, les données et les connaissances qui sont en dehors d'un système peuvent être reliées aux données et aux processus de travail internes; l'apparence de l'intégration est donnée, tandis que la richesse des contenus est préservée par la diversité.
Les années 90 ont vu s'installer la notion d'ontologie comme thème central, lors de l'apparition de problèmes d'intégration de connaissances devenues d'une granularité de plus en plus fine.

Nous verrons au cours d'une description théorique de l'intégration, l'importance d'avoir une méthode d'intégration. Nombreuses sont celles, qui proposées par les laboratoires de recherche, représentent des types d'intégration différents.

1 Description théorique de l'intégration

Si l'intégration est un type particulier de conception, il existe aussi plusieurs types d'intégration, selon le type de résultats escompté, les informations disponibles, le niveau d'intégration, etc.

1.1 La réutilisation d'ontologie

Selon [PINTO00], on peut étudier la réutilisation d'ontologie sous deux points de vue:
La construction d'ontologie, par l'assemblage, l'extension, la spécialisation et l'adaptation d'autres ontologies qui seront des parties de l'ontologie créée.
Ou la construction d'une ontologie par la fusion de diverses ontologies de même sujet ou de sujet proche, en une seule qui les unifie toutes.
Ces deux principes de réutilisation ont pour but de construire une ontologie sur la base d'autres ontologies. Du fait, la réutilisation est liée aux problématiques de construction d'ontologies.

1.1.1 Intégration

Ces deux types de réutilisation ont bien sûr des différences; la première est appelée "intégration d'ontologie". Elle se présente lors de la réutilisation d'une ontologie pré-existante afin d'en construire une nouvelle [PIN99].
L'avantage de l'intégration d'ontologie est que, pourvu qu'un ensemble de petites ontologies modulables et hautement réutilisables soit disponible, de larges ontologies peuvent être plus facilement assemblées. Ces petites ontologies doivent bien entendu être modifiées et adaptées avant d'être assemblées. L'intégration est particulièrement intéressante lorsqu'elle permet de réconcilier les différences entre les diverses sources de connaissances tout en maintenant leur autonomie.

1.1.2 Fusion

Quant à la fusion (merging), elle crée une ontologie unique et cohérente; les différentes ontologies à propos du même sujet sont fusionnées en une seule qui les « unifie » toutes (Pinto dans [VANZYL99]).La fusion serait finalement un type d'intégration particulier.

1.1.3 Utilisation

[NOY99] parle d'"alignment", de "traduction" d'ontologie, par la création de liens entre des ontologies dont les domaines sont le plus souvent complémentaires. [PIN99] décrit alors l'utilisation, qui se retrouve lorsqu'une ontologie ou plus sous-tendent et sont partagées entre différentes applications, ou une ou plusieurs ontologies sont utilisées pour spécifier ou implémenter un système à base de connaissances.

Parfois, plus d'une ontologie est utilisée pour l'intégration [TAM99]. En particulier un système peut être classé selon les catégories suivantes :

a°) Sans ontologie

Il y a une traduction directe d'une source à une autre.

b°) Une ontologie partagée
Une ontologie est utilisée pour intégrer les diverses sources et plusieurs fonctions de mappage sont définies entre la ressource et l'ontologie de partage.

c°) De multiples ontologies partagées
L'intégration est effectuée grâce à l'utilisation de multiples ontologies partagées et il y a des fonctions de mappage à la fois entre les ressources et l'ontologie partagée et entre les différentes ontologies.

1.2 Composants théoriques de l'intégration

1.2.1 Hétérogénéité

La problématique de l'intégration est fondée directement sur les remarques suivantes [HOVY97].
Premièrement, le problème de répétition (duplication). En effet de nombreuses ontologies de même ordre ou non ont été créées pour diverses applications. Cette prolifération fait souffrir les principes de réutilisation et de consistance. Bien que certains chevauchements soient nécessaires pour des raisons techniques, la majorité est inutile.
Ils peuvent être réduits par la mise en place d'un ensemble de concepts de base représentant les distinctions nécessaires et qui méritent d'être nommées. De même, pour un ensemble de termes (en accord avec les noms et des définitions de théorie neutre pour les propriétés d'héritage simple, concepts) et pour une taxinomie de base pour les héritages de propriétés simples.

Vient ensuite le problème de consistance. Chaque expert, selon son domaine de prédilection va référer un concept de manière spécifique, ou utiliser un terme pour se référer à différents concepts selon le domaine ou le sous-domaine.
Il est donc important de développer une terminologie consistante et de créer un modèle de décision consistant là où ont lieu les chevauchements.
Cette notion de correspondance, c'est à dire d'enregistrement de différents lexiques, est plus compliquée que le choix d'un simple lexique, et il est nécessaire d'éviter les raisonnements inconsistants et les conclusions en contradiction.

Troisièmement, le problème d'un modèle de construction efficace. La création d'un modèle de domaine est souvent compliquée par le nombre de décisions à prendre simultanément (les concepts et leurs relations définis par d'autres concepts et relations). La tâche de modélisation est simplifiée s'il est possible d'utiliser directement une ontologie de base large conçue pour couvrir les plus importants phénomènes de plusieurs domaines différents.

Une classification générale des différents types d'hétérogénéité selon leur niveau est mise en place par (Visser et al. 1998) dans [TAM99]. Elle peut être comprise dans une théorie plus générale de [HAKIM01] qui place ses définitions d'hétérogénéité à des niveaux plus élevés. Il définit une hétérogénéité des données et une hétérogénéité sémantique.

L'hétérogénéité des données se réfère aux différences au niveau des définitions locales, comme les types d'attributs, les formats… Ces différences sont aisément résolues.

Les deux types d'hétérogénéité, paradigmatique et du langage de [TAM99] peuvent y être compris.
L'hétérogénéité paradigmatique apparaît si des sources de connaissances différentes expriment un savoir au travers de paradigmes de modélisation différents. Par exemple une source peut formuler un savoir en utilisant des bases de données relationnelles alors qu'une autre utilisera une base de données orientée-objet.
Quant à l'hétérogénéité du langage elle est de mise si les sources de connaissances expriment un savoir par différents langages de représentation. Par exemple, qu'un système soit en LISP alors qu'un autre s'exprime au travers de clauses de Horn.

L'hétérogénéité sémantique se réfère aux différences dans la signification des données locales. Et ce, que des noms identiques soient mis sur des significations différentes, ou que des noms différents soient mis sur des significations identiques.
Cela revient aux deux types d'hétérogénéité de [TAM99] suivants:
L'hétérogénéité du contenu: si deux systèmes représentent des connaissances différentes. Par exemple si un système représente un savoir sur les étudiants de l'Université de Liverpool alors qu'un autre représente les connaissances de la fluctuation du marché de l'or.
L'hétérogénéité ontologique: cette hétérogénéité se présente lorsque différents systèmes utilisent différentes conceptualisations. Par exemple, un système conceptualise les animaux comme un ensemble de mammifères et un ensemble de reptiles, alors qu'un autre système les classera en carnivores et herbivores.

La difficulté dans la réconciliation d'ontologies dépend du type d'hétérogénéité. On trouve plusieurs types de décalage entre les ontologies, mais le décalage au sujet de la conceptualisation est le plus difficile à régler.

1.3. Méthodes d'intégration

1.3.1 Méthode de Pinto

[PINTO00]
L'ontologie résultant d'une intégration devrait répondre à une série de besoins, en plus de:
-l'évaluation habituelle: vérification et validation.
-les critères d'estimation (Gomez-Perez, Juristo, et Pazos 1995): complétude, concision, consistance, extensibilité et robustesse.
-les traits habituels (Gruber 1995): clarté, cohérence, extensibilité, parti pris de codage minimum, engagement ontologique minimal.
L'ontologie résultante devrait être, entre autres, non-ambiguë, établie sur des distinctions fondamentales appropriées, consistante et cohérente partout (bien que composée de connaissances de différentes ontologies intégrées), et devrait avoir un niveau suffisant et approprié de détail dans l'ontologie tout entière, c'est à dire qu'il ne faudra pas trouver d'îlots de niveau de détail exagéré et d'autres de niveau adéquat. La granularité se devra d'être homogène.
Les caractéristiques d'une bonne ontologie intégrée est donc liée aux caractéristiques de conception d'une bonne ontologie comme décrit par [HOVY97].

Une approche en cinq phases a été identifiée par le ANSI Ad Hoc group pour construire une norme, appelée la Reference Ontology:

Aux niveaux supérieurs (approx. 100000 termes): mettre en correspondance (pour aligner) les termes d'un nombre restreint de larges ontologies (taille, environ 100.000 éléments). Faire une telle intégration revient à créer un résultat dans lequel les utilisateurs peuvent choisir quels termes des composants de l'ontologie ils veulent voir et utiliser.

Modèles de domaine (moins de 2000 termes chaque): Les liens dans cette ontologie se font vers des ontologies de domaines spécifiques, développées pour le raisonnement au sujet du temps, de l'espace, de la physique, de la géographie, etc. Cela permet la liaison de divers modèles de temps, d'espace, etc.
Outils d'accès: Créer des outils faciles à utiliser pour l'accès et l'extension d'ontologie.
Diffusion: Placer la Reference Ontology ainsi créée sur le web, disponible librement.
Base théorique: Il est important d'avoir une équipe fortement qualifiée qui trouveraient des généralisations puissantes dans l'ontologie, afin de supprimer les éléments inutiles et contradictoires, et de créer une factorisation maximale des niveaux supérieurs de l'ontologie
Outils de "Merging"/d'"alignment": En parallèle, il est important d'étendre les techniques croisées d'alignement semi-automatisées existantes et d'en inventer de nouvelles. Les rendre disponibles sur le Web pour la fusion et la mise en relation à des modèles supplémentaires de domaine.

1.3.2. Méthode de Hakimpour

D'une façon générale, l'intégration d'ontologie est fondée sur la recherche de similitudes et de différences entre deux définitions intensionnelles.
L'approche présentée dans [HAKIM01] est basée sur la fusion d'ontologies sur des relations similaires parmi des concepts de différentes ontologies. Des définitions formelles des ressemblances basées sur des définitions intensionnelles sont présentées desquelles sont tirées les conséquences extensionnelles. Le procédé de fusion d'ontologies basé sur les relations de similitude détectées est discuté. Dans la théorie de [HAKIM01], l'ontologie fusionnée est finalement employée pour dériver un schéma intégré au sens de base de données.

Afin de découvrir si (et comment) les éléments de différents schémas sont apparentés, [HAKIM01] emploie des relations de similitude. La détection de la similitude est basée sur des définitions intensionnelles des termes représentées dans un langage logique (logique de description). Un système de raisonnement (tel que PowerLoom) peut fusionner des ontologies utilisées par des communautés spécifiques. Des superviseurs humains et une méthode semi-automatique coopèrent pour trouver ces similitudes [HAKIM01].

Dans [HAKIM01], la détection des relations de similitude est basée sur les définitions intensionnelles.
Quatre niveaux de similitudes entre deux définitions intensionnelles cohérentes peuvent être identifiés:

Définitions disjointes*:* Ce niveau a le degré le plus bas de similitude. Deux concepts ou relations sont disjoints si la conjonction de leurs définitions intentionnelles implique *faux.* Il s'ensuit que les extensions des concepts (ou des relations) sont disjoints (ex: rue étroite et voie élevée, camion et employé, ou sœur et père).

Définitions superposables: Concepts, définitions qui ont une relation. Si on ne peut pas prouver que la conjonction de deux définitions intentionnelles est *fausse* (un système de

raisonnement peut nécessairement ne pas la considérer *vraie),* alors elles se superposent. Il est important que les extensions des définitions se croisent (ex: employé et stagiaire, ou collègue et sœur).

Définitions spécialisées (sous-concept ou sous-relation*):* Si la définition intentionnelle de Cj est l'implication de la définition intensionnelle de Ci, alors Ci est une spécialisation de Cj.

Définitions égales*:* Ce niveau a le degré le plus élevé de similitude. Si la définition intensionnelle des deux définitions intentionnelles sont équivalentes, alors les concepts définis sont égaux. Selon la définition ci-dessus, si deux concepts ou relations sont égaux, chacun d'eux spécialise l'autre, respectivement. En outre, les extensions correspondantes sont égales. Par exemple, "véhicule"et "moyen de transport" sont égaux s'ils ont la même définition intensionnelle.

La dérivation des similitudes entre les ontologies exige des références communes aux deux ontologies et un système de raisonnement pour apparier. Les références communes peuvent être fournies par une ontologie de niveau plus élevé tel que la bibliothèque des ontologies sur le site d'Ontolingua, ou par des thesaurus tels que WordNet [HAKIM01]. La recherche de similitudes peut également être faite par des experts proches des deux communautés ou par une méthode hybride semi-automatique. Les relations de similitude sont employées pour fusionner deux ontologies. Toutes les définitions intensionnelles sont prises dans les communautés respectives pour le processus de fusion et pour déterminer explicitement des relations de similitude comme expliqué ci-après dans [HAKIM01]. Les définitions disjointes ne sont pas discutées ici.

1. Si deux définitions sont égales, le résultat de la fusion est une seule définition intensionnelle mentionnée par les deux termes initiaux. C'est-à-dire, les différents termes dans les définitions locales du schéma peuvent se rapporter au même concept - termes synonymes tels que "la personne" et "le résident" selon le contexte.
2. Si une définition intensionnelle de Ci spécialise Cj alors la similitude du sous-concept ou de la sous-relation sera explicitement établie entre eux (par exemple, "étudiant "et "personne" pourraient avoir cette similitude). Si Ci spécialise un sous-concept ou une sous-relation de Cj alors un tel rapport ne sera pas explicitement enregistré, puisqu'il ne peut pas être déduit de la transitivité de la spécialisation (par exemple, "étudiant diplomé" et "personne").
3. Si une définition de Ci se superpose à Cj alors un nouveau concept ou une relation supplémentaire sera déclaré en tant que conjonction des deux définitions intentionnelles. Bien que la conjonction des deux définitions intensionnelles ne puisse pas être prouvée fausse, il n'est pas possible d'avoir aucune instance d'un tel concept. Cependant, si des instances de tels concepts se superposant existent, la pertinence du nouveau concept superposant nécessite une étude par un expert

2. Pré-requis à l'intégration

2.1. Techniques théoriques et pratiques de pinto:

Pour simplifier le procédé général de l'intégration [PINTO00] fait remarquer qu'il est important d'avoir accès aux représentations au niveau de la connaissance de toutes les ontologies réutilisées, puisque la plupart du travail est effectué au niveau de la connaissance.

Si la représentation du niveau de la connaissance d'une ontologie n'est pas disponible, alors un processus de re-développement de l'ontologie (Blazquez et autres 1998) peut être appliqué.

De même, il est plus simple d'intégrer deux ontologies implémentées dans le même langage, pour qu'il n'y ait pas besoin de traduction d'une ontologie à l'autre.
Pour toutes ces raisons on peut avoir des activités d'intégration pour la même ontologie à différentes étapes du processus de construction d'ontologie.
Une autre conclusion importante est que l'intégration devrait commencer dès que possible dans le cycle de vie de construction d'ontologie de sorte que le procédé général de construction soit simplifié.

Si on commence l'intégration dès la conceptualisation, alors on a besoin d'ontologies représentées au niveau de la connaissance, pas au niveau de l'implémentation. Par conséquent, un procédé de re-développement de l'ontologie (Blazquez et autres 1998) peut devoir être appliqué. Ce procédé se compose habituellement de trois phases: une phase de décompilation, une phase de restructuration et une phase de mise à jour.

Les activités initiales pour exécuter l'intégration incluent:
- la recherche et le choix d'ontologies à réutiliser;
- l'évaluation des ontologies candidates par des experts du domaine sur des critères spécialisés orientés vers l'intégration;
- l'estimation des ontologies candidates par des ontologistes sur des critères spécialisés orientés vers l'intégration;
- le choix de l'ontologie la plus adéquate à réutiliser parmi les ontologies candidates analysées.

Ces activités précèdent l'intégration de la connaissance de l'ontologie intégrée dans l'ontologie résultante.
La meilleure ontologie candidate est celle qui peut le mieux (le plus étroitement) ou le plus facilement (en utilisant le moins d'opérations) être adaptée pour devenir l'ontologie nécessaire. Quand ce processus se termine, c'est à dire que l'ontologie appropriée à la réutilisation pour ce processus particulier d'intégration est trouvée, il faut intégrer la connaissance de cette ontologie.
Pour cela, on a besoin d'opérations d'intégration. Celles-ci peuvent être vues en tant qu'opérations de composition, combinaison, de modification ou d'assemblage.
Les opérations d'intégration indiquent comment la connaissance d'une ontologie intégrée va être incluse et combinée avec la connaissance dans l'ontologie résultante, ou être modifiée avant son inclusion.
La connaissance des ontologies intégrées peut être, entre autres:
- utilisée telle qu'elle est;
- adaptée (ou modifiée);
- spécialisée (menant à une ontologie plus spécifique de même domaine)
ou
- augmentée (par une connaissance plus générale ou par une connaissance de même niveau).

Après l'intégration de la connaissance, l'application des opérations d'intégration adéquates, on doit évaluer, estimer et analyser l'ontologie résultante.

Il faut prêter attention à un ensemble de critères spécialisés qui analysent spécifiquement si l'ontologie résultante est de qualité.
Certains de ces critères sont également important dans n'importe quelle méthodologie de construction d'ontologie, tels que: l'identification des modules dans lesquels on peut diviser l'ontologie, l'identification des assomptions et des engagements ontologiques auxquels ces modules devraient se conformer, etc.

Une autre activité du processus d'intégration est l'évaluation des ontologies candidates d'un point de vue d'intégration.

Quelle connaissance est manquante?
(par la connaissance il est entendu n'importe quel fragment de connaissance, comme, les concepts, les critères de classification, les rapports, etc.). Parfois quelques fragments de connaissance qui sont appropriés, importants et utilisés couramment pour caractériser le domaine ne sont pas représentés dans l'ontologie.
Ceci inclut non seulement les classes et les instances, mais également des distinctions importantes interprétées selon les concepts de domaine, de différents critères de classification qui sont largement acceptés pour caractériser le domaine, des relations qui sont appropriées pour représenter la connaissance au sujet du domaine (quelles relations devraient être spécifiées et pour quels concepts doivent-elles l'être), etc.
Les experts du domaine devraient également analyser quelle connaissance importante au sujet du domaine manque dans l'ontologie en vue de l'utilisation particulière que l'ontologie va avoir.

Quelle connaissance devrait être retirée?
Parfois quelques fragments de connaissance représentés dans l'ontologie sont superflus, parce qu'ils ne sont pas importants, ou non appropriés, ou rarement utilisés pour décrire le domaine en question ou parce qu'ils ne sont pas nécessaires pour l'usage particulier que l'ontologie va avoir.

Quelle connaissance devrait être replacée?
Parfois des fragments de connaissance devraient être placés ailleurs dans l'ontologie de sorte que le domaine soit mieux caractérisé.

Quelles sources de la connaissance devraient être changées?
Parfois certaines des sources de connaissance employées pour acquérir de la connaissance ne sont pas les plus sûres ou les plus à jour. La connaissance de ces sources qui est représentée dans l'ontologie devrait être remplacée par une connaissance plus sûre, normalisée et plus à jour.

Quelle documentation devrait être changée?
Parfois la documentation du domaine n'est pas correcte (syntaxiquement et sémantiquement), précise, complète ou ne réfléchit pas les dernières découvertes dans le domaine et devrait être changée. La documentation devrait expliquer le domaine et les fragments de connaissance représentés dans l'ontologie de sorte qu'un non-expert puisse apprendre assez au sujet du domaine pour pouvoir comprendre les concepts qui sont représentés dans l'ontologie.

Quelle terminologie devrait être changée?
Parfois la terminologie utilisée n'est pas la plus couramment acceptée dans le (sous)domaine ou dans un domaine relatif, ou n'est pas la terminologie standard et devrait être changée.

Quelles définitions devraient être changées?
Parfois les définitions utilisées ne sont pas les plus couramment acceptées, standards ou composées des caractéristiques définitoires des fragments de connaissance.

Quelles pratiques devraient être changées?
Parfois les procédures employées pour recueillir la connaissance (acquisition de connaissances) et construire l'ontologie (conception d'ontologie) ne sont pas les plus correctes ou ne suivent pas les pratiques les plus acceptées pour un domaine.

L'Estimation orientée intégration [PINTO00]
Une autre activité du procédé d'intégration est l'estimation des ontologies candidates d'un point de vue d'intégration. Les points suivants sont à prendre en compte dans l'analyse d'une ontologie pour intégration:

Structure générale de l'ontologie
Pour analyser la structure de l'ontologie candidate, six critères doivent être pris en compte:
- la structure est-elle adéquate (une hiérarchie, plusieurs hiérarchies, un graph, etc.) et de préférence bien-équilibrée?
- l'ontologie est-elle divisée en modules adéquats et suffisants, c'est-à-dire, l'ontologie est-elle divisée en sous-ontologies naturelles et appropriées (qualité et quantité)?
- la spécialisation des concepts est-elle adéquate et suffisante, c'est-à-dire, les concepts nécessaires et leurs spécialisations sont-ils représentés?
- l'héritage est-il utilisé correctement et de façon appropriée?
- existe-t-il assez de diversité représentée dans l'ontologie de sorte que de nouveaux concepts soient plus facilement introduits?
- les concepts semblables sont-ils représentés d'une manière proche l'un de l'autre tandis que des concepts moins semblables sont représentés plus à l'écart (minimisation de la distance sémantique entre concepts enfants de mêmes parents)?

Les distinctions fondamentales
Les distinctions fondamentales (critères de classification faits des concepts décrits dans l'ontologie) appropriées et exigées (quantité et qualité) sont-elles représentées? Changer les distinctions fondamentales (habituellement représentées au sommet de l'ontologie) sur lesquelles l'ontologie est basée peut également impliquer une vaste mise à jour de l'ontologie. La seule connaissance qui peut être réutilisée sont les exemples représentant les éléments. Habituellement, dans ce cas-ci, il est préférable de construire une autre ontologie.

La relation structurante
La relation structurante privilégiée lors de la construction de l'ontologie est-elle celle requise? Changer la relation privilégiée selon laquelle l'ontologie est organisée peut également avoir des conséquences profondes. La connaissance entière devrait, probablement, être révisée, puisque la nouvelle relation organise la connaissance dans l'ontologie de façon tout a fait différente. La connaissance au sujet d'un domaine donné qui devrait être représentée en utilisant une certaine relation n'a rien à voir avec celle qui devrait être représentée en utilisant une autre relation. Il est probablement préférable d'établir une nouvelle ontologie à partir de zéro (si aucune ontologie disponible ne répond aux besoins).

Nommage des règles de convention

Les noms des fragments de la connaissance suivent-ils des règles standardisées? Par exemple, dans l'ontologie de référence autant que possible les relations binaires ont été nommées en concaténant le nom de l'ontologie (ou le nom du concept représentant le premier élément de la relation), le nom de la relation et le nom du concept cible.

Les définitions

Les définitions des fragments de connaissance suivent-elles des patrons unifiés, sont-elles claires, concises, conformes, complètes, correctes (lexicalement et syntaxiquement), précises et fines? Sont-elles efficaces? Toutes ces questions traitent de la façon dont la connaissance est représentée dans l'ontologie.

La documentation

La documentation est-elle claire, utile et adéquate? Traite-t-elle des alternatives aux représentations et des choix qui ont été faits pour représenter la connaissance? Est-elle logique par rapport à la définition des fragments de connaissance? Bien que la documentation soit un des constituants d'un fragment de connaissance d'ontologie, c'est habituellement un composant des plus négligé.

Les fragments de connaissance représentés

Est-ce que seuls et d'une manière exhaustive, les fragments appropriés de connaissance sont représentés (ou inclus)? Cette question devrait être analysée prenant en compte des fragments de connaissance que les experts dans ce domaine ont trouvés importants à représenter. S'ils ont trouvé des fragments de connaissance manquants ou superflus, ils doivent être ajoutés ou effacés de l'ontologie.
L'analyse de l'ontologiste doit focaliser sur les deux aspects: les fragments appropriés et nécessaires de connaissance sont représentées et ils sont utilement représentés. Ils sont orientés vers l'intégration de la connaissance qui sera exécutée aux étapes ultérieures du processus.

Choix de l'ontologie adéquate

Ce choix dépend également dans une certaine mesure des autres ontologies qui vont être réutilisées puisque dans un processus d'intégration on peut réutiliser plus d'une ontologie. Il est important que les ontologies réutilisées soient compatibles entre elles, à savoir en ce qui concerne leurs hypothèses, par exemple, leurs engagements ontologiques (théories de Gruber).

2.2. Décomposition de l'intégration

Opérations d'intégration de Pinto

Intégrer de la connaissance nécessite des opérations d'intégration. Pinto a identifié et défini une taxinomie des opérations d'intégration au niveau de la connaissance. Les opérations d'intégration d'ontologies sont classifiées en opérations fondamentales et non-fondamentales (complexes).
Les opérations fondamentales d'intégration sont divisées en:

- opérations sur l'ensemble des ontologies comme, l'inclusion d'une ontologie.

- opérations sur les constituants d'une ontologie subdivisées en opérations qui:

- éliminent ou introduisent un fragment de connaissance de sorte que l'ontologie contienne la connaissance appropriée et nécessaire pour décrire le domaine,
- changent le nom d'un fragment de connaissance pour se conformer aux règles conventionnelles de nommage, ou introduisent une terminologie habituelle ou standard.
- changent la documentation d'un fragment de connaissance pour la mettre à jour, la corriger ou en augmenter sa clarté,
- changent la définition d'un fragment de connaissance pour représenter la connaissance d'un domaine donné plus exactement, simplement, clairement, correctement, précisément, complètement, etc.

Il existe donc des opérations qui introduisent ou suppriment une classe, un exemple, une relation, etc., son nom, sa documentation et sa définition. Toutes ces opérations doivent être utilisées avec parcimonie, sinon, il est plus rentable de construire une ontologie à partir de zéro.
Un soin spécial doit être apporté aux opérations qui relient des fragments de connaissance à d'autres.
Par exemple, l'inclusion d'une classe a besoin en général de spécifications sur l'endroit où le concept devrait être introduit (par exemple, en énonçant les parents de la classe) et la suppression d'une classe peut impliquer ou non la suppression de la hiérarchie entière sous elle. Les changements de la définition incluent, une série d'autres changements. Par exemple, dans le cas des concepts (des classes ou des instances), elle inclut des changements dans les relations dans lesquelles le concept est impliqué (y compris les relations privilégiées structurant l'ontologie), des changements dans les propriétés définissant le concept, etc. Ces opérations fondamentales d'intégration servent de base aux opérations d'intégration non-fondamentales telles que les mappings d'ontologie, le ré-adressage, la "sous-ontologisation", l'effondrement, etc.

3. Les différentes techniques d'intégration

Le processus de fusion d'ontologies est fondé sur deux ontologies d'entrée (ou plus) et retourne une seule ontologie basée sur les ontologies sources. La fusion manuelle d'ontologies à l'aide d'outils conventionnels d'édition sans support est longue, difficile, et sujette à l'erreur. Par conséquent, on a récemment proposé plusieurs systèmes et cadres de travail et de support pour la fusion. (Hovy, Chalupsky, Noy et Musen, MGUINNESS et autres). Les approches se fondent sur l'heuristique appariement syntaxique et sémantique qui sont dérivées des habitudes des ontologistes confrontés avec la tâche de fusionner des ontologies.
Trois grandes techniques d'intégration existent principalement, et servent dans la conception d'outil pour des fusions automatiques ou semi- automatique. Elles sont fondées sur la façon de traiter les opérations de fusion.

3.1. technique "bottom up"

Une première approche pour la fusion d'ontologies est décrite par Hovy [STUM01]. Là, plusieurs heuristiques sont décrites pour identifier les concepts se correspondant dans diverses ontologies, par exemple en comparant les noms et les définitions en langage naturel de deux concepts, et en contrôlant la proximité de deux concepts dans la hiérarchie de concepts.
L'approche de type "bottom-up" est une approche dans laquelle on remonte des concepts spécifiques vers les catégories de haut niveau.

3.2. technique "middle-out"

Une méthodologie de capture des concepts de l'ontologie de type "middle-out" est une méthodologie où les concepts les plus importants sont définis en premier, et sont ensuite affinés ou spécialisés si nécessaire. Cette approche permet de garder le contrôle du niveau de détail (les détails ne sont explicités que si c'est nécessaire). C'est un avantage que l'on ne retrouve pas dans des approches de type bottom-up.
De même, l'approche middle-out augmente la stabilité; les concepts génériques sont définis en fonction des concepts majeurs, supposés les plus stables. On ne retrouve pas cette stabilité dans des approches de type "top-down" [COSI].

3.3. technique "top down"

La troisième approche est de type top-down, c'est à dire où les concepts les plus génériques sont définis au début et affinés par la suite.

De ces points théoriques et ces différentes techniques, plusieurs choix sont à faire, aussi bien sur le type d'intégration à effectuer que sur les opérations à appliquer pour traiter une intégration.
Par exemple, par rapport au travail demandé, certaines opérations sont inclues beaucoup trop profondément dans le système et il n'y aurait que des moyens longs de les mettre en place (comme les re-défintions de concepts), ce qui les rend impossible pour le temps d'un stage.
Les deux premières parties de ce rapport ont permis de poser des jalons théoriques, de mettre en place l'intégration des ontologies d'Artimis et d'EWN.
La partie suivante exposera les choix de type d'intégration, de niveau d'intégration, de types d'opération
sont expliqués ci- après ainsi que les résultats de cette intégration.

(…)

1.2 Description de l'ontologie d'EWN

EuroWordNet est une base de connaissances lexicales multilingue (cf. A-3.6 p.34) Le vocabulaire couvert par Eurowordnet appartient au domaine général.
Dans ces dictionnaires monolingues, ces différents WordNets, on ne trouve pas tant des définitions que des liens de sens entre les concepts (essentiellement les relations «est un type de» et «est une partie de»). Tous ces dictionnaires étant reliés par le biais d'une ontologie, l'Index-Inter-Lingue (ILI), c'est ce dernier qui sera pris en compte pour la mise en correspondance (mappage), l'intégration.

EWN va servir à l'intégration de données nouvelles à l'ancien système afin d'affiner la description du monde au niveau ontologique et d'accéder à un ensemble de données au niveau lexical, grâce à une ontologie étendue (exhaustive). Les données à intégrer sont constituées par les 91600 (22700 pour le français) synsets disponibles dans la base EWN par l'intermédiaire de l'ILI.

Bibliographie

[BAT90] Bateman, J. A. (1990) , *Upper modeling: organizing knowledge for natural language processing*, *in* `5th. International Workshop on Natural Language Generation, 3-6 June 1990'.

[BENJAM99] Benjamin and Perez -*Knowledge system technology: ontology and problem solving methods,* 1999.

[BORG96] A. Borgida. "On the relative expressiveness of description logics and predicate logics". *Artificial Intelligence*, volume 82, number 1-2, pages 353–367, 1996.

[BOUIL98] P. Bouillon, Fr. Vandooren, L. Da Sylva, L. Jacqmin, S. Lehmann, G. Russell and E. Viegas, *Traitement automatique des langues naturelles*, Duculot, 1998.

[CLIPS] http://clips.imag.fr/projets/unl/

[CORBY02] http://www.essi.fr/~riveill/cours/internet/02-ontologie.pdf

[COSI] Bertrand Sereno, Annie Corbel et Jean-Jacques Girardot http://www.emse.fr/~sereno/rjcia_draft.html

[CYC] http://www.cyc.com/products.html

[DAHL88] Dahlgren, Kathleen, *Naive Semantics for Natural Language Understanding*. Boston: Kluwer Academic Press, 1988

[EUROW] http://www.hum.uva.nl/~ewn/index.html#1

[FER99] *Overview of methodologies for building ontologies* Fernandez and Lopez, 1999.

[GARD01] www.loria.fr/~gardent/teaching/semLex/pustejovski4.pdf

[GARD02] http://www.loria.fr/~gardent/teaching/semLex/wdnet4.pdf

[GOM99] Gomez Perez, *Développements récents en matière de conception de maintenance et d'utilisation des ontologies*.. Terminologies Nouvelles, 1999.

[GRUNIG98] Mark S. Fox, Michael Gruninger, *Enterprise Modeling,* in A.A.A.I., fall 1998.

[GUAR95] N. Guarino and P. Giaretta. "Ontologies and knowledge bases. towards a terminological clarification". In N. Mars, editor, *Towards very large knowledge bases:knowledge building and knowledge sharing*, pages 25–32, IOS Press, Amsterdam, 1995.

[GUAR97] N.Guarino et P.Giaretta, *Ontologies and Knowledge Bases: Towards a terminological Clarification. In towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, 1997.(Ed.), IOS press, Amsterdam, pp. 25-32.

[GUAR98] N. Guarino. "Formal ontologies and information systems". In N. Guarino, editor, *Proceedings of FOIS'98*, IOS Press, Amsterdam, 1998.

[GUM] http://www.darmstadt.gmd.de/publish/komet/gen-um/node1.html

[GUMC] viewer DMI/GRI

[GUMW] http://www.darmstadt.gmd.de/publish/komet/gen-um/node11.html#SECTION00352000000000000000

[HABE01] http://www.limsi.fr/Individu/habert/00-01/WordNetEtRessourcesSemantiquesArticle/index.html

[HAKIM01] Farshad Hakimpour, Andreas Geppert, *Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach*, www.ifi.unizh.ch/dbtg/Projects/MIGI/ publication/FOIS2001-final.pdf

[HEIJ97] G. van Heijst, A. Th. Schreiber and B. J. Wielinga., *Using explicit ontologies for KBS development. International Journal of Human-Computer Studies* 1997.

[HEIN02] johannes.heinecke@rd.francetelecom.com

[HOVY97] *A Standard for Large Ontologies*, http://www.isi.edu/nsf/papers/hovy2.htm, Eduard Hovy, 1997.

[JASP99] Jasper and Uschold, *Framework for understanding and classifying ontology application* 1999.

[LOPEZ99] Fernandez and Lopez, *Overview of methodologies for building ontologies* 1999
http://kmi.open.ac.uk/people/bertrand/phd/notices/lopez99overview.html

[MCGUI01] O. Lassila and D. McGuinness, "The role of frame-based representation on the semantic web". *Electronic Transactions on Artificial Intelligence (ETAI) Journal: area The Semantic Web*, volume To appear, 2001.

[MILL90] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. *Introduction to WordNet: an on-line lexical database*. In: *International Journal of Lexicography* 3 (4), 1990, pp. 235 - 244. revised in 1993.

[MIZ95] Mizoguchi Riichiro and Vanwelkenhuysen Johan and Ikeda Mitsuru, *Task Ontology for Reuse of Problem Solving Knowledge* in Towards Very Large Knowledge Bases, N. Mars (editor), pages 46-59, IOS Press, Amsterdam, 1995.

[MMTHS01] http://www-clips.imag.fr/geta/mathieu.mangeot/MM-These/partieA.html

[MUEL98] Mueller, Erik T. (1998). *Natural language processing with ThoughtTreasure*. New York: Signiform. Available: http://www.signiform.com/tt/book/
http://www.signiform.com/tt/htm/faq.htm

[NKOS] http://nkos.slis.kent.edu/KOS_taxonomy.htm

[NOY97] Noy, N.F. and Hafner, C. *The State of the Art in Ontology Design: A Survey and Comparative Review In AI Magazine*, 18(3), 53-74 (1997) http://www.aaai.org/Library/Magazine/Vol18/18-03/vol18-03.html

[NOY99] Natalya, Fridman, Noy, Mark, A., Musen, *Smart : automated support for ontology merging and alignment*, 1999 http://kmi.open.ac.uk/people/bertrand/phd/notices/noy99smart.html:

[ONTOS02] http://crl.nmsu.edu/Staff.pages/Technical/sergei/book/intro.pdf

[OWL02] http://www.w3.org/TR/2002/WD-webont-req-20020307/#section-introduction

[PAN96] Panaget F., *D'un système générique de génération d'énoncés en contexte de dialogue oral à la formalisation logique des capacités linguistiques d'un agent rationnel dialoguant,* 1996

[PATIL] www.arches.uga.edu/~abhijitp/SemWeb/LargeOntology.ppt

[PEREZ01] Mariano Fernández-López; Asunción Gómez-Pérez, Nicola Guarino, *The Methontology & OntoClean merge*, 2001

[PIN99] Pinto, H.S., A. Gómez-Pérez, and J.P. Martins. *Some Issues on Ontology Integration*. In *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods:*Lessons Learned and Future Trends, 1999. Stockholm, Sweden.

[PINTO00] H. Sofia Pinto, J.P. Martins, *Reusing Ontologies*, aifbhermes.aifb.uni-karlsruhe.de/AAAI2000/ CameraReady/HPinto00.pdf

[PROT00] http://protege.stanford.edu/index.html

[SCHUTZ] http://www.unicom.co.uk/3in/ISSUE2/4.ASP#1

[SENSARB] http://www.isi.edu/natural-language/po/c-664.html

[SMI00] William W. Stead, Randolph A. Miller, Mark A. Musen, William R. Hersh, *Integration and Beyond: Linking Information from Disparate Sources and into Workflow,* http://www-smi.stanford.edu/pubs/SMI_Reports/SMI-2000-0834.pdf

[SOWA00] Brooks/Cole, *Knowledge Representation* chp.2 2000 http://www.bestweb.net/~sowa/ontology/toplevel.htm

[SOWA84] J. Sowa. *Conceptual structures: information processing in mind and machine.* Addison-Wesley, 1984.

[SOWA95] J.F. Sowa, *Top-Level Ontological categories. In international Journal of Human-Computer Studies*,43,5-6,pp.669-685, 1995.

[STUM01] Stumme, Maedche FCA-MERGE : Bottom-Up Merging of Ontologies Gerd Stumme, www.aifb.uni-karlsruhe.de/WBS/ gst/papers/2001/IJCAI01.pdf

[SUMO] http://ontology.teknowledge.com/rsigma/FormalSUOdraft.rtf

[SUMO01] http://ontology.teknowledge.com/cgi-bin/cvsweb.cgi/SUO/

[SUMOARB] http://ontology.teknowledge.com/rsigma/arch.html

[TAM01] Valentina A.M. Tamma, *An Ontology Model supporting Multiple Ontologies for Knowledge sharing*, thèse octobre 2001.

[TAM99] Tamma, V.A.M. and P.R.S. Visser, *Integration of Heterogeneous Sources: Towards a Framework for comparing Techniques*, in *Proceedings of the IJCAI-99 Workshop on Intelligent Information Integration, Stockholm,* 199 http://www.csc.liv.ac.uk/~kraft/publications/AIIA98.ps.gz

[USCH96] Uschold M. et Grüninger M. *ONTOLOGIES: Principles, Methods and 16 applications, Knowledge engineering review*, vol.11, N.2, 1996.

[VAILL] http://www.lalic.paris4.sorbonne.fr/stic/data/Vaillant(etal).htm

[VANZYL99] J.Van Zyl, D.Corbett, *Framework for comparing methods for using or reusing multiple ontologies in an application* 1999.

[VOSS97] Piek Vossen, Wim Peters, et al., *The Multilingual design of the EuroWordNet Database,* Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, 1997 http://citeseer.nj.nec.com/cache/papers/cs/343/http:zSzzSzwww.let.uva.nlzSz~ewnzSzdocszSzP013.pdf/vossen97multilingual.pdf

[WIEDERHOLD94] Wiederhold, Gio, *Interoperation, Mediation, and Ontologies* Proceedings *International Symposium on Fifth Generation Computer Systems (FGCS94)*, Workshop on Heterogeneous Cooperative Knowledge-Bases, Vol.W3, pages 33-48, ICOT, Tokyo, Japan, Dec. 1994; to be published in a Springer Verlag volume.