



TOTh 09

Terminologie & Ontologie : Théories et Applications

Actes de la troisième conférence TOTh - Annecy - 4 & 5 juin 2009



Institut Porphyre
Savoir et Connaissance

Dans un monde où la communication et le partage d'information sont au cœur de nos activités, les besoins en terminologie se font de plus en plus pressants. Il est devenu impératif d'identifier les termes employés et de les définir de façon consensuelle et cohérente tout en préservant la diversité langagière.

La terminologie, en tant que discipline scientifique, se fonde sur une conceptualisation d'un domaine et sur les mots pour en parler. Elle se doit donc de concilier un point de vue linguistique et un point de vue ontologique. Elle doit également, dans une société numérique où les connaissances constituent la principale richesse, pouvoir être opérationnalisée à des fins de traitement de l'information.

Les conférences TOTh se situent dans le prolongement des colloques annuels de la Société française de terminologie organisés en décembre à Paris (Ecole normale supérieure de la rue d'Ulm). Planifiées à mi-parcours, au mois de juin à Annecy (Polytech'Savoie), elles en complètent l'offre et proposent des conférences avec appel à communications, comité de lecture et publication des actes.

Les conférences TOTh ont pour objectif de rassembler industriels, chercheurs, utilisateurs et formateurs dont les préoccupations relèvent à la fois de la terminologie et de l'ontologie et, de façon plus générale, de la langue et de l'ingénierie des connaissances. Elles se veulent un lieu d'échange et de partage où sont exposés problèmes, solutions et retours d'expériences tant sur le plan théorique qu'applicatif ; ainsi que les nouvelles tendances et perspectives des disciplines associées : terminologie, langues de spécialité, linguistique, intelligence artificielle, systèmes d'information, ingénierie collaborative, etc.

Christophe Roche, Président du Comité Scientifique

<http://www.porphyre.org>



Institut Porphyre
Savoir et Connaissance

ISBN 978-2-9536168-0-4
EAN 9782953616804

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1^{er} juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy - 5 et 6 juin 2008

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2009. *Actes de la troisième conférence TOTh - Annecy - 4 & 5 juin 2009*

Editeur : Institut Porphyre, *Savoir et Connaissance*

<http://www.porphyre.org>

Annecy, 2009

ISBN 978-2-9536168-0-4

EAN 9782953616804

© Institut Porphyre, *Savoir et Connaissance*



Actes de la conférence

TOTh 2009

Annecy – 4 & 5 juin 2009

avec le soutien de :

- Société française de terminologie
- Association Européenne de Terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France



Institut Porphyre
Savoir et Connaissance

<http://www.porphyre.org>

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

| | |
|------------------|---|
| Loïc Depecker | Professeur, Université de Sorbonne nouvelle |
| André Manificat | Directeur, GRETh |
| Christophe Roche | Professeur, Université de Savoie |
| Philippe Thoiron | Professeur émérite, Université de Lyon II |

Comité de programme

| | |
|----------------------|---|
| Bruno de Bessé | Professeur, Université de Genève |
| Pierre Blanc | EDF SEPTEN |
| Danièle Bourcier | CNRS, CERSA Paris |
| Marc van Campenhoudt | Professeur, Termisti, ISTI, Bruxelles |
| Danielle Candé | CNRS, Université Paris Diderot |
| Stéphane Chaudiron | Professeur, Université de Lille III |
| Viviane Cohen | France Télécom, Paris |
| Rute Costa | Professeur, Université Nouvelle de Lisbonne |
| Luc Damas | MCF, Université de Savoie |
| Sylvie Desprès | MCF, Université Paris XIII |
| François Gaudin | Professeur, Université de Rouen |
| Anne-Marie Gendron | Chancellerie fédérale suisse, Section de terminologie |
| Jean-Yves Gresser | ancien Directeur à la Banque de France |
| Olivier Haemmerlé | Professeur, Université de Toulouse |
| Jean-Paul Haton | Professeur, Université de Nancy 1 |
| Michèle Hudon | Professeur, Université de Montréal |
| John Humbley | Professeur, Université Paris 7 |
| Michel Ida | Directeur MINATEC, CEA |
| Hendrik Kockaert | Professeur, Lessius Hogeschool (Anvers) |
| Michel Léonard | Professeur, Université de Genève |
| Pierre Lerat | Professeur honoraire, Université Paris XIII |
| Widad Mustafa | Professeur, Université de Lille III |
| Henrik Nilsson | Terminologikum TNC, Suède |
| Jean Quirion | Professeur, Université du Québec en Outaouais |
| Renato Reinau | Suva, Lucerne |
| François Rousselot | MCF, Université de Strasbourg |
| Gérard Sabah | CNRS, Orsay |
| Michel Simonet | CNRS Grenoble |
| Marcus Spies | Professeur, Université de Munich |
| Dardo de Vecchi | Professeur associé, Euromed-Management |

Comité d'organisation :

Responsable : Luc Damas
Samia Chouder, Joëlle Pellet

Avant propos



Dès la troisième édition, les conférences TOTh ont trouvé une structuration qui traduit bien à la fois le caractère scientifique et pluridisciplinaire de la terminologie et l'intérêt de notre communauté pour d'autres domaines partageant des préoccupations communes.

Ainsi, la conférence d'ouverture a été donnée par une personnalité invitée issue d'une discipline différente de la nôtre – ici la phylogénèse – mais pour laquelle le langage et la pensée jouent également un rôle primordial.

Les contributions se sont réparties naturellement, et par le jeu des évaluations de façon équitable, en trois groupes ayant donné lieu à trois sessions.

Le premier groupe a rassemblé les articles portant principalement sur la dimension linguistique de la terminologie. Ont été abordés l'extraction terminologique à partir de dictionnaire, la place accordée aux corpus dans la construction de terminologies, l'acquisition de connaissances à partir de textes et l'apport des ressources linguistiques issues du web.

La deuxième session s'est donc logiquement intéressée à la dimension conceptuelle de la terminologie. Les notions de concept, de relation, d'ontologie ont été au cœur des présentations portant sur les cartes conceptuelles pour les bibliothèques numériques, les relations dynamiques et les graphes conceptuels, l'alignement d'ontologies et l'accès multilingue aux ontologies.

Enfin, la troisième session a été consacrée à la présentation de plusieurs applications terminologiques pour des secteurs aussi différents que l'ingénierie nucléaire, l'informatique, le domaine bancaire ou l'agriculture biologique. Il est à souligner que ces applications ont permis d'aborder différents points théoriques tels que la variation terminologique, la diachronie ou la structure des dictionnaires.

La richesse des débats qui ont animé ces deux jours de conférence – chaque présentation, questions comprises, s'est vue allouer plus de quarante cinq minutes de temps de parole – a été certainement une des plus belles récompenses pour les participants de TOTh 2009.

Christophe Roche

Président du Comité Scientifique

Table des matières

CONFERENCE INVITEE

| | |
|---|---|
| <i>La nomenclature biologique aujourd'hui : que reste-t-il de Linné ?</i> | 1 |
| Michel Laurin | |

SESSION 1

| | |
|--|----|
| <i>Approche lexico-sémantique de l'extraction terminologique : utilisation de ressources lexicographiques et validation sur corpus</i> | 19 |
| Bertrand Gaiffe, Evelyne Jacquey, Laurence Kister | |
| <i>Quelle place accorder aux corpus dans la construction d'une terminologie ?</i> | 33 |
| Marie Calberg-Challot, Pierre Lerat, Christophe Roche | |
| <i>Extraction de connaissances orientées évolution dans les textes techniques</i> | 53 |
| Kata Gabor, François Rousselot, François De Bertrand de Beuvron | |
| <i>Corpus et Web : deux alliés pour la construction de l'enrichissement automatique de classes conceptuelles</i> | 73 |
| Nicolas Béchet, Mathieu Roche, Jacques Chauché | |

SESSION 2

| | |
|--|-----|
| <i>Following the path between conceptual maps and visual thesauri</i> | 93 |
| Olga Bessa Mendes | |
| <i>Dynamic concept relations: a definition and representation proposal</i> | 107 |
| Chiara Messina | |
| <i>Construction et alignement d'ontologies pour évaluer le risque alimentaire</i> | 127 |
| Liliana Ibanescu, Patrice Buche, Juliette Dibie-Barthélemy | |
| <i>Accès multilingue à une ontologie par des correspondances avec un lexique pivot</i> | 143 |
| David Rouquet, Hong-Thai Nguyen | |
| <i>La reformulation : processus dynamique d'acquisition des connaissances. Le cas du discours technique arabe d'Internet</i> | 161 |
| Andrée Affeich | |

SESSION 3

| | |
|--|-----|
| <i>Structuration d'un dictionnaire de spécialité pour sa publication sur internet. Bénéfices du langage XML</i> | 181 |
| Jacques Joseph | |
| <i>Mémoire du Club informatique des grandes entreprises françaises (CIGREF) : nouveau plan de classement</i> | 197 |
| Jean-Yves Gresser, M.P. Lacroix | |
| <i>Les secteurs d'activité à l'épreuve du discours</i> | 217 |
| Frédéric Erlos | |
| <i>De l'agriculture biologique aux espaces naturels : une étude des syntagmes terminologiques à l'intérieur des textes de spécialité</i> | 235 |
| Elisa Lavagnino | |
| <i>Pages blanches</i> | 253 |

Multilinguïisation d'ontologies par des correspondances avec un lexique pivot

David Rouquet - Hong-Thai Nguyen

Résumé : Les ontologies sont parmi les représentations formelles de connaissance les plus utilisées en informatique. Le besoin d'un accès multilingue à ces connaissances apparaît tant au niveau des applications (Web Sémantique, RI, TA, etc.) que de la création et de l'enrichissement des ontologies. Après avoir précisément délimité le problème de la multilinguïisation d'ontologies, nous définissons formellement les objets de notre étude. Une rapide revue de l'état de l'art montre que les approches existantes pour répondre au problème posé ne sont pas satisfaisantes. Nous proposons une méthode basée sur une correspondance entre l'ontologie et un langage pivot, disposant d'un espace lexical autonome, et pouvant représenter les énoncés de la langue de façon formelle. Une application de cette méthode dans le projet OMNIA pour l'extraction d'information de textes multilingues, en vue d'une indexation et d'une recherche d'images est présentée. Nous exposons enfin la mise en œuvre de notre méthode, supportée par la plateforme de gestion de ressources lexicales PIVAX et utilisant le langage pivot UNL.

Mots-clés : ontologie, multilinguïisme, langage pivot

1. Introduction

Les ontologies formelles font partie des représentations informatiques de connaissances les plus utilisées. Les domaines d'application sont de plus en plus nombreux et mettent en avant des difficultés bien spécifiques. On peut se demander, avec l'avènement du Web Sémantique, comment les utilisateurs pourront contribuer à la construction d'ontologies et les utiliser sans pré-requis en logique formelle ? Comment utiliser automatiquement les ontologies dans des applications (RI, TALN, etc.) qui traitent des textes "tout venant" ? Comment les conceptualisations décrites avec des terminologies monolingues peuvent-elles être accessibles dans d'autres langues (CLIR, TA, etc.) ?

On voit que l'accès (contribution et utilisation) aux ontologies par le biais de la langue naturelle est un problème clef, tant pour les humains que pour les logiciels traitant des textes. Le bénéfice d'un accès multilingue est alors évident dans le contexte d'utilisation distribuée des ontologies. Les méthodes visant à la "multilinguisation d'ontologies" doivent selon nous répondre à certains critères que nous précisons dans cet article. Elles doivent être *modulaires* et *dynamiques*, sans contraindre l'ontologie ni interférer avec la conceptualisation.

On commencera dans cet article par définir précisément le problème de la multilinguisation d'ontologies. Nous proposerons ensuite une définition formelle des ontologies informatiques adaptée au problème traité et suffisamment générale pour englober toutes les ontologies rencontrées. Une brève revue de l'état de l'art montrera qu'aucune des approches que nous avons trouvé ne répond au problème dans son intégralité. Nous proposerons donc une méthode de multilinguisation d'ontologies par des correspondances avec un langage pivot. Cette méthode est illustrée par son utilisation dans le projet ANR OMNIA (recherche et indexation d'images accompagnées de textes), pour une extraction d'information dans des textes multilingues. Nous décrirons enfin la mise en œuvre de notre méthode, avec le langage pivot UNL et des ressources complémentaires comme WordNet, au sein de la plate-forme de gestion de ressources lexicales PIVAX.

2. Définition du problème

Selon (Gruber 1993), les ontologies "informatiques" sont des spécifications explicites et formelles d'une conceptualisation d'un domaine de connaissances.

Une ontologie comporte :

1. un treillis conceptuel, enrichi par des relations et des propriétés obéissant à des axiomes logiques (*T-box*) ;
2. un ensemble d'objets peuplant les concepts et décrits par les relations et les propriétés (*A-box*).

Chaque concept, objet et relation de l'ontologie est désigné par une *étiquette* constituée à partir de lexèmes ou locutions une langue naturelle. Les étiquettes ne sont pas nécessairement bien formées dans une langue (abréviations, mots "agglutinés", etc.) mais sont suffisamment explicites pour permettre une interprétation en langue naturelle. Idéalement, les ambiguïtés dans l'interprétation de ces étiquettes sont levées par le contexte au sein de l'ontologie. La connaissance contenue dans une ontologie est accessible tant par des agents logiciels (grâce à une sémantique formelle) que par des humains (grâce aux étiquettes et à leur interprétation pragmatique). Par *accès* à une ontologie, nous entendons : contribuer à la connaissance ontologique et utiliser cette connaissance.

L'ajout de connaissances dans une ontologie peut, en outre, se faire manuellement via un éditeur spécialisé (par exemple Protégé¹) ou de façon automatisée grâce à des processus d'extraction d'information. La contribution manuelle est effectuée par un humain qui retranscrit naturellement une conceptualisation via sa langue maternelle ; l'extraction d'information est souvent faite à partir de textes. Aussi, dans ces deux cas, le traitement d'une langue naturelle est nécessaire. D'autre part, la majorité des utilisateurs humains ne sont pas en mesure de formuler des requêtes formelles SPARQL² pour consulter une ontologie et certaines applications (RI, TA, etc.) doivent utiliser automatiquement une ontologie pour traiter des textes. Ici encore, l'ontologie doit être accessible dans la langue de l'utilisateur ou du document.

Il est donc intéressant de proposer des méthodes permettant l'ajout de données multilingues à une ontologie donnée, pour en favoriser l'accès (par des agents humains ou logiciels). Selon nous, ces méthodes doivent répondre à certains critères. Premièrement, elles ne doivent pas contraindre la création de l'ontologie pour ne pas décourager les contributions humaines ou complexifier les contributions automatiques. Ensuite, la conceptualisation spécifiée dans l'ontologie sert un but bien précis que nous ne connaissons pas nécessairement, les méthodes de

1 <http://protege.stanford.edu/>

2 www.w3.org/TR/rdf-sparql-query/

multilinguisation ne doivent donc pas interférer avec cette conceptualisation. Enfin, les méthodes doivent idéalement être *modulaires* et *dynamiques*. C'est à dire qu'elles doivent respectivement : permettre l'ajout de nouvelles langues sans recours aux spécialistes qui ont fourni la conceptualisation et s'adapter automatiquement à des modifications incrémentales de l'ontologie sans repartir de zéro.

Nous cherchons donc à résoudre le problème suivant :

Étant donnée une ontologie informatique quelconque, permettre l'ajout de données multilingues à cette ontologie, sans interférer avec la conceptualisation, de façon modulaire et dynamique.

3. Définition formelle des ontologies

3.1. Objectifs

Bien que certains aspects, comme la nécessité d'un caractère consensuel, soient encore sujets à débat, la définition intuitive de (Grubber 1993) est aujourd'hui communément admise pour les ontologies informatiques. Cependant, aucune définition formelle des ontologies ne s'est imposée et l'on en trouve de multiples concurrentes (par exemple (Maedche *et al.* 2003) et (Euzenat *et al.* 2007)). Il parait en fait raisonnable d'en choisir une selon ses besoins spécifiques.

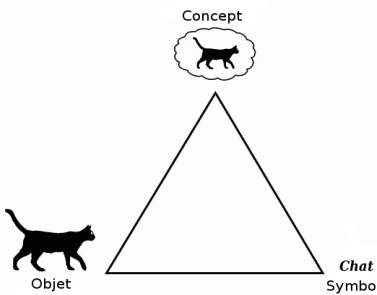


Figure 1. Le triangle sémiotique entre ces éléments, comme présenté dans le triangle sémiotique de la figure 1 (Ogden *et al.* 1923).

Dans notre cas, il s'agit de développer une méthode pour permettre l'ajout de données multilingues. La définition doit donc identifier clairement les éléments linguistiquement pertinents d'une ontologie. D'autre part, une ontologie est constituée de trois éléments principaux : concept, objet, symbole (ou étiquette). Une bonne définition formelle doit donc également faire une distinction claire

entre ces éléments, comme présenté dans le triangle sémiotique de la figure 1 (Ogden *et al.* 1923). La définition doit enfin être suffisamment générale pour englober toutes les ontologies informatiques, indépendamment du point de vue adopté pour leur utilisation ou leur développement.

3.2. Définitions formelles

Nous adoptons comme définition formelle pour les ontologies informatiques celle de (Maedche *et al.* 2003). Nous l'avons complétée pour répondre au mieux aux critères présentés précédemment. Elle est composée des notions d'*ontologie abstraite*, d'*instanciation* et de *lexique*.

Définition 1 : Une *ontologie abstraite* est une structure $\mathcal{O} = (\mathcal{C}, \top, \mathcal{R}, \sigma, \leq_{\mathcal{C}}, \leq_{\mathcal{R}}, \mathcal{L}, \mathcal{T})$ avec :

- \mathcal{C} et \mathcal{R} des ensembles finis disjoints d'étiquettes de concepts et d'étiquettes de relations ;
- $\sigma : \mathcal{R} \rightarrow \mathcal{C} \times \mathcal{C}$ une fonction signature, retournant le domaine d'une relation ;
- $\leq_{\mathcal{C}}$ et $\leq_{\mathcal{R}}$ des ordres partiels sur \mathcal{C} et \mathcal{R} ;
- $\top \in \mathcal{C}$ une borne supérieure pour $\leq_{\mathcal{C}}$ de sorte que cet ordre forme un semi-treillis sur \mathcal{C} , nommé treillis conceptuel ;
- \mathcal{L} une théorie logique, dotée d'une sémantique formelle, dont la signature contient les constantes de \mathcal{C} et \mathcal{R} , les ordres $\leq_{\mathcal{C}}$ et $\leq_{\mathcal{R}}$ ainsi que σ ;
- \mathcal{T} un ensemble d'axiomes exprimés dans la logique \mathcal{L} . On l'appelle aussi la T-box (terminological box) ;

Exemple : voici une ontologie très simple :

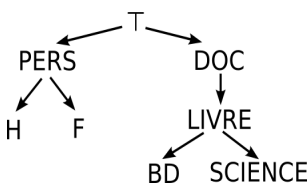


Figure 2. Treillis des concepts

Les étiquettes des concepts s'interprètent comme : personne, homme, femme, document, livre, bande dessinée, ouvrage scientifique. La relation AUT exprime qu'une personne est l'auteur d'un document.

$\mathcal{R} = \{AUT\}$

$\mathcal{C} = \{PERS, H, F, DOC, LIVRE, BD, SCIENCE\}$

$\sigma(AUT) = PERS \times DOC$

La T-box, exprimée avec la théorie des ensembles, est :

1. $H \cap F = \emptyset$
2. $PERS = H \cup F$
3. $\forall d \in DOC \{a \in PERS, AUT(a, d)\} \neq \emptyset$

Les deux premiers axiomes expriment qu'une personne est soit un homme soit une femme. Le troisième exprime qu'un document a au moins un auteur.

Définition 2 : une *instanciation* pour une ontologie abstraite $\mathcal{O} = (C, \top, R, \sigma, \leq_C, \leq_R, \mathcal{L}, \mathcal{T})$ est une structure $Inst = (E, \mathcal{A})$ avec :

- E un ensemble fini d'individus ;
- \mathcal{A} un ensemble d'axiomes exprimés dans la logique \mathcal{L} . On l'appelle aussi la A-box (assertional box) ;

L'ontologie *instanciée* pourra être notée :

$$\mathcal{O} = (C, \top, R, E, \sigma, \leq_C, \leq_R, \mathcal{L}, \mathcal{T}, \mathcal{A})$$

Exemple : une instanciation de l'ontologie prise en exemple peut être donnée par la A-box suivante :

- $NicolasBourbaki \in PERS$
- $FrankMiller \in PERS$
- $ElementsdeMathematiques \in SCIENCE$
- $SinCity \in BD$
- $AUT (ElementsdeMathematiques, NicolasBourbaki)$
- $AUT (SinCity, FranckMiller)$

Définition 3 : un *lexique* pour une ontologie $\mathcal{O} = (C, \top, R, E, \sigma, \leq_C, \leq_R, \mathcal{L}, \mathcal{T}, \mathcal{A})$ est une structure $Lex = (D, Ref)$ avec :

- D un ensemble de lexies (mots ou locutions) ;
- $Ref \subseteq ((C \cup R \cup E) \times D)$ une relation appelée affectation lexicale.

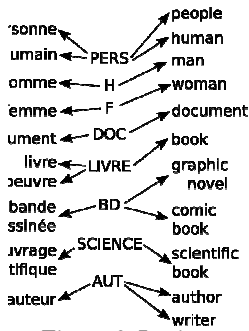


Figure 3. Lexiques

Exemple : la figure 3 présente deux lexiques (français et anglais) pour l'ontologie prise en exemple.

Maintenant que nous avons délimité le problème et défini formellement les objets de notre étude, nous présentons les approches existantes pour l'accès multilingue à une ontologie et montrons en quoi elles ne semblent pas satisfaisantes.

4. Approches existantes du problème

4.1. Traduction vers des langues cibles

(Espinoza *et al.* 2008) propose une méthode pour traduire les étiquettes de l'ontologie, directement vers d'autres langues. Un service adapté propose d'abord des traductions possibles en consultant des ressources linguistiques (dictionnaires bilingues, bases lexicales, etc.) ; la liste des traductions est ensuite classée selon leur qualité probable en utilisant les voisinages dans le treillis conceptuel.

Les méthodes de désambiguïsation utilisant le treillis conceptuel sont intéressantes mais doivent être appliquées de nouveau pour chaque langue cible. Le travail de désambiguïsation n'est pas factorisé entre les différentes langues et le simple apport de ressources lexicales ne suffit pas à augmenter le nombre de langues couvertes.

4.2. Greffe de "sous-ontologies" linguistiques

(Buitelaar *et al.* 2006) propose une trame de "sous-ontologie" linguistique permettant de stocker la traduction d'un concept accompagnée de données morpho-syntactiques. Il faut, pour chaque concept de l'ontologie source, instancier la trame dans la langue cible et la greffer au concept comme illustré dans les figures 4 et 5.

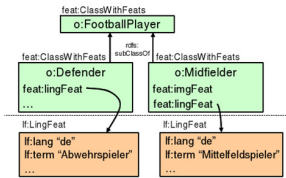


Figure 4. Greffe de la trame linguistique

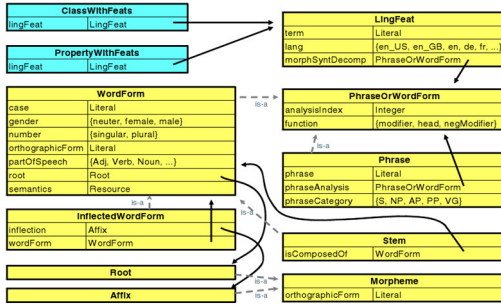


Figure 5. Détail de la trame linguistique

Cette approche rend cependant l'ontologie source bien plus complexe. De plus, l'instanciation et la greffe des sous-ontologies linguistiques doivent être faites pour chaque langue cible sans qu'aucune méthode automatisée n'ait été proposée.

4.3. Relier l'ontologie à des WordNets³

(Niels *et al.* 2003) décrit la correspondance entre SUMO (Suggested Upper Merged Ontology) et WordNet, calculée pour rendre l'ontologie accessible à des humains et utilisable automatiquement par des applications traitant des textes. Cette correspondance comprend des relations de synonymie, d'hyponymie et d'instanciation entre les concepts de l'ontologie et les *synsets* de WordNet.

Cette méthode ne permet pas la multilinguisation. Elle est cependant intéressante puisqu'elle permet l'ajout de données monolingues à une ontologie, ce qui est un sous-problème de celui qui nous concerne. Dans cette méthode, les ressources ontologiques et linguistiques ne sont pas clairement séparées (des relations d'hyponymie et d'instanciation sont déjà présentes dans l'ontologie), ce qui risque d'interférer avec la conceptualisation proposée dans l'ontologie. Il se trouve d'ailleurs qu'un des objectifs de cette correspondance avec WordNet est de réviser et de compléter l'ontologie, ce qui n'est pas souhaitable pour nous. D'autre part, en vue d'une multilinguisation, le travail décrit est à effectuer pour chaque nouvelle langue, ce qui est loin de se résumer à l'ajout de ressources linguistiques.

Le système KYOTO (Vossen *et al.* 2008) propose un environnement Wiki pour le développement collaboratif d'une ontologie interlingue et de sa correspondance avec des WordNets (actuellement sept langues

³ <http://wordnet.princeton.edu/>

supportées : basque, chinois, allemand, anglais, italien, japonais et espagnol).

Ici encore, l'approche ne permet pas de traiter séparément les aspects de conceptualisation et de multilinguisation (et c'est même un de ses buts) puisque des experts de chaque langue doivent proposer, à partir des entrées des WordNets, des liens vers les concepts existants ou vers de nouveaux concepts consensuels à insérer dans l'ontologie.

5. Approche avec un langage pivot

5.1. Principe

En vue de permettre l'accès multilingue à une ontologie, développée *a priori*, nous proposons de passer par une représentation pivot de la langue. Notre objectif est de construire un lexique non ambigu pour l'ontologie, dans un langage pivot approprié, et d'utiliser ce lexique interlingue comme portail vers les langues naturelles. Pour permettre cela, le langage pivot doit disposer d'un espace lexical autonome et non ambigu qui est mis en correspondance avec les étiquettes de l'ontologie par des affectations lexicales. Il doit également permettre la construction de syntagmes pour traiter les concepts portant des étiquettes "composées".

Cette méthode présente plusieurs avantages. Premièrement, l'inévitable travail de désambiguïsation pour relier les concepts à un lexique est "factorisé". Il est nécessaire pour le calcul du lexique pivot (et ses mises à jour en cas de modification de l'ontologie), mais pas pour l'ajout de nouvelles langues. Une fois le lexique pivot calculé, l'ajout de nouvelles langues peut se faire par la simple acquisition de dictionnaires reliant la langue cible au langage pivot. En outre, la construction de ces ressources ne requiert pas un expert du domaine de l'ontologie compétent dans la langue cible, c'est une tâche linguistique (la méthode est bien modulaire). D'autre part, la méthode proposée ne contraint en aucune manière les processus de création ou de contribution pour l'ontologie puisque les correspondances sont calculées ou mises à jour *a posteriori*. Enfin, la méthode est respectueuse de la conceptualisation proposée dans l'ontologie car les affectations lexicales sont clairement distinguées des relations initialement présentes dans l'ontologie.

La mise en œuvre concrète du stockage et de la gestion des correspondances est décrite dans la partie 6. Le paragraphe suivant

explique la méthode adoptée dans le projet ANR OMNIA⁴ pour permettre l'accès multilingue à une ontologie en utilisant cette approche par correspondance avec un langage pivot.

5.2. Exemple d'accès à une ontologie

Un des buts du projet OMNIA est de développer un outil de recherche pour des entrepôts d'images en ligne. A cet effet, une ontologie de catégorisation des images est construite. Les images seront décrites dans la A-box grâce à la fusion des données issues de l'analyse visuelle et de l'analyse des légendes et des textes compagnons écrits en langue naturelle "tout venant". Plusieurs langues seront par ailleurs proposées à l'utilisateur pour formuler ses requêtes librement (mots clefs, phrases, etc.).

L'accès multilingue à l'ontologie dans le projet concerne donc l'indexation des images dans la A-box à l'aide des légendes textuelles, et le traitement des requêtes de l'utilisateur. La même méthode est employée pour traiter ces deux aspects, comme illustré dans la figure 6.

Il s'agit d'annoter les textes avec les lexies interlingues non ambiguës (lexèmes ou locutions) du langage pivot sans procéder à une analyse syntaxique poussée. On sait avec (Daoud 2006) que c'est une approche viable pour initier une extraction de contenu. L'annotation est réalisée de façon automatique grâce à un dictionnaire "langue naturelle" – "langage pivot" et à des procédés de désambiguïsation. On réalise ensuite l'extraction de contenu. Ce processus prend en entrée les annotations interlingues et la correspondance "langage pivot" – "ontologie". Il retourne les informations pertinentes (i.e. qui peuvent être représentées dans l'ontologie) formatées dans le langage de description ou de requête de l'ontologie. Les informations peuvent alors, selon le cas, être stockées dans la A-box ou soumises à un raisonneur pour résoudre la requête.

⁴ www.omnia-project.org

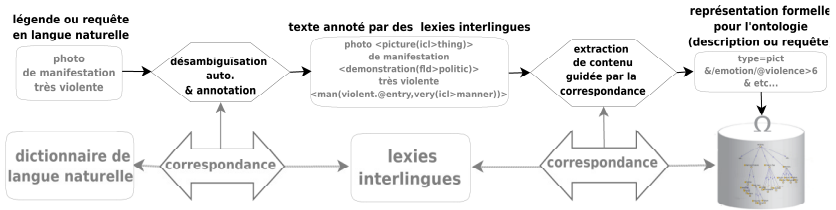


Figure 6. Accès à l'ontologie à partir de textes

6. Mise en œuvre

6.1. La plate-forme PIVAX

PIVAX (Nguyen *et al.* 2007) est une plateforme en ligne pour la gestion des ressources lexicales de systèmes de TA utilisant un pivot lexical. Elle a été développée à partir de la plateforme générique Jibiki (Sérasset 2005) qui différencie l'organisation des volumes de la base lexicale (*macrostructure*) et l'organisation des éléments de chaque volume (*microstructure*). Les ressources qui respectent une syntaxe XML peuvent être importées directement par la simple adjonction d'un fichier de métadonnées « Xpath » décrivant leur microstructure. Un exemple de fichier de métadonnées est présenté dans la figure 8 au paragraphe 6.2.

La macrostructure de PIVAX est composée de trois couches. Pour chaque langue supportée, on trouve :

- un ou plusieurs volumes de *lexies*. Les *lexies* correspondent à des sens de mots dans un dictionnaire.
- un unique volume d'*axèmes* (acceptions monolingues). Un *axème* relie des lexies synonymes dans un même langage.
- Un volume partagé d'*axies* (acceptions interlingues). Une *axie* relie des axèmes synonymes.

Au niveau de la microstructure, les lexies contiennent un lemme et des informations complémentaires (classe, définition, statut, etc.). Les axèmes et les axes sont des liens, représentés simplement comme des ensembles de lexies et d'axèmes respectivement. Nous exploiterons également la possibilité de représenter des relations entre axèmes ou entre axes.

L'organisation des notions précédentes au sein de PIVAX est illustrée dans la figure 2. Le "langage pivot" n'occupe pas une place centrale mais est représenté comme les autres langues. Les paragraphes suivant décrivent les ressources proposées pour la mise en œuvre de l'accès multilingue à une ontologie ainsi que leur intégration et leur utilisation dans la plate-forme PIVAX .

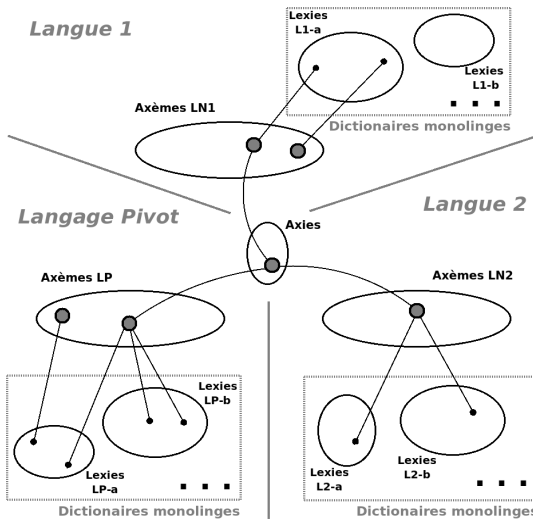


Figure 7. La plate-forme PIVAX

6.2. L'ontologie

L'ontologie que l'on cherche à multilingualiser est représentée dans PIVAX comme un dictionnaire. Les concepts et les instances forment un volume d'axèmes dans lequel les relations de l'ontologie sont également représentées (spécialisation, instanciation, etc.). Les étiquettes des concepts, instances et relations sont stockées dans un volume de lexies.

Les langages de description d'ontologies basés sur XML permettent une importation immédiate dans PIVAX. La figure 8 présente le fichier de métadonnées pour importer dans un volume de lexies les (étiquettes des) concepts d'une ontologie décrite en OWL. D'autres métadonnées doivent être ajoutées pour importer les instances et les relations dans les volumes de lexies et d'axèmes.

```

<volume-metadata>
<comments>PIVAX OWL - OMNIA Ontology for OMNIA project</comments>
  <cdm-elements>
    <cdm-volume      xpath="/rdf_RDF"/>
    <cdm-entry       xpath="/rdf_RDF/owl_Class"/>
    <cdm-entry-id    xpath="/rdf_RDF/owl_Class/@rdf_ID"/>
    <cdm-headword    d:lang="owl"
  xpath="/rdf_RDF/owl_Class/text()"/>
  </cdm-elements>
  <administrators>
    <user-ref      name="nguyenht"/>
  </administrators>
  <system      name="omnia">
    <copyright>Copyright by OMNIA project</copyright>
    <description>OWL Ontology dictionary</description>
    <organization>GETALP-LIG</organization>
    <adress/>
    <responsability>Christian.Boitet@imag.fr</responsability>
    <url_link>http://www.ellemme.org/</url_link>
  </system>
  <xmllschema-ref      xlink:href="pivax_local.xsd"/>
  <volume-ref      xlink:href="Pivax_owl-omnia.xml"/>
  <template-entry-ref      xlink:href="Pivax_owl-omnia-
template.xml"/>
</volume-metadata>

```

Figure 8. Métadonnées pour PIVAX

6.3. Le langage pivot UNL⁵

Le terme UNL (Universal Networking Language) recouvre en particulier deux choses différentes :

- le projet international UNL, lancé en novembre 1996 par l'UNU (Université des Nations Unies) à Tokyo ;
- le langage UNL, qui est un langage pivot "anglo-sémantique", et pas une langue humaine naturelle ou construite (comme l'espéranto) ;

Les expressions du langage UNL représentent le sens d'un énoncé par une structure sémantique abstraite (un graphe) d'un énoncé anglais équivalent.

Le vocabulaire de UNL est constitué de lexies interlingues appelées UW (Universal Words, en français, Unités de Vocabulaire Virtuel). Idéalement, elles déterminent de manière non ambiguë un concept existant dans l'ensemble des langues considérées (certains concepts, comme atterrir ou amerrir n'existent que dans certaines langues, les autres ne référant qu'à des concepts moins fins. Une UW est composée de :

1. un *mot-vedette*, si possible dérivé de l'anglais, qui peut être un mot, une expression ou encore une phrase entière.

5 <http://www.undl.org/>

2. une *liste de restrictions* servant à délimiter le concept précis porté par l'UW.

Exemples :

- `book(icl>thing)` et `book(icl>do, agt>human, obj>thing)` pour lesquels le sens de l'UW est précisé par des restrictions ;
- `ikebana(icl>flower_arrangement)` dont le mot-vedette a été importé du japonais ;
- `go_down` dont le mot-vedette est une expression et dont le sens n'a pas besoin d'être précisé par des restrictions.

Les UW sont organisées dans un réseau sémantique nommé UNLKB6 (UNL Knowledge Base). Ce réseau contient des relations sémantiques et syntaxiques pondérées décrivant le comportement des UW les unes par rapport aux autres. Il facilite l'interprétation des expressions UNL.

Plusieurs dictionnaires d'UW sont disponibles et/ou en cours de développement (projet initial de l'UNU, consortium U++⁷, etc.). Chacun de ces dictionnaires constitue un volume de lexies dans PIVAX. Ils devront être reliés entre eux par des axèmes. Certaines UW sont en effet équivalentes mais n'ont par exemple pas le même mot-vedette d'un volume à l'autre, comme `book(agt>human,obj>thing)` et `reserve(agt>human,obj>thing)`. Le volume d'axèmes contient également les relations de l'UNLKB. Les dictionnaires "UNL" – "langue naturelle" disponibles sont représentés par des axes entre les axèmes "UNL" et les axèmes "langue naturelle".

6.4. WordNet

WordNet est une base lexicale de l'anglais développée à l'Université de Princeton (Fellbaum 1998), toujours en développement et portée dans d'autres langues (voir par exemple le projet EuroWordNet⁸). Bien que cette base lexicale ne soit pas fondamentale pour notre méthode de multilinguisation, il est intéressant d'en disposer parmi nos ressources au sein de PIVAX. WordNet est en effet largement utilisé pour de nombreuses applications de traitement de la langue (désambiguïsation, recherche d'information, etc.) que l'on pourra alors exploiter.

Les éléments de WordNet sont des *synsets*. Ce sont des ensembles de termes qui présentent une interprétation commune dans au moins un

6 <http://www.undl.org/unlsys/unl/unl2005/UNLKB.htm>

7 <http://www.unl.fi.upm.es/consorcio/>

8 <http://www.illc.uva.nl/EuroWordNet/>

contexte d'utilisation. Idéalement, ces synsets représentent des concepts. Ils sont par ailleurs reliés par des relations sémantiques et lexicales.

Sous PIVAX, les lexies des volumes "WordNet" sont constituées d'un couple (mot, synset) comparable à une entrée dans un dictionnaire (mot, sens de mot). Le volume d'axèmes est constitué de l'ensemble des synsets et des relations entre ces synsets. Pour faciliter l'importation dans PIVAX, nous avons utilisé la conversion de WordNet sous forme RDFS/OWL⁹ dont la syntaxe est une spécification XML. D'autre part, les UW de certains volumes "UNL" sont construites à partir de synsets de WordNet (par exemple les "UW++" du consortium U++). Ces volumes "UNL" et "WordNet" sont alors mis en correspondance par des axes.

6.5. Organisation générale

La mise en œuvre de notre méthode dans PIVAX revient donc à calculer des axes entre les volumes "ontologie" et "UNL". Pour une ontologie dont les étiquettes sont formées à partir d'une langue naturelle donnée "L1", cela est fait grâce à la combinaison des axes "ontologie" – "L1" et des axes "L1" – "UNL", ainsi qu'à des procédés de désambiguïsation. L'accès à l'ontologie par une autre langue est alors réalisé grâce aux axes entre UNL et cette autre langue.

La figure 8 décrit l'instance spécifique de la plateforme PIVAX qui supporte les correspondances entre une ontologie donnée, UNL, les langues naturelles et des ressources complémentaires (pour l'instant WordNet uniquement).

⁹ <http://www.w3.org/TR/wordnet-rdf/>

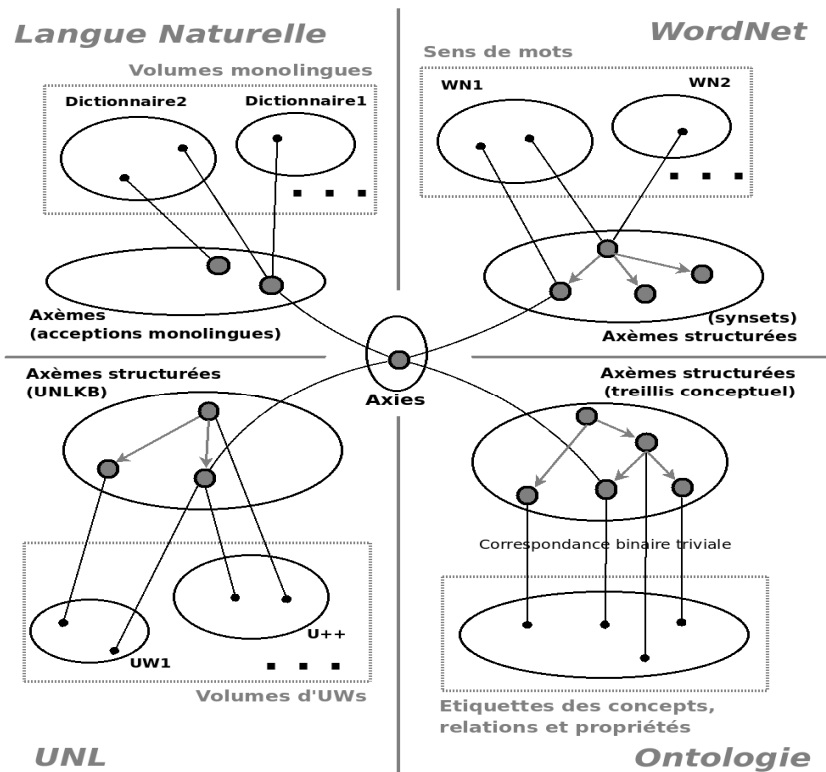


Figure 9. PIVAX pour la multilinguisation d'ontologies

7. Conclusion et perspectives

Nous avons montré dans cet article que l'ajout de ressources multilingues aux ontologies informatiques constitue un problème en soi. Disposer de telles ressources est pertinent pour de multiples applications utilisant les ontologies (recherche d'information, traduction automatique, etc.), ainsi que pour les processus de création et d'enrichissement des ontologies. Des caractéristiques requises par les méthodes visant à résoudre ce problème ont été mises en avant (caractères modulaire et dynamique, respect de la conceptualisation formalisée, liberté dans la création et l'enrichissement des ontologies). Une rapide revue de l'état de l'art a montré en quoi les méthodes existantes ne répondaient que partiellement au problème posé, et nous avons proposé une approche plus satisfaisante par correspondance entre l'ontologie et un langage pivot dont le vocabulaire est un ensemble d'acceptions interlingues (UW++). Cette

méthode est appliquée dans le projet OMNIA pour l'extraction d'information à partir de textes multilingues "compagnons" d'images, en vue de l'indexation et de la recherche d'images. La mise en œuvre de notre méthode est supportée par la plate-forme de gestion de ressources lexicales PIVAX.

Si le stockage et la gestion des correspondances sont concrètement résolus par l'utilisation de PIVAX, le calcul et la mise à jour de ces correspondances reste une tâche à explorer en détail. En particulier, l'étude de la nature des correspondances et de leurs propriétés pour assurer un caractère dynamique à la méthode fait partie de l'objet d'une thèse en cours. Des procédés de désambiguïsation automatique sont également étudiés en utilisant des vecteurs conceptuels (Schwab 2005). Cette dernière ressource pourrait également être exploitée avec PIVAX.

Remerciements

Les auteurs tiennent à remercier leurs directeurs et/ou collègues Christian Boitet, Valérie Belynyck et Didier Schwab pour les relectures et les précieux conseils.

Notre gratitude va également aux partenaires du projet ANR OMNIA (ANR-07-MDCO-009-02) qui nous offrent le cadre applicatif et les ressources financières pour mener ces travaux.

Bibliographie

Buitelaar P., Sintek M. et Kiesel M. (2006) : A Multilingual/Multimedia Lexicon Model for Ontologies, *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, vol. 4011, pp. 502-513, ISBN: 978-3540-34544-2, Springer

Daoud D. (2006) : Il faut et on peut construire des systèmes de commerce électronique à interface en langue naturelle restreints (et multilingues) en utilisant des méthodes orientées vers les sous-langages et le contenu, *Thèse de Doctorat*, 290 p., Université Joseph Fourier, Grenoble

Espinoza M., Gómez-Pérez A. et Mena E. (2008) : Enriching an Ontology with Multilingual Information, *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, vol. 4011, pp. 333-347, ISBN: 978-3-540-34544-2, Springer

Euzenat J. and Shvaiko P. (1998) : *Ontology Matching*, ISBN: 3540496114 Springer, 2007.

Fellbaum C. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, ISBN : 026206197X, The MIT Press

Gruber T.R. (1993) : A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition*, vol. 5-2, pp. 199-220, ISSN: 1042-8143, Academic Press Ltd.

Nguyen H.T., Boitet C., Sérasset G. (2007) : PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot, *The Seventh Symposium on Natural Language Processing (SNLP-2007)*, Bangkok, Thailand

Maedche A., Neumann G. et Staab S. (2003) : Bootstrapping an ontology-based information extraction system, *Intelligent exploration of the web*, pp. 345-359, ISBN: 3-7908-1529-2, Physica-Verlag GmbH

Niles I. et Pease A. (2003) : Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, pp. 23--26, Las Vegas

Ogden C. et Richards C. (1923) : The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism, *Harcourt*

Schwab D. (2005) : Approche hybride –lexicale et thématique– pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de textes, *Thèse de Doctorat*, 363 p., Université Montpellier 2

Vossen P. et al. (2008) : KYOTO : A system for Mining, Structuring and Distributing Knowledge Across Languages and Cultures, *Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco

A propos des auteurs

David Rouquet

GETALP – LIG

385 av. de la Bibliothèque

Domaine Universitaire, BP 53

38041 Grenoble Cedex 9

David.Rouquet@imag.fr

<http://www.liglab.fr>

Hong-Thai Nguyen

GETALP – LIG

385 av. de la Bibliothèque

Domaine Universitaire, BP 53

38041 Grenoble Cedex 9

Hong-Thai.Nguyen@imag.fr

<http://www.liglab.fr>