Modèle d'entrepôt de ressources hétérogènes pour le traitement sémantique des documents

Nizar Ghoula, Gilles Falquet, Jacques Guyot

Dans Document numérique 2010/2 Vol. 13 , pages 97 à 124
Éditions JLE

ISSN 1279-5127
ISBN 9782746232334
Date de mise en ligne : 13/12/2010

Modèle d'entrepôt de ressources hétérogènes pour le traitement sémantique des documents

Nizar Ghoula — Gilles Falquet — Jacques Guyot

Centre universitaire d'informatique
Université de Genève
7, route de Drize
CH-1227 Carouge, Suisse
{Nizar.Ghoula, Gilles.Falquet}@unige.ch, jacques@simple-shift.com

RÉSUMÉ. *Les ressources documentaires sont riches en connaissances d'un domaine donné. L'extraction et la représentation de ces connaissances est un problème largement exploré dont la solution est basée sur l'utilisation de ressources ontologiques, terminologiques et linguistiques. Ces ressources ont des types et des modèles de représentations hétérogènes. Nous effectuons des représentations multiples des ressources à l'aide de différents modèles de contenus. Ceci est modélisé par une ontologie générique qui formalise les modèles des ressources, les opérations que nous pouvons effectuer sur ces ressources et les processus de gestion de connaissances. L'ontologie de ressources permet de construire un entrepôt de ressources hétérogènes et facilite leur interopérabilité.*

ABSTRACT. *Multiple sources of information can improve knowledge mangement if they are properly combined and processed. Knowledge engineering usually relies on knowledge resources, typically ontologies. We propose a domain-independent framework which models, combines and represents heterogenous sources of information. Our aim is to build a resources repository and afford operations of loading, storing, indexing, translating, generating and matching different resources. We propose an ontology as a model of these resources and we explain how can we represent, annotate and load new resources into our repository. These resources are treated to fit a specific need in a knowledge management process.*

MOTS-CLÉS : *ontologie de ressources, multilingues, terminologie, alignement, entrepôt de ressources.*

KEYWORDS: *ontology of resources, multilingual, terminology, alignment, resources repository.*

DOI:10.3166/DN.13.2.97-124

1. Introduction

Le traitement sémantique des documents, qu'il s'agisse d'indexation sémantique, d'alignement de textes, de désambiguïsation, de traduction, etc., requiert des connaissances de nature linguistique, terminologique et ontologique. Ces connaissances existent actuellement sous forme de ressources de différents types, telles que les terminologies, les glossaires, les ontologies (générales ou de domaine), les dictionnaires multilingues ou encore les corpus de textes (simples ou parallèles). À cette hétérogénéité des types il faut ajouter l'hétérogénéité des représentations. Il existe en effet, pour chaque type de ressource, de multiples formalismes, langages et formats de représentation. Par exemple, les ontologies peuvent s'exprimer, en logique des prédicats, dans une logique de description, sous forme de réseaux sémantiques ou avec d'autres formalismes. Tout système de gestion de connaissances pour le traitement sémantique des documents devra donc prendre en compte cette hétérogénéité.

Le développement et la mise à disposition d'un nombre croissant de ressources de connaissances sur le web pose la question de la recherche des ressources les plus adéquates pour un traitement sémantique donné. Des moteurs et services de recherche spécialisés ont été développés à cet effet. Le système Swoogle[1] indexe environs 10 000 ontologies, d'autres services comme DAML[2] et BioPortal[3] (Noy *et al.*, 2008), offrent une possibilité de recherche plus fine sur les ontologies en exprimant des requêtes sur les entités des ontologies (Ding *et al.*, 2001). Dans le même contexte d'utilisation, le moteur de recherche d'ontologies Watson[4] permet de repérer et d'indexer les ontologies du web sémantique en gardant des références vers leurs entités. Pour une requête basée sur des mots-clés, Watson (Sabou *et al.*, 2007) renvoie une réponse sous forme d'une liste d'entités ontologiques avec des liens vers les ontologies correspondantes. D'autres systèmes existants offrent l'accès à des thésaurus, glossaires ou dictionnaires multilingues dont le portail TermSciences[5] (pour les ressources terminologiques) et le portail CNRTL[6] (pour les ressources linguistiques).

L'utilisation d'une ressource de connaissance se réduit parfois à l'utilisation d'un fragment de celle-ci, par exemple, un ensemble d'entités (concepts, propriétés et/ou axiomes) (Bouquet *et al.*, 2003) d'une ontologie, ou un sous-ensemble des textes d'un corpus. Dans d'autres cas, on aura besoin de plusieurs ressources, éventuellement hétérogènes et dans différentes langues, pour réaliser un traitement (Lopez *et al.*, 2009). À titre d'exemple, si une tâche d'indexation sémantique des documents nécessite une ontologie en français[7] inexistante dans les entrepôts de ressources, un utilisateur peut vouloir la générer à partir d'une ontologie existante de même type, en anglais, et d'un

1. http://swoogle.umbc.edu.
2. http://www.daml.org/ontologies.
3. http://bioportal.bioontology.org/.
4. http://watson.kmi.open.ac.uk/WatsonWUI/.
5. http://www.termsciences.fr/spip.php?rubrique23.
6. Centre national de ressources textuelles et lexicales : http ://www.cnrtl.fr/.
7. Par "ontologie en français" nous entendons une ontologie dont les concepts sont étiquetés par des termes en français.

dictionnaire de traduction anglais-français et/ou d'un ensemble d'alignements entre des concepts d'ontologies. De ce fait, les nouveaux usages créent de nouvelles ressources qui pourront elles-mêmes, à condition d'être correctement décrites et stockées, servir pour d'autres traitements.

Notre objectif est de créer un système de gestion de connaissances capable de gérer les différents types de ressources intervenant dans les traitements sémantiques. Il s'agit d'une part de décrire les caractéristiques des ressources, sous forme de métadonnées, pour faciliter leur indexation. D'autre part il faut permettre de générer, par sélection, compositions, et d'autres opérations, de nouvelles ressources répondant à des besoins particuliers.

Dans cette contribution, nous présentons notre approche fondée sur la définition des métadonnées et de modèles de représentation dans une ontologie de ressources et la définition d'opérateurs de traitement des ressources.

La deuxième section est consacrée à l'identification des ressources que nous traitons et des modèles proposés dans la littérature. La troisième section décrit les niveaux de notre approche de représentation et modélisation des ressources hétérogènes. La quatrième section décrit le modèle de ressources hétérogènes que nous avons réalisé sous forme d'ontologie appelée *TOK_Onto*. La dernière section décrit un processus d'importation et de stockage de ces ressources et montre une implémentation de ce modèle.

2. Les ressources

Ce travail fait partie d'une réflexion sur l'exploitation des ressources de connaissances et leur traitement dans un même contexte. Autour de cette réflexion, on identifie plusieurs problématiques : Comment utiliser ces ressources dans un même contexte ? Comment représenter des ressources hétérogènes ? Peut-on prendre en compte "toutes" les connaissances que contiendrait une ressource ? Quelle est la nature des ressources d'information ou de connaissances ? ...

Nous traitons la problématique relative à l'hétérogénéité des représentations des ressources[8]. Nous allons tout d'abord déterminer la nature des ressources d'information et de connaissances dont on veut définir un modèle commun.

2.1. *Périmètre des ressources*

L'objet de base de notre approche est un modèle de description des ressources. Ces ressources ont plusieurs utilisations et définitions. Leur représentation dépend de leur

8. Le terme ressource dans ce qui suit désigne une ressource terminologique, ontologique ou linguistique.

usage. En nous basant sur une étude des ressources, nous avons organisé celles-ci en deux catégories principales.

2.1.1. *Ressources autonomes*

Les ressources autonomes désignent la catégorie des ressources dont l'existence est indépendante des autres ressources.

– *Les ressources ontologiques* : une ontologie a pour but de représenter une conceptualisation d'un domaine (Gruber, 1995). Cette conceptualisation consiste essentiellement en une définition des concepts du domaine et des relations existant entre ces concepts. Les ontologies sont exprimées à l'aide de formalismes (Wang *et al.*, 2007) qui fournissent des constructeurs pour la définition des entités ontologiques. Suivant le formalisme utilisé, les entités peuvent être des classes, propriétés, individus et axiomes (dans les logiques de description), des concepts et relations (dans les réseaux sémantiques), des classes, objets et associations (dans les modèles à objets), etc. Le choix du formalisme dépend de l'objectif pratique poursuivi lors de la construction de l'ontologie : échange de connaissances, référence commune, raisonnement automatique (inférences logiques), structuration de données, etc.

Dans le traitement des documents les ontologies servent, entre autres choses, à représenter le sens des termes d'un document (levée d'ambiguïté), comme référence lors de l'annotation sémantique des documents (principe du web sémantique), comme base de raisonnement pour les systèmes de recherche d'information précise ou les systèmes question-réponse.

– *Ressources terminologiques* : elles représentent des termes rigoureusement définis pour un domaine spécifique (Wright *et al.*, 1997). Ces ressources sont le résultat d'une étude théorique des dénominations des objets ou des concepts utilisés par un domaine de l'activité humaine. Cette étude se focalise sur le fonctionnement dans la langue des unités terminologiques et sur les problèmes de traduction, de classement et de documentation. Beaucoup de travaux de recherche se sont focalisés sur l'étude des terminologies (Zhu *et al.*, 2009) surtout dans le domaine biomédical. Parmi ces ressources, on trouve les thesaurus pour les systèmes d'indexation automatique, les référentiels terminologiques pour les systèmes de gestion de données techniques, les bases de données terminologiques pour l'aide à la traduction, etc. Les thesaurus sont généralement utilisés pour la recherche d'information. Chaque ressource de connaissances peut être associée à un ou plusieurs concepts représentés à l'aide d'un ensemble de termes. Dans les thesaurus les termes sont organisés suivant un nombre restreint de relations (hiérarchiques, d'équivalence et associatives) (Foskett, 1980).

– *Ressources linguistiques* : elles représentent les types de données et informations sur la langue. Ces ressources sont plus généralement utilisées pour le traitement automatique de la langue, l'apprentissage (pour entraîner les programmes de traduction automatique par des approches statistiques). Dans ce type de ressources on trouve les documents, les corpus, les hyperdocuments, etc. Les corpus sont des ressources contenant d'autres ressources (documents) et caractérisés par la taille, le langage, le registre de langue et le temps couvert par les entités de corpus.

En résumé, les ressources ontologiques définissent les concepts, les ressources terminologiques permettent de décrire les termes associés à chaque concept dans une langue et les ressources linguistiques servent à décrire les langues (dans lesquelles les concepts sont exprimés).

2.1.2. *Ressources d'enrichissement*

Les ressources d'enrichissement désignent les ressources résultant de l'application d'un processus (automatique ou humain) sur les ressources autonomes.

– *Ressources d'indexation* : résultent d'un processus par lequel les ressources appartenant à une collection sont étiquetées pour représenter les caractéristiques des ressources et les rendre exploitables par des services de recherche d'information. Les index peuvent avoir plusieurs formes en fonction des ressources utilisées. Parmi ces ressources on trouve : (i) les index par mots-clés basés sur les ressources linguistiques et les index hypertextuels (tels que les cartes des sites) structurés pour la navigation dans les documentations techniques électroniques ou sur les sites web ; et (ii) les index ontologiques ou conceptuels (annotations sémantiques) qui enrichissent la ressource initiale en associant à son contenu des éléments conceptuels lui permettant d'être utilisable, accessible et reconnue par un ensemble d'acteurs ou d'agents. Une annotation sémantique est une formalisation de l'interprétation du texte sous forme de métadonnées (Kiryakov *et al.*, 2004).

– *Ressources d'alignement* : des ressources ayant un degré d'expressivité variable et des formes simples ou complexes et résultant de l'application d'une procédure de mise en correspondance entre deux ressources de même type. Cette catégorie de ressource est utilisée dans les applications de gestion de connaissances. L'alignement sert à trouver des entités similaires dans des ressources différentes tout en préservant l'indépendance et l'intégrité de ces ressources. Parmi ces ressources on trouve (i) les alignements des termes et des ressources terminologiques ; (ii) les alignement des ressources linguistiques telles que les corpus de textes alignées dans différentes langues ; et (iii) les alignements d'ontologies, qui servent à mettre en correspondance les concepts des deux ontologies. Ces correspondances peuvent être l'inclusion, l'équivalence, la disjonction etc. (Euzenat *et al.*, 2007).

À ces deux catégories de ressources on peut ajouter un autre type de ressources autonomes qui s'inscrit sous le cadre de ressources complexes ou composées. Ce type de ressources peut combiner des ressources linguistiques, terminologiques et ontologiques ou des ressources autonomes avec des ressources d'enrichissement. Par exemple, un corpus comparable ou un corpus parallèle est une sorte de ressource complexe puisqu'il contient des documents (ressources autonomes) et des alignements entre textes (ressources d'enrichissement). Les hypertextes sémantiques sont des ressources complexes combinant des ressources linguistiques indexées par des concepts ou termes figurant dans des ressources ontologiques. Wikipédia est une ressource complexe composée de plusieurs types de ressources autonomes (documents, catégories etc.) et des ressources d'enrichissement (alignements de textes pour la traduction, etc.).

La diversité de représentation des connaissances dans les ressources s'explique par leurs utilisations différentes. Lorsque les connaissances à construire sont issues de documents, l'ingénierie de connaissances s'appuie sur des méthodologies développées dans le domaine de la linguistique et du traitement automatique des langues pour assurer une compréhension du contenu des documents considérés. Pour répondre à ces besoins en termes d'ingénierie de connaissance et recherche d'information, il faut offrir des modèles et des systèmes capables de représenter et d'utiliser les connaissances provenant de plusieurs ressources.

2.2. *Modèles de représentation des ressources*

S'il existe de nombreux modèles et langages de représentation des connaissances, ceux-ci sont généralement centrés sur un ou deux aspects : ontologique, terminologique, lexical, textuel, documentaire, etc. On trouve plus difficilement des modèles permettant de représenter divers aspects de la connaissance ou des ressources de différentes natures : (i) très peu de formalismes supportent l'utilisation de ressources complexes, beaucoup de formalismes se focalisent sur un niveau de représentation (linguistique, terminologique ou ontologique) ; (ii) même dans le cas où on peu mélanger les formalismes pour avoir des ressources plus riches, très peu d'utilisateurs le font, par exemple, on peut mélanger du OWL et SKOS pour avoir une ontologie et un thesaurus.

Le modèle proposé par (Jimenez-Ruiz *et al.*, 2007), permet de représenter les ontologies et leurs entités indépendamment du formalisme de nomenclature. Ce modèle est lié à un langage de requêtes appelé *OntoPath* qui extrait des fragments des ontologies larges avec une possibilité de spécifier le niveau de détail dans la hiérarchie des concepts. Les fragments extraits sont stockés dans une base sous forme de graphes. La généralité de ce modèle est due à sa capacité de reprendre les éléments communs dans les modèles d'ontologies et sa définition des relations abstraites entre ces entités. L'utilisation de ce modèle engendre une création de nouvelles classes et relations explicites à partir des axiomes de l'ontologie d'origine. Dans le contexte de gestion de ressources hétérogènes, ce modèle n'est pas applicable sur d'autres ressources à part les ontologies.

Une modélisation de l'aspect multilingue dans les ontologies à été proposée par (Montiel-Ponsoda *et al.*, 2008). Le modèle conçu est une association entre le métamodèle des ontologies et un modèle linguistique. Un modèle pour unifier la gestion de ressources linguistiques en contexte multilingue à été élaboré afin de centraliser la gestion des ressources linguistiques dans la plate-forme Intuition (Cailliau, 2006). Ce modèle se caractérise par son exploration de la structure des formes linguistiques. L'application de ce modèle permet de représenter des entités ontologiques et de les identifier par des unités lexicales en tenant compte des relations syntaxiques, sémantiques et multilingues. Cette représentation est centrée sur les ontologies, chaque représentation commence par l'entité conceptuelle dans une ontologie et décrit par la

suite l'unité lexicale correspondante. Ce modèle n'ayant pas de lien avec des entités ontologiques, ne permet pas de représenter des ressources linguistiques pures.

Dans le contexte de mise en correspondance de ressources linguistiques et ontologiques, (Suchanek *et al.*, 2007) ont proposé une approche d'intégration et de fusion de Wikipédia et WordNet pour étendre une ontologie (YAGO[9]). L'ontologie est extraite de ces deux ressources par l'ajout des nouveaux faits[10] extraits de Wikipédia sous forme d'individus et de classes issus des catégories conceptuelles de Wikipédia et de chaque "synset" de WordNet. Le processus d'extraction est basé sur l'identification d'un certain nombre de relations tels que ; *Type, SubClassOf, Means, Context*. Le modèle de l'ontologie résultante est défini en fonction de la ressource à utiliser et est dédié à la représentation de faits. Cette approche montre que la combinaison de plusieurs ressources permet de construire et d'enrichir de nouvelles ressources. Si on dispose d'un modèle générique capable de représenter les ressources, l'extraction et les combinaisons de leurs entités deviendraient des tâches moins complexes et moins coûteuses que celles proposées.

Pour l'intégration de ressources hétérogènes, (Vandenbussche *et al.*, 2009) ont proposé un métamodèle de représentation de terminologies et d'ontologies. Ce modèle propose un formalisme de représentation plus général que les formalismes existants et fournit de nouveaux constructeurs qui apportent une expressivité supplémentaire aux ressources terminologiques. Cette représentation est basée sur la différenciation des entités de ressources, et reste fidèle à la représentation de chaque modèle de ressource, mais en utilisant des entités abstraites communes. Une partie de ce modèle est centrée sur la terminologie et reprend des entités des modèles de thésaurus (Hall, 2001; Manh Hung, 2004).

Des outils comme OWLIM (Kiryakov *et al.*, 2005) de Ontotext[11] et ITM (Delaporte *et al.*, 2004) de Mondeca[12] permettent de regrouper des connaissances provenant des ontologies hétérogènes dans les formats RDF(S), OWL ou Topic Maps. Les modèles permettant de représenter ces connaissances et sur lesquels ces outils sont basés représentent les entités ontologiques ou terminologiques. OWLIM est un entrepôt sémantique, utilisant la plate-forme sesame[13] pour le stockage des triplets RDF. ITM est un outil basé sur les ontologies pour la classification du contenu et la gestion des taxonomies, thésaurus, lexiques, etc.

Dans la majorité des modèles que nous venons de décrire nous remarquons un attachement aux ressources, on ne peut pas représenter de nouvelles ressources différentes de celles pour lesquelles ces modèles ont été prévus. Le modèle à proposer doit avoir un niveau d'abstraction plus élevé afin de pouvoir représenter toutes les ressources avec des métadonnées communes. C'est le critère de notre modèle de ressources, il permet de représenter les ressources indépendamment de leurs types. L'originalité de

9. *Yet Another Great Ontology*.
10. Relatifs à l'ensemble des données existants dans une base de connaissances.
11. http://www.ontotext.com/.
12. http://www.mondeca.com/.
13. Un entrepôt RDF très populaires : http://www.openrdf.org/.

notre modèle est sa capacité à représenter les contenus des ressources avec des modèles multiples. Un modèle de contenu utilise l'ensemble des entités de la ressource qu'il décrit dans le but de rendre cette ressource utilisable par plusieurs processus de gestion de connaissances.

3. Approche de représentation des ressources

L'approche que nous proposons repose sur un modèle d'entrepôt de ressources constitué de trois niveaux : ressource, représentation et définition, présentés dans le tableau 1. Lorsqu'une nouvelle ressource est importée dans le système on en stocke une copie (niveau ressource). Si la ressource est très volumineuse (p.ex. Wikipedia) on peut ne garder qu'une référence pointant vers la ressource originale. Une représentation de la ressource est ensuite générée et stockée au niveau représentation. Cette représentation joue deux rôles : 1) décrire globalement la ressource par des métadonnées et 2) décrire le contenu de la ressource. Le niveau définition sert à définir les métadonnées et les modèles de représentation du contenu d'une ressource.

Niveau	Fonction
Définition	définition des métadonnées et des modèles de représentation du contenu
Représentation	représentation des ressources - métadonnées - représentation (abstraite) du contenu
Ressource	stockage de chaque ressource (dans son format d'origine)

Tableau 1. *Niveaux du modèle TOK*

Les niveaux définition et représentation forment une base de connaissances sur les ressources. Le niveau définition est assuré par la partie terminologique d'une ontologie (*TOK_Onto*) exprimée en logique de description. Elle comprend des descriptions de classes, propriétés et axiomes qui permettent la représentation des ressources.

La représentation d'une ressource est une instance (d'une sous-classe) de la classe TOK_Resource de l'ontologie, associée à des instances d'autres classes représentant les métadonnées et le contenu de la ressource.

3.1. *Métadonnées*

Les éléments de métadonnées sont utilisés pour décrire une ressource et faciliter son indexation dans l'entrepôt. Ces informations vont permettre d'effectuer des recherches avancées en tenant compte des critères spécifiques. Pour élaborer un modèle

ou un formalisme pivot capable de représenter des ressources hétérogènes nous avons suivi une démarche d'observation et de spécification des caractéristiques de ces ressources. Nos travaux précédents nous ont permis d'avoir la base théorique pour la description des ressources.

En effet, nos travaux sur (i) les annotations sémantiques et les documents techniques (Ghoula *et al.*, 2007), (ii) les bibliothèques numériques sémantiques (Falquet *et al.*, 2009) et (iii) l'indexation conceptuelle (Guyot *et al.*, 2008) et la désambiguïsation (Guyot *et al.*, 2005), nous ont permis de modéliser, manipuler et générer ces ressources. En se basant sur ces travaux et sur l'étude des ressources (cf. 2.1) nous avons pu déterminer les éléments des métadonnées et les caractéristiques internes et externes de chaque type de ces ressources :

– *le domaine* sert à déterminer le secteur d'activité humaine décrit par la ressource. Il existe des ressources qui couvrent plusieurs domaines tels que Wikipédia. Ce type de ressource fait référence à une liste de domaines. Nous pouvons représenter cette ressource comme une collection d'autres ressources décrivant chacune un domaine particulier. Pour la représentation des domaines nous pouvons faire référence à une ontologie ou une classification des domaines de l'activité humaine ;

– *le formalisme* sert à la représentation de connaissances associées à la ressource. Une ressource peut être représentée par plusieurs formalismes, dans ce cas elle est représentée par des représentations selon plusieurs modèles de contenu. Pour les formalismes on trouve les approches logiques (logique de description, logique des prédicats, etc.) et les approches non logiques (graphes conceptuels, réseaux sémantiques, etc.) ;

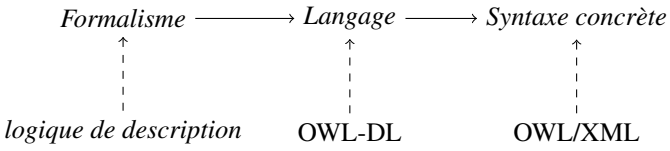


Figure 1. *Chaque formalisme est représenté par un langage ayant une syntaxe donnée*

– *le Langage* sert à déterminer le langage de représentation dans lequel le formalisme de la ressource est exprimé (cf. figure 1) ;

– *la catégorie* sert à déterminer le type de ressource. Une ressource peut être de type ontologie, terminologie, linguistique ou ressource d'indexation, d'alignement ou d'annotation. Ce critère permet de classer les ressources pour pouvoir les réutiliser et les associer à des formats bien déterminés ;

– *la langue* sert à indiquer la liste des langues de la ressource. Pour les ressources multilingues ce critère est défini par des valeurs multiples au niveau de la ressource et spécifié aussi chez ses entités ;

– *l'usage* sert à indiquer les ressources dont l'usage est bien défini. Par exemple, un corpus peut être utilisé pour l'apprentissage ou le test. Une ontologie peut être utilisée

pour l'annotation ou la recherche d'information. Un alignement peut être utilisé pour la fusion des ressources ou la réécriture des requêtes, etc. ;

– *la version* sert à spécifier une version de la ressource. Une ressource peut avoir plusieurs versions, ce critère assure une bonne exploitation des ressources afin de gérer la compatibilité, par exemple, si un alignement a été élaboré entre deux ontologies, cet alignement n'est plus forcément utile avec une nouvelle version d'une des deux ontologies ;

– *la source* sert à spécifier la personne ou l'organisme qui a conçu la ressource. L'origine de la ressource permet de savoir pour quelle raison et pour quelle utilisation une ressource a été créée ;

– *la taille* ou volume et le degré d'expressivité de la ressource. Ces critères permettent de nous donner une information sur l'importance de la ressource et son utilité pour des opérations particulières.

À titre d'exemple, un ontologue veut enrichir une ontologie dans le domaine de l'aéronautique. Cette ontologie est sous la forme d'une hiérarchie de concepts. Il veut ajouter des définitions dans deux langues ; anglais et français, aux concepts de cette ontologie. Il veut également raffiner la classification par l'ajout de nouvelles classes. Pour réaliser cette tâche, il lui faut des ressources externes telles que des glossaires, terminologies ou dictionnaires bilingues dans les langues en question.

Afin d'avoir accès aux ressources pertinentes le concepteur peut interroger l'ontologie *TOK_Onto* pour chercher toutes les ressources décrivant le domaine de l'aéronautique, ayant pour langues le français et/ou l'anglais. Comme résultat de sa requête, le système de recherche basé sur *TOK_Onto* retourne un certain nombre de ressources, par exemple, un corpus parallèle anglais-français de textes concernant l'aéronautique, des articles de Wikipedia dans ce même domaine, classés par catégorie, et un dictionnaire des synonymes en anglais de l'aéronautique, etc.

3.2. *Contenu*

Étant donné la diversité des ressources de connaissances terminologiques, ontologiques et linguistiques et la variété des formalismes et langages de représentation des connaissances, il serait vain de tenter de définir un modèle unifié capable de représenter le contenu de n'importe quelle ressource. L'approche que nous proposons consiste plutôt à définir un ensemble de modèles abstraits de contenus et à représenter le contenu d'une ressource à l'aide d'un ou de plusieurs modèles, en fonction des besoins. Lors de l'importation dans l'entrepôt on pourra choisir les modèles de représentation nécessaires à l'exécution des tâches pour lesquelles la ressource est requise. Ces représentations ne préservent en général pas toute la connaissance contenue dans la ressource mais en extraient les parties nécessaires à un traitement donné.

Un exemple typique du besoin de modèles simplifiés est l'alignement d'ontologies. La majorité des algorithmes d'alignement actuels peuvent aligner des ontologies en OWL mais ils n'utilisent pas toute la sémantique exprimée par ce formalisme. Ils

sont souvent basés sur les étiquettes textuelles attachées à chaque classe dans la structure de l'ontologie. La structure est généralement un graphe représentant la hiérarchie des classes et les propriétés qui font le lien entre deux classes (e.g. Il y a un lien d'étiquette P entre les classes C_1 et C_2, s'il existe un axiome de la forme $C_1 \sqsubseteq P$ $\mathbf{only/some}$ C_2). Dans ce cas, il est plus approprié de représenter une ontologie en OWL par un graphe de structure au lieu d'utiliser le modèle complet de la logique de description OWL. Les algorithmes d'alignement vont être plus faciles à écrire et ils vont permettre d'aligner plusieurs types d'ontologies pouvant être représentées par un graphe étiqueté.

Au niveau de la base de connaissances, nous créons une instance représentant la ressource et des instances représentant ses entités. Selon les traitements que nous avons besoin d'appliquer, cette représentation peut utiliser un certain type de modèle.

En outre, la même ressource peut être impliquée dans des processus qui supportent chacun un format spécifique. Ainsi, grâce aux représentations multiples une même ressource peut être utilisée dans plusieurs processus car son contenu est représenté par plusieurs modèles. Par exemple, un algorithme d'alignement ne peut accepter des ontologies au format OWL, tandis qu'un autre algorithme nécessite des ontologies dans un format de type WordNet.

3.3. *Traitement des ressources*

La gestion et le traitement des ressources dans l'entrepôt consistent essentiellement à importer des ressources, puis à appliquer des processus sur leurs représentations pour générer de nouvelles ressources.

Si l'on revient à l'exemple de la section 3.1 concernant l'ontologie aéronautique, le processus d'extension ou d'enrichissement pourrait être décrit comme suit :

– recherche de glossaires, terminologies ou dictionnaires de termes en anglais et français relatifs au domaine de l'aéronautique (par sélection sur les métadonnées)

– application d'opération de transformation (*mapping*) pour obtenir des représentations de leur contenu sous forme d'ontologies lexicales (chaque terme donne lieu à un concept avec sa définition sous forme d'annotation) ;

– application d'opérations d'alignement d'ontologies pour faire correspondre les concepts de ces ontologie avec ceux de l'ontologie à étendre ;

– fusion des ontologies ainsi alignées pour produire une nouvelle ontologie enrichie ;

– exportation de cette ontologie dans le format désiré.

Chaque processus de traitement de ressources peut être décrit comme une séquence d'opérations élémentaires sur les représentations de ressources. Ces opérations peuvent être de différents types : transformations de représentations (pour passer d'un modèle à un autre), sélection d'un sous-ensemble des entités d'une représentation, fu-

sion, alignement, composition d'alignement, annotation, etc. Mis à part l'importation et l'exportation toutes ces opérations agissent au niveau représentation et non pas directement sur les ressources elles-mêmes. Chaque opération est caractérisée par le ou les modèles de représentation auxquels elle s'applique et les algorithmes ou heuristiques utilisés.

La modélisation des processus et opérations a deux objectifs principaux : 1) trouver les opérations applicables à une ressource ou inversement trouver les ressources sur lesquelles on pourrait appliquer une opération ; 2) mémoriser les processus de création de ressources dérivées, ce qui permettra, entre autres, de ré-exécuter les processus sur de nouvelles versions des ressources.

La figure 2 montre comment les différents nivaux du modèle TOK sont impliqués dans le traitement des ressources tels que l'importation, la recherche et la génération, etc.

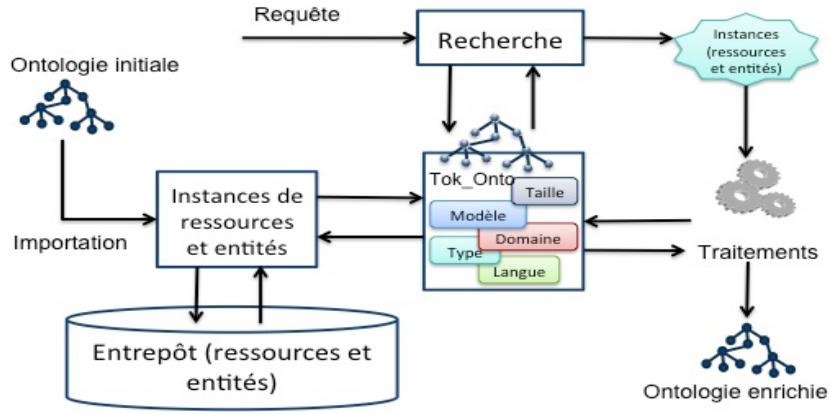


Figure 2. *Exemple d'un scénario d'utilisation de TOK_Onto*

4. Ontologie de ressources

4.1. *Modélisation des métadonnées des ressources*

L'étude des ressources selon les critères identifiés précédemment nous a permis de construire une classification des ressources et d'élaborer la première couche de notre ontologie générale *TOK_Onto*[14].

La classe '*TOK_Resource*' permet de modéliser les ressources, elle comporte plusieurs sous-classes en fonction du type des ressources étudiées. Les critères communs

14. Disponible sur internet à http://cui.unige.ch/isi/onto/tok/OWL_Doc/.

de ces ressources sont représentés dans cette classe et les critères spécifiques font l'objet d'une description dans des sous-classes.

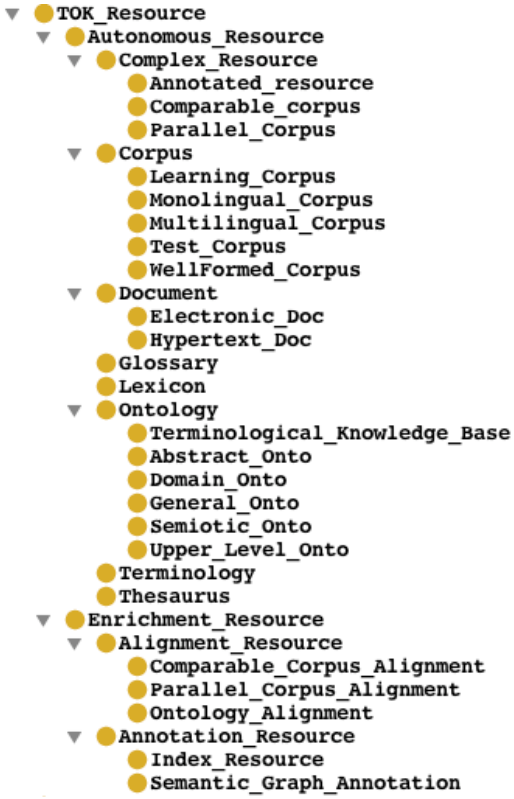


Figure 3. *Vue partielle de la classification des ressources de connaissances dans TOK_Onto*

Cette ontologie décrit l'ensemble des ressources de connaissances hétérogènes. *TOK_Onto*, a été développée en format *OWL* avec le degré d'expressivité *SRIQ(D)* en utilisant l'éditeur d'ontologies Protégé. *TOK_Onto* contient 195 concepts (nommés et non nommés), 120 propriétés, 450 axiomes, et 2 000 annotations.

La figure 4 décrit la classe '*TOK_Resource*' et ses liens avec les autres classes à travers des propriétés. Une ressource peut contenir, importer ou être alignée avec d'autre ressources. Les entités d'une ressource sont modélisées par la classe '*TOK_Entity*', ces entités peuvent avoir des relations entre elles de type association, alignement, traduction, description. La relation entre une classe et une propriété dans une ontologie est de type '*source -> destination*'.

Chaque élément (concept, propriété, axiome, individu, terme, etc.) est traité comme une entité de connaissances ontologiques, terminologiques ou linguistiques

(*TOK_Entity*) et lié à une ressource à travers la relation "*uses*" entre le modèle de la ressource et l'entité.

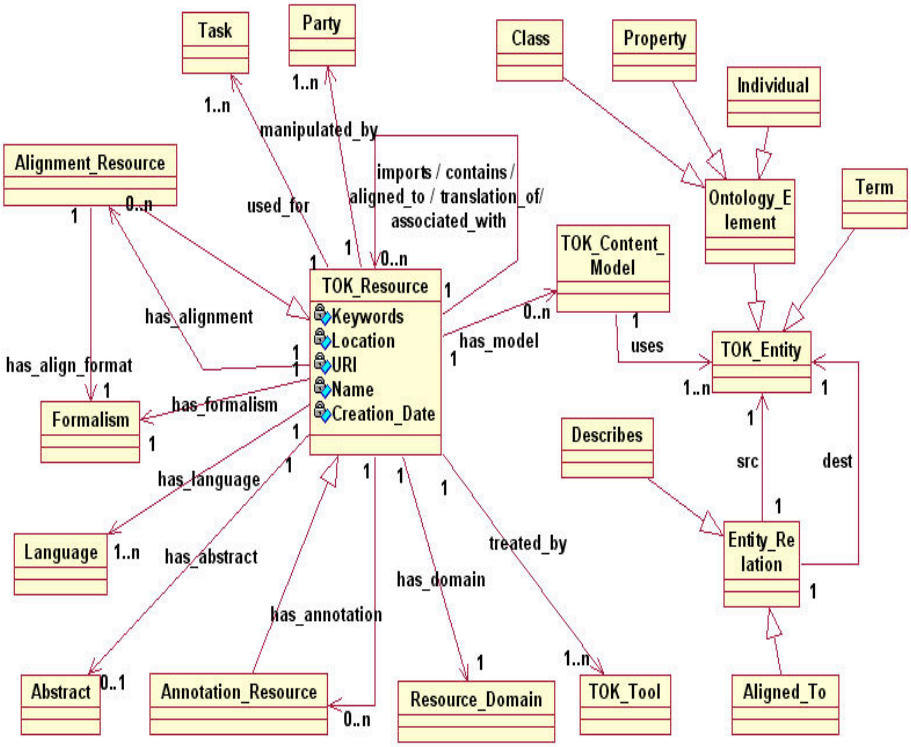


Figure 4. *Modèle de ressources TOK*

Les relations entre les entités sont représentées par des instances de "*Entity_Relation*" qui permettent de catégoriser ces relations par type (relation entre concepts, relation entre termes, relation entre concepts et propriétés, relation entre concepts et individus, relation entre termes et concepts, etc.). La relation entre les concepts peut avoir plusieurs types comme l'équivalence, la subsomption, l'intersection, la disjonction, etc.

4.2. *Modélisation multiple du contenu des ressources*

Un modèle de contenu est composé d'un ensemble de classes, correspondant aux diverses entités du modèle, de propriétés et d'un ensemble d'axiomes définissant les relations entre ces classes. La représentation du contenu d'une ressource est composée d'instances des classes satisfaisant les axiomes du modèle. Le modèle de contenu

joue le rôle du lien entre la ressource et ses entités, puisqu'une représentation d'une ressource par un modèle spécifique utilise une partie ou la totalité de ses entités.

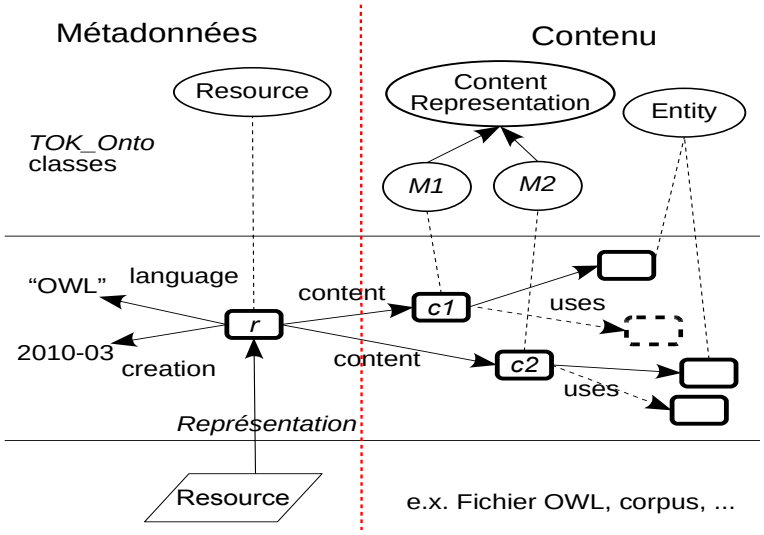


Figure 5. *Représentation d'une ressource avec ses métadonnées et deux représentations de son contenu (*c1 *et* c2*). Les éléments de représentation sont des instances des classes de TOK_Onto*

Nous avons décrit un ensemble de modèles de représentation du contenu à l'aide des axiomes et concepts dans notre ontologie de ressources. *TOK_Onto* permet de décrire :

– les modèles (figure 6) relatifs à la représentation de la ressource selon la démarche décrite dans la section (3.2) ;

Figure 6. *La classification des modèles de représentation du contenu des ressources dans l'ontologie TOK_Onto*

– les détails des entités appartenant aux ressources (figure 7) et leurs particularités. Cette description est modélisée par la classe *TOK_Entity*. Les types d'entités sont décrits comme des sous-classes de *TOK_Entity*.

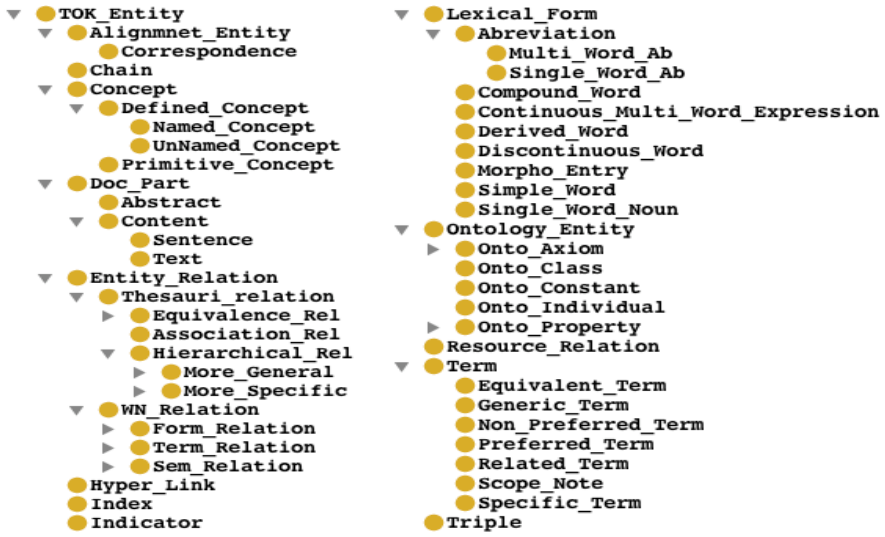


Figure 7. *La classification des entités dans l'ontologie TOK_Onto*

Exemple : représentation de la ressource WordNet à l'aide de l'ontologie *TOK_Onto*.

WordNet (Fellbaum, 1998) est un ensemble de formes lexicales ayant des liens entre elles. La composante atomique sur laquelle repose le système entier est le synset (synonym set), un groupe de mots interchangeables, dénotant un sens ou un usage particulier. Nous avons construit un modèle de représentation du contenu de la ressource WordNet et nous l'avons appelé *WN_Like* (WordNet like). Ce modèle est composé des classes Concept, Term, LexicalForm, Sentence, Part_of_speech et des classes associatives Sem_Relation, Term_Relation et Form_Relation (entre autres).

Dans ce type de modèle on part des entités conceptuelles, liées entre elles par des relations sémantiques, vers les entités terminologiques reliées avec les concepts par des relations de description ou étiquetage. Les termes sont décrits par des formes lexicales qui sont des entités linguistiques permettant de désigner un terme dans une langue donnée.

La figure 8 présente le modèle de cette ressource dont les correspondances avec ses entités initiales.

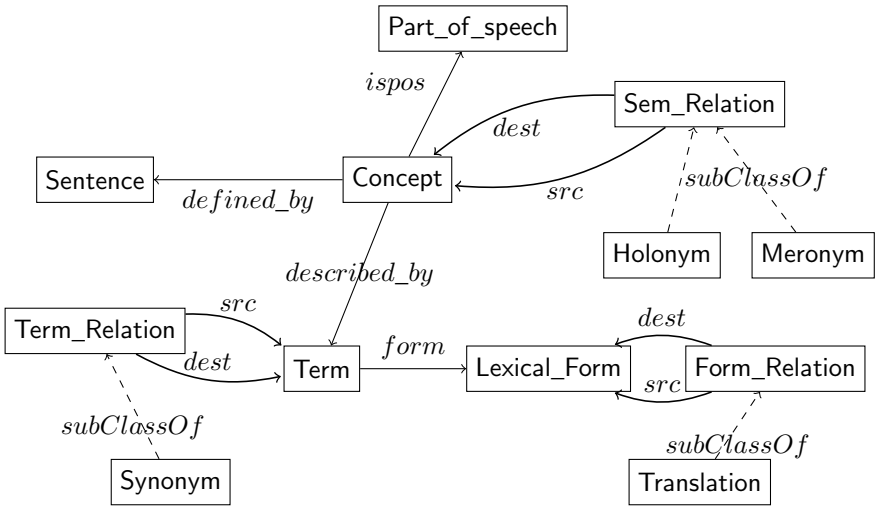


Figure 8. *Partie de la description du modèle WordNet_Like*

4.3. *Représentation des opérateurs*

Pour compléter la définition d'un modèle de représentation du contenu, il faut lui associer un ensemble d'opérateurs. Les opérateurs permettent d'utiliser une représentation du contenu pour effectuer un traitement sur les entités qu'elle utilise. Par exemple, à chaque modèle de représentation du contenu on associe un opérateur ou plusieurs opérateurs d'importation et d'exportation qui permettent de représenter une ressource exclusivement à l'aide des éléments de ce modèle. Nous avons conçu un modèle d'opérateurs élémentaires et complexes (cf. figure 9) en nous basant sur nos travaux antérieurs sur des opérateurs pour la gestion des ontologies (Falquet *et al.*, 2008).

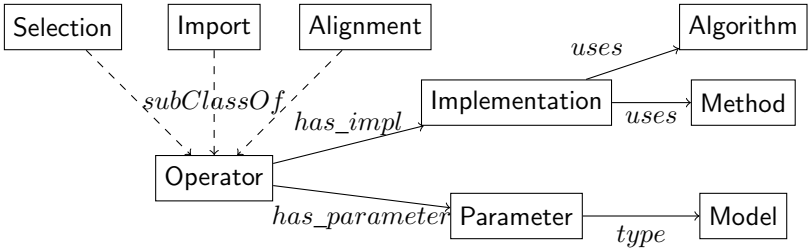


Figure 9. *Partie de la description du modèle d'opérateurs*

5. Stockage des représentations des ressources

Nous avons construit un espace de stockage permettant à la fois de représenter des ressources, leurs entités et les relations entre elles. Cet espace de stockage est la base du reste du travail qui sera la détection des alignements entre concepts ou entités, la traduction et l'alignement multilingue des unités lexicales, ainsi que d'autres opérations. La possibilité de stocker directement des triplets RDF dans l'entrepôt a été explorée grâce au modèle RDF offert par Oracle[15]. En réalisant une expérimentation avec ce mécanisme de stockage en triplet nous avons réalisé une expansion importante de la taille de données ce qui nécessite un espace mémoire plus volumineux.

5.1. *Modèle de stockage*

Nous avons utilisé la structure de stockage correspondant au modèle des ressources car elle est équivalente au modèle RDF. Cette structure est sous la forme d'un modèle nœud-lien qui ressemble à des triplets RDF[16]. Le modèle nœud-lien se caractérise par (i) des éléments qui sont sous formes de nœuds (entités) tels que les concepts, les termes, les formes lexicales, les phrases, etc., (ii) des liens entre les nœuds tels que les relations hiérarchiques, les relations sémantiques et d'autres types de relations provenant des modèles de représentation des ressources, (iii) les sources de données et (iv) les modèles de représentation de ces sources. Ce modèle a servi pour la définition du schéma de la base de données de l'entrepôt de ressources.

L'utilisation des bases de données se justifie par la taille importante des ressources à traiter. Nous voulons exploiter les performances de ce type de stockage avec son langage de requêtes simple et efficace. Les instances permettent de faire le lien avec les ressources physiques dans la base de données. Chaque élément d'une ressource est associé à une classe de *TOK_Onto*.

La table NODES est décrite de la façon suivante :
NODES (
IDN : *identifiant unique de l'entité*,
IDN_EXTERN : *identifiant original de l'entité dans la ressource*,
KIND : *type de l'entité (Concept, Terme, . . . ,*
LANG : *langue de l'entité si défini*,
SOURCE : *le ressource d'origine de l'entité (clé étrangère vers la table SOURCES)*,
STATUS : *statut de l'entité dans le version de la ressource (valable ou invalide)*,
LIB : *label de l'entité si défini*,
EXTENSION : *référence vers la description de l'entité*,
COM : *commentaires et annotations*,
EXT_TYPE : *type de l'extension de référence (fichier, texte, . . .)*
)

15. http://www.oracle.com/technology/tech/semantic_technologies/index.html
16. Nous sommes entrain de migrer vers les entrepôts RDF sous Jena ou Sesame

La table LINKS est décrite de la manière suivante :
LINKS (
REL : *identifiant de la relation utilisée (clé étrangère vers la table RELATIONS*,
IDNFROM : *identifiant unique de l'entité source (clé étrangère vers la table NODES)*,
IDNTO : *identifiant unique de l'entité cible (clé étrangère vers la table NODES)*,
RELINV : *identifiant unique de la relation inverse source (clé étrangère vers la table RELATIONS)*,
SOURCE : *identifiant unique de la ressource en question (clé étrangère vers la table SOURCES)*,
SEQ : *numéro de séquence de la relation (exemple : n° de synset pour WordNet)*,
CONFIDENCE : *degré de confidence du lien entre les 2 entités*,
STATUS : *statut du lien entre les 2 entités dans le version de la ressource (valable ou invalide)*,
COM : *commentaires et annotations*,
)

5.2. *Importation des ressources*

Notre technique d'importation des ressources dans la structure de stockage permet de formaliser et stocker les ressources *TOK* dans un seul entrepôt. Cette méthodologie est décrite à travers un processus de chargement de ressources composé de quatre modules.

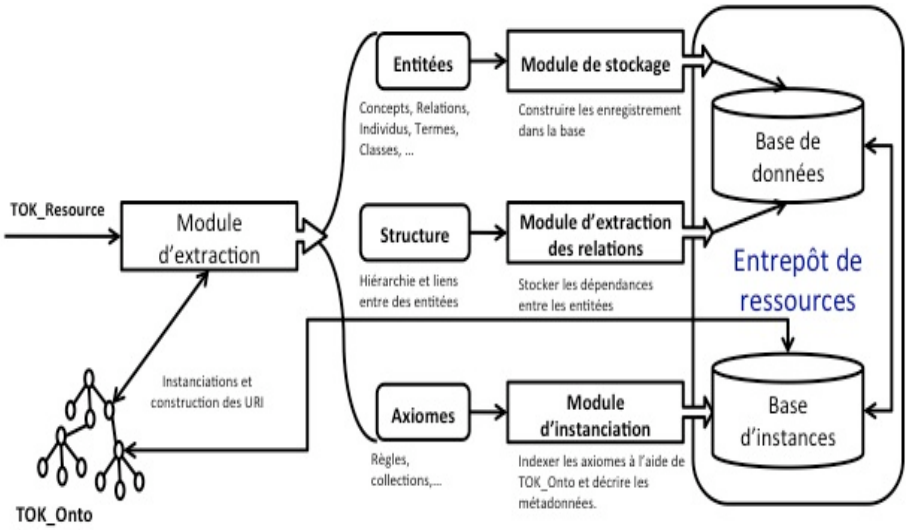


Figure 10. *Description de processus de chargement de ressources dans l'espace de stockage*

1) Un module d'extraction, basé sur une interaction entre trois niveaux de représentation de connaissances :

– identifier les entités de chaque ressource en utilisant l'ontologie *TOK_Onto*, chaque nouvelle entité est considérée comme instance d'un concept de l'ontologie ;

– extraire les relations hiérarchiques et de structure de la ressource. Ceci étant utile pour l'exportation de la ressource en question et pour la classification des entités ;

– extraire les axiomes et les représenter à l'aide de l'ontologie *TOK_Onto* pour garder la sémantique des concepts de la ressource.

2) Un module de stockage et d'indexation des entités, qui permet de construire de nouveaux enregistrements dans la base de données pour stocker les entités extraites (instances de *TOK_Entity*).

3) Un module d'extraction des relations, assurant l'inférence des dépendances entre les entités stockées. Ces dépendances sont généralement des subsomptions, des relations hiérarchiques simples, ou des relations complexes entre concepts et propriétés.

4) Un module d'indexation d'axiomes permettant de les décrire et d'identifier les entités utilisées dans chaque axiome.

Le modèle TOK est en cours d'utilisation. Nous ajoutons progressivement de nouvelles ressources. Nous avons ajouté AGROVOC[17] en 17 langues, WordNet en anglais, allemand, Catalan, Espagnol, etc., UNL[18] en Français, Arabe, Japonais, CityGML[19], URBAMET[20], etc.

Nous avons stocké ces ressources et nous avons pu générer un ensemble d'entités conceptuelles et terminologiques. Ces entités ont été reliées entre elles par des relations d'indexation de catégorie "*Term_Concept*" et des relations de traduction et de hiérarchie de catégorie "*Term_Term*". Ces liens ont été établis par l'implémentation du module d'extraction des relations. L'algorithme d'extraction permet de repérer les relations entre entités, nous allons l'étendre pour détecter des relations complexes et des alignements multilingues.

La figure 11 montre un extrait des ressources importées dans l'entrepôt. La colonne de gauche est une liste des modèles utilisés. La colonne de droite montre la

17. AGROVOC est un vocabulaire multilingue structuré, développé par la FAO, couvrant la terminologie de tous les domaines ayant trait à l'agriculture, à la pêche, à l'alimentation et aux domaines connexes (l'environnement, par exemple).
18. Universal Networking Language, est un langage artificiel qui peut être utilisé comme langage pivot pour des systèmes de traduction automatique ou comme un langage de représentation des connaissances dans les applications de recherche d'information.
19. CityGML est un modèle d'information commun pour la représentation des objets 3D en milieu urbain.
20. URBAMET est une base de données bibliographiques française sur l'urbanisme, l'aménagement du territoire, les villes, l'habitat et le logement, l'architecture, les équipements collectifs, les transports, les collectivités locales etc.

liste des ressources qui utilisent un modèle sélectionné, dans ce cas c'est le modèle *WordNet_Like*.

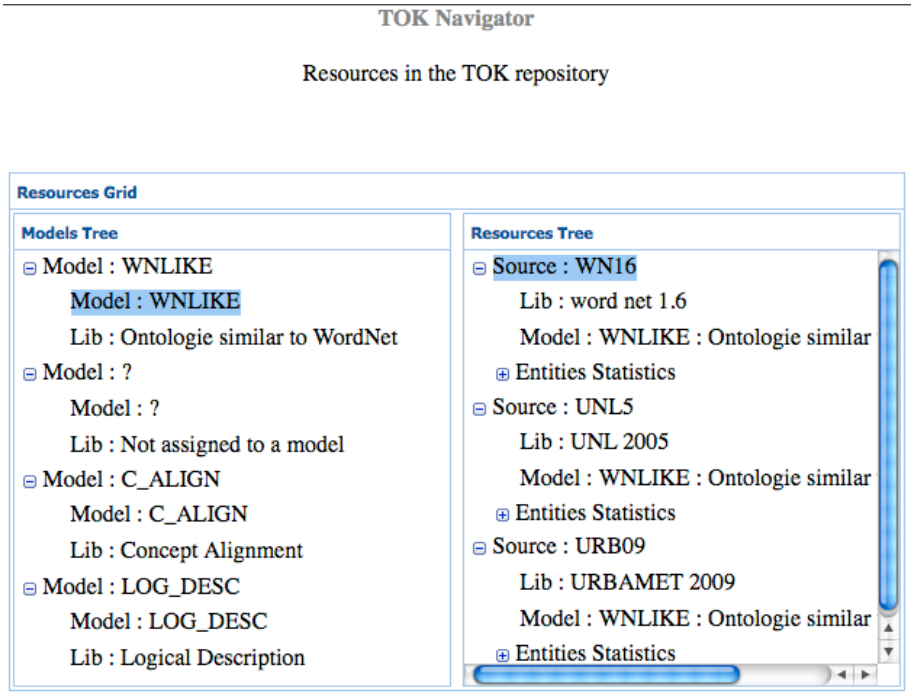


Figure 11. *Les ressources importées dans l'entrepôt classées par modèle*

5.3. *Usage de l'entrepôt*

Actuellement, notre entrepôt compte environ 950 000 formes lexicales différentes dans 24 langues, 173 000 concepts d'ontologies et 335 000 phrases provenant tous de 13 ressources différentes. Pour importer ces ressources, nous avons développé plusieurs outils pour faire la correspondance entre les formats et les langages de représentation des ressources (OWL/XML, WordNet, AGROVOC, XML Schema, les pages HTML liées, . . .) pour supporter des modèles tels que *WordNet_Like*, *mémoire de traduction*, . . .

Exemple : génération d'une ontologie lexicale à partir de Wikipédia

Pour importer des éléments de la ressource Wikipédia nous avons utilisé un modèle de représentation simplifié (*Wikipédia_Like*) et nous l'avons enrichi par le biais du processus d'extraction de termes, des descriptions et des liens dans les articles.

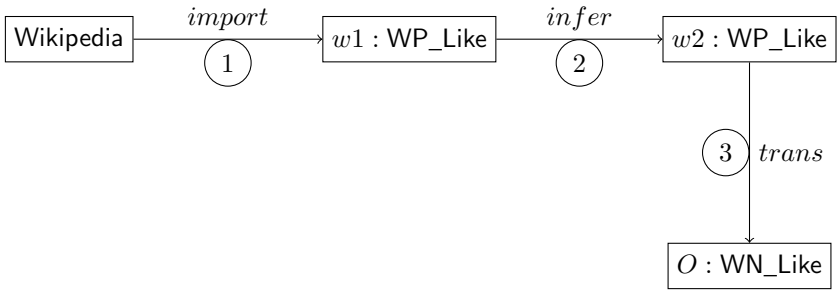


Figure 12. *Importation des articles de la ressource wikipédia dans l'entrepôt*

Wikipédia est une ressource de type corpus structurée et organisée par catégories. C'est une collection d'articles ayant la forme d'un document hypertexte, ce document contenant plusieurs sections relatives à la définition, traduction, classification et désambiguïsation d'un terme.

1) L'importation d'un ensemble d'éléments de la ressource Wikipédia se fait par l'identification des documents hypertextes, leur type et leurs métadonnées dans le modèle WP_Like :
Page de Wiki —> Hypertext_Doc (*classe définie dans TOK_Onto*) ;
Suffixe de l'URL —> name (*propriété définie dans TOK_Onto*) ;
Liens vers des pages dans d'autres langues —> Translation_Link ;
Contenu HTML —> Doc_Part ;

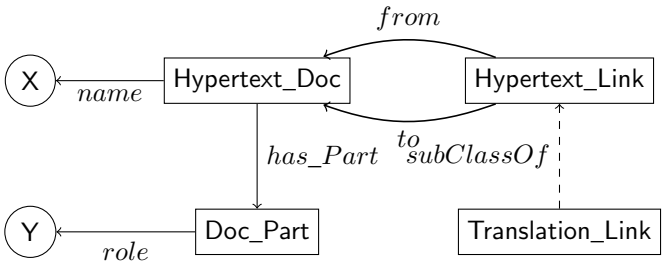


Figure 13. *Partie de la description du modèle Wikipédia_Like*

2) L'identification des éléments d'annotation et de traduction se fait pas une heuristique qui attribue le rôle 'Definition' (définition) à la partie du document qui permet de décrire le concept représenté par la page. Cette heuristique permet de parcourir toutes les pages wikipedia, figurant comme traduction de la page de référence, et d'extraire les formes lexicales avec leurs langues correspondantes comme des candidats de traduction pour la forme lexicale source.

3) Le changement de modèle se fait par la transformation de la représentation de la ressource importée du modèle WP_Like vers le modèle WN_Like. Le mapping entre les deux modèles se fait sur la base des correspondances suivantes :

Hypertext_Doc	→	Concept	*l'URL du document devient* un concept
name	→	Lexical_Form	*avec spécification de la langue*
Translation_Link	→	form	*avec la construction du* Term *qui fait le lien entre le concept et la forme lexicale*
. . . ;			

La figure 14 représente une interface permettant de parcourir les entités importées à partir de la ressource Wikipédia dans l'espace de stockage *TOK_Base*.

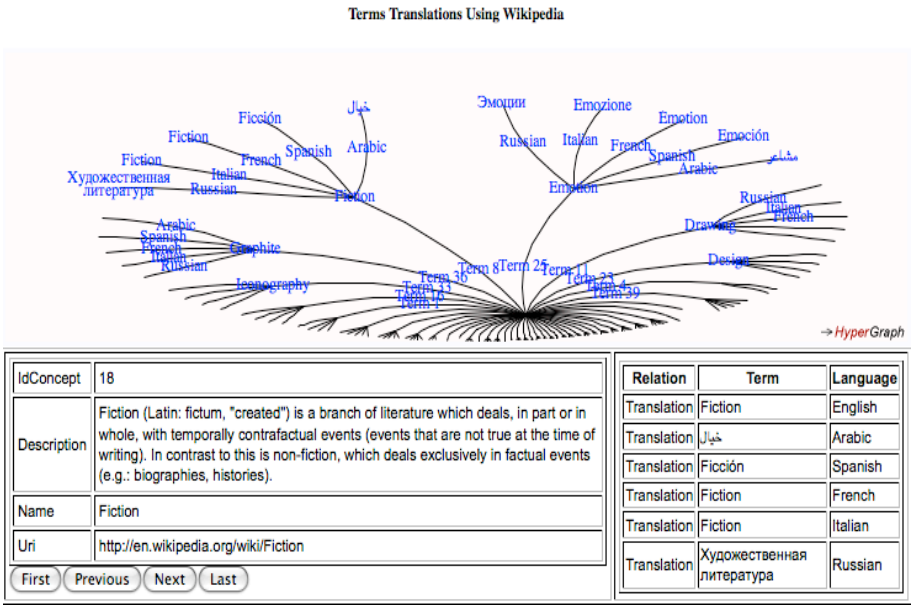


Figure 14. *Parcours des termes et concepts extraits à partir des pages Wikipédia*

Les opérations d'enrichissement des ressources permettent de générer de nouveaux alignements ou annotations sur des ressources existantes. Elles sont généralement basées sur des algorithmes spécifiques (ou des heuristiques) et utilisent des ressources auxiliaires.

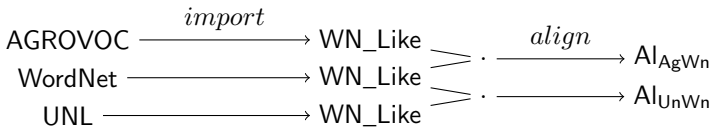
Une opération d'alignement prend comme entrée deux contenus de ressources, représentés par le même modèle (M_1) et produit un contenu de type alignement dans le même modèle que nous notons $\mathsf{AIn} < M_1 >$

Exemple : enrichissement de la ressource WordNet en anglais par des formes lexicales d'autres langues.

Ceci est un exemple plus spécifique d'une opération d'alignement de ressources représentées par le modèleWN_Like. Nous avons conçu un algorithme d'alignement simple que nous avons appelé AL_HS permettant de collecter les alignements évidents par similarité de parent sur le modèle *WordNet_Like*. Le modèle de cet opérateur est appelé ALG-ISI1. La signature de cet opérateur est :

$align_{AL\_HS} : (AL\_SP : WN\_Like, WN\_Like \rightarrow \mathsf{AIn} < WN\_Like >)$

$align_{AL\_HS} :\rightarrow implem_{AL\_HS} < ALG - ISI1 >$ (*implémentation*)



Nous avons aligné la version en Anglais de WordNet avec d'autres ressources, comme AGROVOC, URBAMET, UNL, qui ont des formes lexicales dans plusieurs langues. Cela nous a permis l'importation de ces formes lexicales dans la version en Anglais de WordNet et de les associer aux concepts correspondants, obtenant ainsi un enrichissement WordNet.

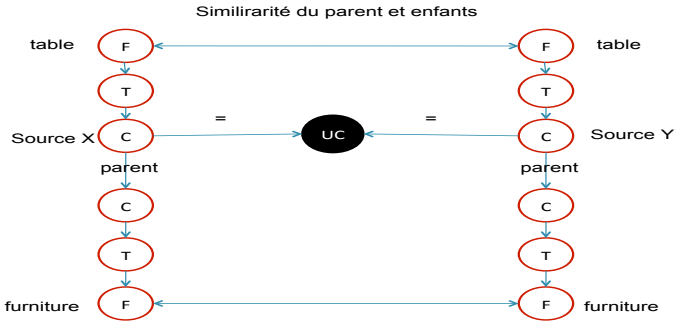


Figure 15. *Alignement par similarité dans le modèle TOK*

Des alignements et des correspondances s'effectuent par le biais de la similarité des formes. Ces correspondances déduites (819 alignements), permettent la désambiguïsation des termes. Dans cet exemple (figure 16), le concept numéro 161185 est décrit par le terme "table" en Anglais qui est un terme ambigu, son alignement avec le

terme "mesa" en Espagnol (non ambigu), permet de déduire que le concept "161185" appartenant à la catégorie des meubles.

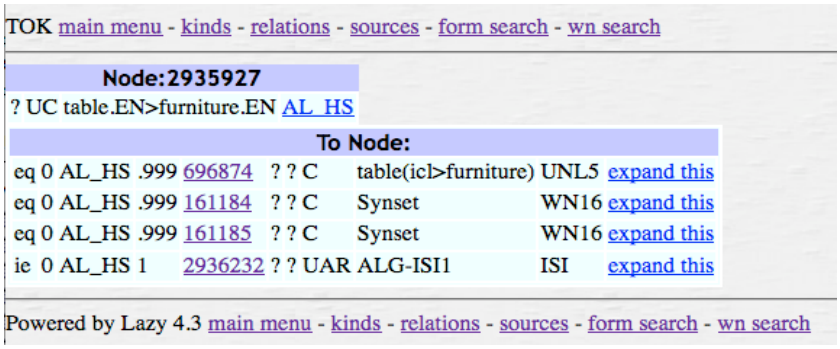


Figure 16. *Exemple d'entités alignées dans deux ressources*

L'importation du contenu d'une ressource ne permet pas nécessairement de préserver tout son contenu. En particulier, si le modèle de représentation du contenu est moins expressif que le modèle original de la ressource, il est évident qu'au cours du processus d'importation certaines informationss vont être perdues. À titre d'exemple, en important des ontologies en OWL vers le modèle *WN_Like* nous avons perdu la sémantique des concepts exprimée par les axiomes. Le problème de perte d'information n'est pas considéré comme un handicape puisque nous gardons une version originale de la ressource.

6. Conclusion

Notre travail est centré sur les ressources de connaissances terminologiques, ontologiques et linguistiques. Nous avons proposé un modèle de représentation de ces ressources et nous avons expliqué sa construction et son usage. Ce modèle intervient dans (i) le traitement d'un large spectre de ressources représentées dans différents formalismes ; (ii) la définition d'un processus de transformation et de sauvegarde des ressources ; (iii) la perspective de définir un ensemble d'opérateurs pour la traitement sémantique des ressources et la détection des alignements.

L'objectif principal de notre approche est de pouvoir générer de nouvelles ressources à partir de la composition des ressources existantes dans l'entrepôt et instanciées dans l'ontologie. Ainsi, l'enrichissement des connaissances dans l'entrepôt s'effectue à chaque utilisation. En se basant sur l'espace de stockage élaboré, les traitements sur les connaissances devront permettre l'utilisation, la génération, l'intégration de connaissances et la production de nouvelles ressources dans différents formalismes. Cette boîte à outils est basée sur l'entrepôt de données *TOK_Base*, l'ontologie *TOK_Onto* et l'implémentation de l'ensemble des opérateurs. L'entrepôt a été implé-

menté à l'aide des technologies de bases de données relationnelles et d'applications Java. Il possède une interface web pour son utilisation interactive.

Comme nous l'avons décrit au début, nous voulons par la suite modéliser l'usage des ressources. Ces usages vont permettre d'associer à une tâche de gestion de connaissances un type de ressource bien déterminé. Les tâches sont à définir sous forme d'opérateurs abstraits. Ces opérateurs devront permettre à l'utilisateur de générer la connaissance qui répond à son besoin en toute simplicité et transparence. De ce fait, il faut concevoir par la suite un métalangage des opérateurs. La définition de ces opérateurs dépend des traitements sur les ressources collectées. Ainsi, faut-t-il connaître les besoins des utilisateurs potentiels d'un tel système de gestion de connaissances. Chaque opérateur peut être implémenté de plusieurs façons en fonction de la nature des ressources utilisées. Ces implémentations doivent construire un Framework d'outils que l'utilisateur peut parcourir afin de sélectionner un opérateur de son choix.

L'application de notre approche crée de nouvelles connaissances et fournit plusieurs composantes pour l'entrepôt de ressources interagissant toute avec *TOK_Onto* : (1) le dictionnaire de modèles de représentation du contenu, modélisant des formalismes existants ou fournissant de nouvelles représentations des ressources ; (2) l'entrepôt de ressources, utilisant l'ontologie *TOK_Onto* avec les modèles de représentations ; (3) le dictionnaire des opérations, qui contient des opérations simples ou complexes pour la manipulation des ressources en fonction du modèle de représentation et (4) le dictionnaire de processus, basé sur la modélisation des processus de gestion de connaissances. Cette modélisation prend en compte la possibilité de combiner des opérateurs fournis par le dictionnaire des opérations et des modèles existants dans le dictionnaire des modèles et des instances des ressources dans l'entrepôt. Dans ce dictionnaire on peut modéliser les processus d'alignement, d'annotation sémantique, l'indexation conceptuelle, la traduction, etc.

Une prochaine étape du travail consiste à définir des règles et des axiomes permettant d'associer à chaque tâche l'ensemble des ressources à utiliser, la représentation correspondante, les opérateurs disponibles ou la combinaison des opérateurs permettant d'effectuer cette tâche. Pour assurer la réalisation de cette perspective nous devons : (i) définir un modèle pour chaque tâche de traitement de connaissances utilisant les ressources *TOK*, ces modèles de tâches seront le résultat d'une réflexion sur un ensemble de cas d'utilisation ; (ii) étudier les besoins et définir des règles permettant de rattacher à chaque tâche l'ensemble des ressources candidates pour être utiles à son accomplissement ; (iii) définir une algèbre ou un langage de composition d'opérateurs de sélection, génération, intégration, projection ou alignement afin de construire un nouvel opérateur relatif à la tâche demandée par l'utilisateur ; (iv) définir et appliquer un ensemble d'heuristiques pour la déduction des correspondances pour construire des alignements entre les ressources lors de l'exécution d'une tâche quelconque.

7. Bibliographie

Bouquet P., Giunchiglia F., van Harmelen F., Serafini L., Stuckenschmidt H., « COWL : Contextualizing Ontologies », *Second International Semantic Web Conference*, vol. 2870 of *Lecture Notes in Computer Science*, Springer Verlag, p. 164-179, 2003.

Cailliau F., « Un modèle pour unifier la gestion de ressources linguistiques en contexte multilingue », *in* P. Mertens (ed.), *Verbum ex machina : actes de la 13e Conférence sur le Traitement Automatique des Langues Naturelles*, Presses univ. de Louvain, p. 454-461, 2006.

Delaporte G., Amardeilh F., « ITM et intelligence ééconomique : MONDECA = ITM software and competitive intelligence : MONDECA », , vol. 2, p. 365-366, 2004.

Ding Y., Fensel D., « Ontology Library Systems : The key to successful Ontology Re-use », *Stanford University 2001 ; S*, p. 93-112, 2001.

Euzenat J., Shvaiko P., *Ontology matching*, Springer-Verlag, Heidelberg (DE), 2007.

Falquet G., Jiang C.-L. M., Guyot J., « Un modèle et une algèbre pour les systèmes de gestion d'ontologies », *EGC*, p. 697-702, 2008.

Falquet G., Nerima L., Ziswiler J.-C., « Hyperbooks », *in* S. R. Kruk, B. McDaniel (eds), *Semantic Digital Libraries*, Springer, p. 179-196, 2009.

Fellbaum C. (ed.), *WordNet : An Electronic Lexical Database*, Language, Speech, and Communication, MIT Press, Cambridge, Mass., 1998.

Foskett D. J., « Thesaurus », *in* A. Kent, H. Lancour, J. E. Daily (eds), *Encyclopedia of Library and Information Science*, vol. 30, Marcel Dekker, New York, p. 416-462, 1980.

Ghoula N., Khelif K., Dieng-Kuntz R., « Supporting Patent Mining by using Ontology-based Semantic Annotations », *Web Intelligence*, IEEE Computer Society, p. 435-438, 2007.

Gruber T. R., « Toward principles for the design of ontologies used for knowledge sharing ? », *Int. J. Hum.-Comput. Stud.*, vol. 43, n° 5-6, p. 907-928, 1995.

Guyot J., Falquet G., Radhouani S., Benzineb K., « Analysis of Word Sense Disambiguation-Based Information Retrieval », *in* C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Peñas, V. Petras (eds), *CLEF*, vol. 5706 of *Lecture Notes in Computer Science*, Springer, p. 146-154, 2008.

Guyot J., Radhouani S., Falquet G., « Conceptual Indexing for Multilingual Information Retrieval », *in* C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke (eds), *CLEF*, vol. 4022 of *Lecture Notes in Computer Science*, Springer, p. 102-112, 2005.

Hall M., « CALL Thesaurus Ontology in DAML », 2001.

Jimenez-Ruiz E., Llavori R. B., Nebot V., Sanz I., « OntoPath : A Language for Retrieving Ontology Fragments. », *in* R. Meersman, Z. Tari (eds), *OTM Conferences (1)*, vol. 4803 of *Lecture Notes in Computer Science*, Springer, p. 897-914, 2007.

Kiryakov A., Ognyanov D., Manov D., « OWLIM - A Pragmatic Semantic Repository for OWL », *WISE Workshops*, p. 182-192, 2005.

Kiryakov A., Popov B., Ognyanoff D., Manov D., Goranov K. M., « Semantic annotation, indexing, and retrieval », *Journal of Web Semantics*, vol. 2, p. 49-79, 2004.

Lopez P., Romary L., « Multiple Retrieval Models and Regression Models for Prior Art Search », *CoRR*, 2009.

Manh Hung N., « Thesaurus Implementation in Integrated System of Information Resources (ISIR) », *Program. Comput. Softw.*, vol. 30, n° 4, p. 230-240, 2004.

Montiel-Ponsoda E., Aguado de Cea G., Gómez-Pérez A., Peters W., « Modelling Multilinguality in Ontologies », *Companion volume : Posters*, Coling 2008 Organizing Committee, Manchester, UK, p. 67-70, August, 2008.

Noy N. F., Shah N., Dai B., Dorf M., Griffith N., Jonquet C., Montegut M., Rubin D. L., Youn C., Musen M. A., « BioPortal : A Web Repository for Biomedical Ontologies and Data Resources », *in* C. Bizer, A. Joshi (eds), *International Semantic Web Conference (Posters & Demos)*, vol. 401 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2008.

Sabou M., Dzbor M., Baldassarre C., Angeletou S., Motta E., « WATSON : A Gateway for the Semantic Web », *Poster session of the European Semantic Web Conference, ESWC*, 2007.

Suchanek F., Kasneci G., Weikum G., « YAGO : A Core of Semantic Knowledge - Unifying WordNet and Wikipedia », *in* C. L. Williamson, M. E. Zurko, P. J. Patel-Schneider, Peter F. Shenoy (eds), *16th International World Wide Web Conference (WWW 2007)*, ACM, Banff, Canada, p. 697-706, 2007.

Vandenbussche P.-Y., Charlet J., « Méta-modèle général de description de ressources terminologiques et ontologiques », *in* F. L. Gandon (ed.), *Actes d'IC*, PUG, p. 193-204, 2009.

Wang Y., Haase P., Bao J., « A Survey of Formalisms for Modular Ontologies », *International Joint Conference on Artificial Intelligence Workshop SWeCKa*, Hyderabad, India, JAN, 2007.

Wright S. E., Budin G. (eds), *Handbook of Terminology Management*, vol. 1 — Basic Aspects of Terminology Management, John Benjamins, Amsterdam, 1997.

Zhu X., Fan J.-W., Baorto D. M., Weng C., Cimino J. J., « A review of auditing methods applied to the content of controlled biomedical terminologies », *Journal of Biomedical Informatics*, vol. 42, n° 3, p. 413 - 425, 2009. Auditing of Terminologies.