N° d'Ordre : D.U. ...
EDSPIC : ...

Université Paris 13 - Sorbonne Paris Cité
Ecole Doctorale Galilée

THÈSE

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITÉ PARIS 13

Spécialité : ***Informatique***

Construction d'une cartographie de domaine à partir de ressources sémantiques hétérogènes

présentée et soutenue publiquement par :

Sarra Ben Abbès

Le 25 octobre 2013

Composition du jury :

Directrice de thèse :	Adeline Nazarenko	-	Professeur, LIPN, Univ. Paris 13
Encadrant de thèse :	Haïfa Zargayouna	-	Maître de Conférences, LIPN, Univ. Paris 13
Rapporteurs :	Chantal Reynaud	-	Professeur, Université Paris Sud 11
	Gaël Dias	-	Professeur, Université de Caen Basse-Normandie
Examinateurs :	Pierre Zweigenbaum	-	Directeur de recherche, LIMSI-CNRS
	Claire Nédellec	-	Chargée de recherche, MIG, INRA
	François Lévy	-	Professeur, LIPN, Univ. Paris 13

Remerciement

C'est avec un grand plaisir que je réserve ces lignes en signe de gratitude et de sincère reconnaissance à tous ceux qui y ont participé de près ou de loin ainsi qu'à l'élaboration de ce mémoire de thèse.

Ce travail n'aurait pu aboutir sans mes encadrantes. Je tiens à remercier vivement Adeline Nazarenko, ma directrice de thèse pour m'avoir confié cette thèse, ainsi que pour son aide et ses conseils très précieux qui m'ont permis d'avancer au cours de ces années. Je la remercie énormément d'avoir confiance en mes compétences et de m'avoir fait évoluer avec son expérience qu'elle n'a de cesse à partager.

Je remercie aussi Haïfa Zargayouna, mon encadrante de thèse pour la confiance qu'elle m'a accordée en me proposant un stage de master et ensuite un sujet de thèse, pour sa disponibilité, pour son appui total lors de la réalisation de ce manuscrit de thèse et pour l'aide qu'elle m'a apportée, pour ses conseils et ses commentaires précieux durant ces années qui m'ont permis de surmonter mes difficultés et de progresser sur le plan professionnel et personnel.

Je tiens à remercier Chantal Reynaud, professeur à l'université de Paris Sud et Gaël Dias, professeur à l'université de Caen d'avoir accepté de rapporter ce travail. Je remercie également Pierre Zweigenbaum, directeur de recherche CNRS, Claire Nédellec, chargée de recherche à l'INRA et François Lévy, professeur à l'université Paris 13 d'avoir accepté d'examiner cette thèse.

Je remercie tout le personnel du LIPN et je tiens à remercier plus particulièrement tous les membres de l'équipe RCLN pour leur sympathie.

Un grand merci à mon adorable famille. Mes parents Hattab et Latifa qui ont été pour moi un exemple de courage, de persévérance et d'honnêteté. Je remercie également mes frères Bilel et Aymen et ma sœur Soumaya pour leur soutien moral et leurs encouragements permanents et je vous dis que tellement je vous aime, nulle expression ne peut définir mes sentiments envers vous. Je vous souhaite tout le bonheur du monde. Mes vifs remerciements vont aussi : à la famille Ben Abbès et notamment à la mémoire de ma grand-mère Beya et mon grand-père Jilani, à la famille Gharbi et notamment à la mémoire de mon grand-père Ahmed.

Je voudrais aussi remercier mes amis. Je n'oublierai jamais les bons moments qu'on a vécu ensemble ainsi que votre aide et plus particulièrement Zayd, Nouha, Abdoulaye, Nada, Hanene et Sondes.

Je vous dédie ce travail et je vous souhaite une vie pleine de joie et de prospérité.

Résumé

Ces dernières années, un effort considérable a été accompli pour le partage et la réutilisation des connaissances dans le cadre du Web sémantique. Un nombre important de ressources sémantiques ont été mises à disposition mais cette richesse et cette diversité compliquent la réutilisation de ressources existantes.

Avant de chercher à réutiliser des ressources, un ingénieur de la connaissance doit commencer par dresser un état des lieux du domaine qui l'intéresse, identifier les ressources disponibles et les positionner les unes par rapport aux autres. Il lui faut se repérer dans un web foisonnant dont l'hétérogénéité sémantique ne cesse de croître.

C'est le défi que cette thèse a cherché à relever. Nous proposons de cartographier le web sémantique sur un domaine particulier et nous avons mis au point une méthode qui permet de construire automatiquement de telles « cartographies de domaine » destinées à des ingénieurs de la connaissance souhaitant prendre connaissance des ressources disponibles pour un domaine particulier. Pour que ces cartographies soient centrées sur leurs centres d'intérêt, les ingénieurs fournissent en entrée un texte représentatif du domaine et de l'application visée, ce texte servant ensuite de pivot à l'ensemble de la méthodologie de construction de la cartographie.

Nous avons défini un *processus global de construction de cartographies* qui se décompose en trois étapes, en supposant résolue l'étape préalable de sélection des ressources sémantiques. Le texte d'acquisition servant de point de départ, la première étape consiste à lier les ressources sélectionnées au texte : c'est une phase d'annotation sémantique qui projette les entités d'ontologies sur le texte afin d'identifier celles qui y sont mentionnées. La deuxième étape permet d'aligner les différentes ressources et ce processus d'alignement est lui-aussi guidé par le texte : les entités de ressources différentes sont rapprochées sur la base de leur proximité distributionnelle dans le texte. La troisième étape permet de détecter et au besoin de corriger les anomalies sur les correspondances obtenues à l'issue de la phase d'alignement ; cette phase permet aussi de détecter et d'afficher les correspondances remarquables qui peuvent aider l'ingénieur à prendre connaissance du matériau existant.

Notre *méthode d'alignement d'ontologies* présente la particularité d'être guidée par le texte. L'alignement consiste classiquement à identifier des relations ou « correspondances » existant entre les entités issues d'ontologies différentes et il est utilisé dans plusieurs applications, comme l'enrichissement d'une ontologie, la fusion de plusieurs ontologies ou, ici, pour aider l'ingénieur de la connaissance à appréhender les ressources disponibles pour un domaine particulier. L'exploitation d'un texte permet de rapprocher des entités ontologiques qui n'ont ni la même étiquette – comme dans les méthodes d'alignement lexical – ni la même position dans les ressources d'origine – comme dans les approches structurelles. Nous proposons de différencier deux types de correspondance selon la nature de la relation existant entre les termes associés : si les termes ou étiquettes tendent à apparaître dans les mêmes phrases nous présumons l'existence d'une relation associative entre eux ; à l'inverse, s'ils se substituent l'un à l'autre, on peut faire l'hypothèse d'une relation d'équivalence.

La *construction d'une cartographie de domaine* se fait à partir des résultats de l'alignement. Cette phase consiste à repérer les anomalies et les correspondances remarquables

à présenter à l'ingénieur de la connaissance. Nous faisons en effet l'hypothèse que ces deux types de configurations sont importantes à repérer : les anomalies font apparaître des différences notables dans les choix de conceptualisation sur lesquels reposent les deux ressources alignées ; les configurations remarquables font au contraire ressortir les zones les plus centrales et les plus cohérentes de l'alignement proposé. Repérer et analyser ces configurations doit donc permettre à l'ingénieur de comprendre comment les ressources alignées se positionnent l'une par rapport à l'autre.

L'ensemble de cette approche a été *implémenté et testé* sur différents cas d'usage dans les domaines de la biologie, de la géographie et de l'alimentation.

Mots clés alignement, ontologie, hétérogénéité, ressource sémantique, cartographie, interopérabilité.

Abstract

In recent years, a large effort has been devoted to the sharing and reuse of knowledge in the semantic web. A number of semantic resources have been made available significantly but this richness and diversity complicate the reuse of existing resources.

Before trying to reuse resources, the knowledge engineer must begin by drawing up an inventory of the domain's interest, identify available resources and put it against each other. He must locate in a "teeming" web where the semantic heterogeneity is growing.

This is the challenge that this thesis has sought to address. We propose to map the semantic web on a particular domain and we have developed an automatic method to create such "domain cartographies" for engineers wishing to gain insught in the available resources on a particular domain. In order to have these cartographies suited to their interests, engineers give as an input a text which is representative of the domain and the target application and the text is a pivot to the whole the cartography process.

We have defined an *overall process of building cartographies* that is composed of three steps, assuming resolved the prior selection step of semantic resources. The text acquisition is a starting point, the first step is to link the selected resources to the text : it is a step of semantic resources annotation which projects ontological entities on the text in order to identify those which are mentioned. The second step is to align different resources and the alignment process itself is also based on the text : the entities of different resources are aligned based on their distributional proximity in the text. The third step is to detect and correct the mapping anomalies obtained after the alignment step ; this step can also detect and display the remarkable mappings that can help the engineer to take knowledge of the existing resources.

The originality of our *ontology alignment method* is to use the text. The alignment consists typically on identifying relations or mappings between entities from different ontologies. It is used in several applications, such as the ontology enrichment, the ontology merging or here, to help the knowledge engineer to understand the available resources of a particular domain. Using a text allows closer ontological entities that do not have the same label – as in the lexical alignment methods – or the same position in the original resources – as in structural approaches. We propose to distinguish between two types of mappings with the nature of the relationship between related terms : if the terms or labels tend to appear in the same sentences, we conclude an associative relationship between them ; on the contrary, if they are substituted to each other, we can conclude an equivalence relation.

The *building cartography process* is based on the results of the alignment process. This step consists on identifying anomalies and to present the remarkable mappings to the knowledge engineer. We make the hypothesis that these two types of configurations are important to identify : anomalies show significant differences in the choice of conceptualization represented on the two aligned resources ; remarkable configurations are instead highlighting the most central parts and most consistent of the proposed alignment. Identify and analyze these configurations must allow the engineer to understand how resources are aligned to each other.

This overall approach was *implemented and tested* on different use cases in the biology, geography and food domains.

Keywords : alignment, ontology, heterogeneity, semantic resource, cartography, interoperability

Table des matières

Table des figures

Introduction générale

Contexte et motivation

Ces dernières années, un effort considérable a été accompli pour le partage et la réutilisation des connaissances dans le cadre du Web sémantique. Un nombre important de structures de connaissances ont été mises à disposition. Ce sont des « ressources sémantiques » qui peuvent être exploitées dans de nombreux contextes applicatifs, pour annoter des pages web et des documents, comme bases de connaissances pour des systèmes qui intègrent du raisonnement, ou même pour construire de nouvelles ressources à partir de celles qui existent. La diversité des applications visées et des champs thématiques couverts fait la richesse du web sémantique mais complique la réutilisation de ressources existantes pour une nouvelle application, nouvelle tâche et/ou nouveau domaine.

Pour construire une ressource sémantique, on cherche cependant souvent à réutiliser les ressources disponibles – quitte à les adapter et à les combiner – plutôt que de partir de rien. Dans certains cas, on préfère même « faire avec » les ressources existantes, aussi limitées et biaisées soient-elles, plutôt que d'en produire de nouvelles. Tout dépend évidemment de l'application visée et de l'importance de la cohérence sémantique des ressources utilisées : il y a en effet une marge entre l'annotation sémantique grossière d'un texte à l'aide d'un thesaurus généraliste et le raisonnement sur des ontologies formelles.

Avant de décider de réutiliser une ressource, le premier défi consiste à prendre connaissance de l'état des connaissances sur un sujet donné : quelles sont les ressources existantes ? sont-elles redondantes ou complémentaires ? quel(s) domaine(s) ou sous-domaine(s) couvrent-elles et selon quel point de vue ? quel est leur degré de formalisation ? comportent-elles un volet lexical ? Il s'agit donc de commencer par dresser un état des lieux, identifier les ressources disponibles et les positionner les unes par rapport aux autres. Il faut se donner les moyens de se repérer dans un web foisonnant dont l'hétérogénéité sémantique ne cesse de croître. C'est le défi que cette thèse a cherché à relever.

Nous proposons de cartographier le web sémantique sur un domaine particulier et nous avons mis au point une méthode qui permet de construire automatiquement de telles « cartographies de domaine » destinées à des ingénieurs de la connaissance souhaitant prendre connaissance des ressources disponibles pour un domaine particulier.

Problématique

Cette thèse propose une méthode qui permet de dresser, sous la forme d'une cartographie de domaine, l'état des ressources sémantiques disponibles pour un domaine donné. Ce projet comportait différentes difficultés.

L'*hétérogénéité des ressources* est la première d'entre elle. Cartographier est en effet d'autant plus complexe que les ressources disponibles sur un domaine particulier sont généralement hétérogènes dans leurs formats et leur degré de formalisation, leur couverture du domaine visé et leur degré de généralité, et même dans la nature des connaissances

véhiculées : connaissances lexicales ou conceptuelles, génériques ou factuelles, par exemple. L'approche proposée prend cette hétérogénéité en compte. En pratique, nous avons essentiellement travaillé à partir d'ontologies lexicalisées mais des ressources moins riches (terminologies ou thesaurus) sont également exploitables.

Une autre difficulté concerne la caractérisation des *domaines* d'intérêt qui sont toujours difficiles à cerner, à désigner et à décrire. On sait par exemple que le domaine de la « géographie de la Méditerranée » n'est pas vu de la même manière par le cartographe qui cherche à produire des cartes de navigation et l'économiste qui analyse les flux marchands : les contours de la région visée, le niveau de détail requis et les entités géographiques à prendre en compte diffèrent et on n'aura pas recours à la même ressource sémantique dans les deux cas. Un mot-clé ou une liste de mots-clés ne suffisent généralement pas à décrire le domaine visé parce qu'il faut pouvoir indiquer assez précisément dans quelle perspective le domaine est abordé. Nous avons fait l'hypothèse que l'ingénieur de la connaissance peut fournir un texte pour caractériser ce domaine auquel il s'intéresse. Selon les cas, il s'agira d'un texte décrivant l'application visée (document de spécification, par exemple), d'un document pédagogique ou d'un document qui reflète l'activité de ce domaine (articles scientifiques, rapports de panne, etc.). Ce texte d'acquisition sert à sélectionner des ressources sémantiques et à les positionner les unes par rapport aux autres.

La troisième difficulté tient à la forme de la *cartographie* produite. Indépendamment des interfaces ou métaphores de visualisation choisies, la question se pose de savoir 1) quelle est l'information qui aide effectivement l'ingénieur à prendre connaissance d'un domaine et à s'y positionner et 2) comment représenter l'information utile sans surcharger l'ingénieur de bruit. Nous avons pris le parti de la simplicité pour laisser la plus grande marge d'interprétation à l'ingénieur. Une cartographie n'a pas vocation à être une nouvelle ressource sémantiquement cohérente mais elle doit préserver les ressources sur lesquelles elle s'appuie, qui ont été publiées et qui ont *a priori* chacune leur cohérence propre : elle doit montrer à la fois ce qui rapproche les ressources disponibles et ce qui les distingue. Sur la base de cette analyse, nous proposons de cartographier un domaine du web sémantique en alignant les différentes ressources pertinentes pour ce domaine et en soulignant les zones ou configurations remarquables dans l'alignement produit, c'est-à-dire les zones de convergence qui pourraient servir de point de départ pour, si besoin est, fusionner les ressources, mais aussi les zones de contraste qui témoignent de choix de conceptualisation différents et qui empêcheraient la fusion au contraire.

Contributions

Nous construisons une cartographie de domaine représentant un ensemble de liens entre les connaissances des différentes ressources recensées. L'objectif est d'aider l'ingénieur de la connaissance à analyser les ressources retournées les unes par rapport aux autres. La méthodologie proposée repose sur l'exploitation de la richesse et de la diversité des ressources sémantiques en préservant la cohérence propre à chacune et en les articulant entre elles.

Nous avons défini un *processus global de construction de cartographies* qui se décompose en trois étapes, en supposant résolue l'étape préalable de sélection des ressources

sémantiques [1] :

- le texte d'acquisition servant de point de départ, la première étape consiste à lier les ressources sélectionnées au texte : c'est une phase d'annotation sémantique qui consiste à projeter les entités sémantiques sur le texte afin d'identifier celles qui sont ancrées, c'est-à-dire qui sont mentionnées dans le texte ;
- la deuxième étape permet d'aligner les différentes ressources et ce processus d'alignement est lui-aussi guidé par le texte : les entités de ressources différentes sont rapprochées sur la base de leur proximité distributionnelle dans le texte ;
- la troisième étape permet de détecter et de corriger les anomalies sur les correspondances obtenues à l'issue de la phase d'alignement ; cette phase permet aussi de détecter et d'afficher les correspondances remarquables qui peuvent aider l'ingénieur à prendre connaissance du matériau existant.

Notre *méthode d'alignement d'ontologies* présente la particularité d'être guidée par le texte. L'alignement consiste classiquement à identifier des relations ou « correspondances » existant entre les entités issues d'ontologies différentes et il est utilisé dans plusieurs applications, comme l'enrichissement d'une ontologie, la fusion de plusieurs ontologies ou, ici, pour aider l'ingénieur de la connaissance à appréhender les ressources disponibles pour un domaine particulier. L'exploitation d'un texte permet de rapprocher des entités ontologiques qui n'ont ni la même étiquette – comme dans les méthodes d'alignement lexical – ni la même position dans les ressources d'origine – comme dans les approches structurelles. Nous proposons de différencier deux types de correspondance selon la nature de la relation existant entre les termes associés : si les termes tendent à apparaître dans les mêmes phrases nous présumons l'existence d'une relation associative entre eux ; à l'inverse, s'ils se substituent l'un à l'autre, on peut faire l'hypothèse d'une relation d'équivalence.

Notre troisième contribution concerne la *construction d'une cartographie de domaine* à partir des résultats de l'alignement. Cette phase consiste à repérer les anomalies et les correspondances remarquables à présenter à l'ingénieur de la connaissance. Nous faisons en effet l'hypothèse que ces deux types de configurations sont importantes à repérer : les anomalies font apparaître des différences notables dans les choix de conceptualisation sur lesquels reposent les deux ressources alignées ; les configurations remarquables font au contraire ressortir les zones les plus centrales et les plus cohérentes de l'alignement proposé. Repérer et analyser ces configurations doit donc permettre à l'ingénieur de comprendre comment les ressources alignées se positionnent l'une par rapport à l'autre.

L'ensemble de cette approche a été *implémenté et testé* sur différents cas d'usage qui sont présentés au fur et à mesure dans le corps de la thèse.

Structure du document

Le présent mémoire est organisé en six chapitres (voir figure 1).

- Le premier chapitre est consacré aux travaux de l'état de l'art qui portent sur la gestion des ressources sémantiques. Nous décrivons les ressources sémantiques ainsi que leur contenu. Nous présentons ensuite la problématique de la gestion de ces res-

1. Différents moteurs de recherche de ressources sémantiques existent à cet effet (voir le *Linked Open Vocabularies* (LOV, http ://lov.okfn.org/dataset/lov/).

sources, reliée à leur hétérogénéité et leur réutilisation. Nous exposons également les travaux sur l'alignement et les types de techniques utilisées (terminologiques et structurelles). Nous abordons enfin la question des cartes et cartographies conceptuelles pour positionner notre travail bien que notre approche soit assez différente de ces dernières. Nous concluons ce chapitre par notre positionnement par rapport à l'état de l'art.

- Le deuxième chapitre présente une vue d'ensemble de la méthodologie proposée. Nous mettons l'accent sur les ontologies lexicalisées, dans la famille des ressources sémantiques. Notre méthodologie comporte trois phases 1) d'annotation, pour relier le texte aux ontologies, 2) d'alignement, pour rapprocher les entités d'ontologies en se fondant sur le texte, comme support de travail, et (3) de construction de la cartographie de domaine, pour réviser la sortie d'alignement et la présenter à l'ingénieur de la connaissance. Un exemple illustratif est présenté.
- Le troisième chapitre présente notre méthode d'alignement guidé par le texte, *TOM* (*Text-based Ontology Mapping*). Cet alignement est fondé sur la cooccurrence dans le texte des termes associés aux entités ontologiques et comporte lui-même deux phases : 1) le calcul de correspondances, qui consiste à chercher les correspondances entre les entités des ontologies, et 2) le filtrage, qui permet d'éliminer des correspondances périphériques. Une description de l'implémentation de la méthode proposée est présentée dans ce chapitre.
- Le quatrième chapitre montre ce qu'est une cartographie de domaine. Il s'agit d'analyser l'ensemble des correspondances obtenues en sortie d'alignement. Différentes anomalies et configurations de liens remarquables sont modélisées et étudiées. Nous présentons à la fin de ce chapitre une maquette présentant une cartographie de domaine.
- Le cinquième chapitre présente les expériences que nous avons conduites. Trois cas d'usage sont décrits. Ils relèvent de différents domaines, biologique, géographique et alimentaire. Nous présentons ensuite les expériences en détaillant les trois grandes étapes de la méthodologie de construction de la cartographie proposée.
- Le sixième chapitre décrit les expériences complémentaires que nous avons menées pour évaluer notre méthode d'alignement. La première approche repose sur la comparaison des résultats de *TOM* avec ceux d'une approche de l'état de l'art (TaxoMap). La deuxième partie de ce chapitre présente le protocole que nous avons conçu pour faire valider par un juge l'ensemble de correspondances : le but est de pouvoir comparer les résultats de TOM et de TaxoMap à cette référence. La troisième approche consiste à comparer TOM aux systèmes d'alignement disponibles dans la campagne d'évaluation de l'alignement d'ontologies (OAEI). Nous étudions aussi le choix du texte d'acquisition et son influence sur les résultats d'alignement.

FIGURE 1 – Synthèse du plan de la thèse

Chapitre 1

Gestion de ressources sémantiques

Sommaire

1.1 Introduction

Le Web sémantique (WS) est défini par Tim Berners-Lee [Berners-Lee et Hendler, 2001] comme « une extension du web actuel dans laquelle l'information se voit associée à un sens bien défini améliorant la capacité des ordinateurs et des hommes à travailler en coopération » [1]. Son développement et la définition de différents langages de représentation des connaissances ont favorisé la création et la publication de nombreuses bases de connaissances : des ontologies, thésaurus, taxonomies et dictionnaires. Nous utilisons le terme de « ressources sémantiques » pour désigner ces structures de connaissances dès lors qu'elles sont publiées et peuvent être utilisées comme base de connaissances dans des applications. Une fois construites et publiées, ces ressources sémantiques peuvent également être réutilisées, adaptées et partagées.

1. Traduction de C. Dubois reprise sur le site du Centre National de la Documentation Pédagogique (http ://www.cndp.fr/savoirscdi/societe-de-linformation/le-monde-du-livre-et-de-la-presse/histoire-du-livre-et-de-la-documentation/biographies/tim-berners-lee-1955.html).

Il existe une grande hétérogénéité de ressources. Elles contiennent des connaissances terminologiques (thésaurus et dictionnaires) ou conceptuelles (ontologies et taxonomies). Elles concernent des domaines différents et, même pour un même domaine, elles n'en donnent pas la même couverture. Le degré de granularité de la description peut aussi varier d'une ressource à une autre. Enfin, ces ressources peuvent être représentées et manipulées dans différents langages du web sémantique (voir figure 1.1) : le langage RDF (Resource Description Framework) décrit les ressources du web sous la forme de triplets (ressource, propriété, valeur) alors que les langages SKOS (Simple Knowledge Organization System) et OWL (Web Ontology Language) sont dédiés respectivement à la représentation de structures thésauriques et de connaissances ontologiques (toutes les ressources du Web peuvent être interrogées *via* des requêtes SPARQL (SPARQL Protocol and RDF Query Language)).

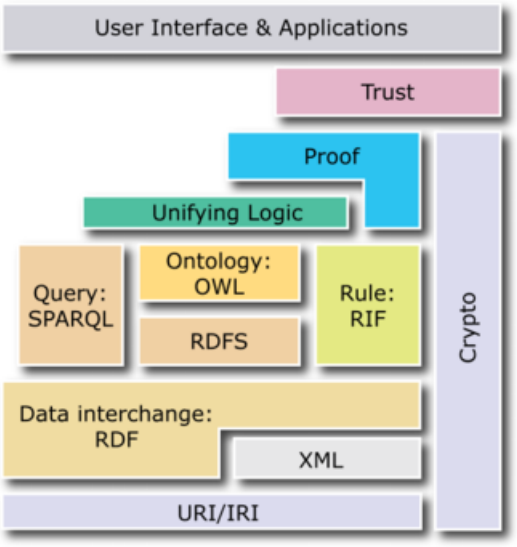


FIGURE 1.1 – Pyramide du web sémantique [Berners-Lee et Hendler, 2001]

Le développement de ces ressources nécessite de créer en parallèle des méthodes et des outils pour en assurer la gestion, c'est-à-dire les construire, les décomposer, les mettre à jour, voire les combiner. Notre objectif est de proposer un outil de construction de cartographies qui serve à prendre connaissance des ressources sémantiques existant sur un domaine particulier et à les confronter les unes aux autres. Au sein des méthodes de gestion des ressources sémantiques, nous nous intéressons donc tout particulièrement aux méthodes d'alignement sur lesquelles nous nous appuyons pour construire des cartographies de domaine.

Nous présentons, dans ce chapitre, une revue des travaux de l'état de l'art concernant la gestion de ces ressources. Nous commençons par présenter les différents types de ressources sémantiques disponibles et les méthodes qui ont été proposées pour en assurer la gestion (construction, modularisation, fusion, etc.). La principale difficulté vient de l'hétérogénéité des ressources disponibles qui soulève un problème d'interopérabilité sémantique. Nous exposons dans la section 1.4 les travaux sur l'alignement d'ontologies comme une so-

lution possible au problème de l'interopérabilité sémantique : nous présentons à la fois les techniques terminologiques et structurelles et celles qui exploitent des ressources externes telle que WordNet. Nous présentons pour finir les travaux sur les cartographies conceptuelles qui se rapprochent de la cartographie de domaine que nous cherchons à produire en ce qu'ils « donnent à lire » un domaine ou un ensemble de notions interalliées. Nous concluons ce chapitre par un bilan où nous nous positionnons par rapport à cet état de l'art.

1.2 Ressources sémantiques

La masse d'informations sur le Web reste difficile à exploiter. Le Web Sémantique est apparu comme une solution intéressante pour structurer et partager les informations sur la toile. Ces informations sont stockées et représentées dans les ressources sémantiques. Une ressource sémantique *RS* se définit généralement comme un ensemble d'entités sémantiques (ex. concept, terme, descripteur, instances, propriétés) plus ou moins reliées entre elles par un ensemble de relations (ex. relation hiérarchique, relation associative). Une ressource a une nature et une spécificité. Nous décrivons dans ce qui suit ces ressources comme nous les entendons dans notre travail.

1.2.1 Terminologie

[Lefèvre, 2000] a défini une terminologie [2] comme « une liste de termes d'un domaine ou d'un sujet particulier, faisant référence à des notions qui sont fréquemment utilisées. Cette liste est non-structurée ». Dans le but de structurer les termes d'un domaine, il est important de les relier. On parle dans ce cas d'un réseau terminologique. En effet, [Bourigault et Charlet, 2005] ont défini le réseau terminologique comme « un ensemble de termes (des mots) reliés entre eux par des relations lexicales » à savoir [Cruse, 1986], la synonymie, l'hyperonymie/hyponymie, l'antonymie.

La terminologie désigner, pour nous, une liste de termes d'un domaine spécifique reliés par des liens lexicaux (ex. voir la figure 1.2). La construction du réseau terminologique est considérée comme une étape essentielle dans le processus de la construction d'ontologies à partir de textes comme dans [Després et Szulman, 2008].

1.2.2 Thésaurus

Dans le but d'indexer les documents d'un domaine spécifique, une terminologie contrôlée et structurée est utilisée, appelée le *thésaurus*. Il existe différentes normes pour représenter les connaissances d'un thésaurus à savoir AFNOR 1987, ISO 2788-1986, SKOS, etc. Chacune de ces normes donne une définition. La définition des normes ISO 2788-1986 et ANSI Z39 nous parait la plus complète : « un thésaurus est une liste d'autorité organisée de descripteurs et de non-descripteurs obéissant à des règles terminologiques propres et reliés entre eux par des relations sémantiques (hiérarchiques, associatives ou d'équivalence) ».

2. A terminology is « a list of domain or subject terms referring to concepts or notions which are frequently used. This list is non-structured ».

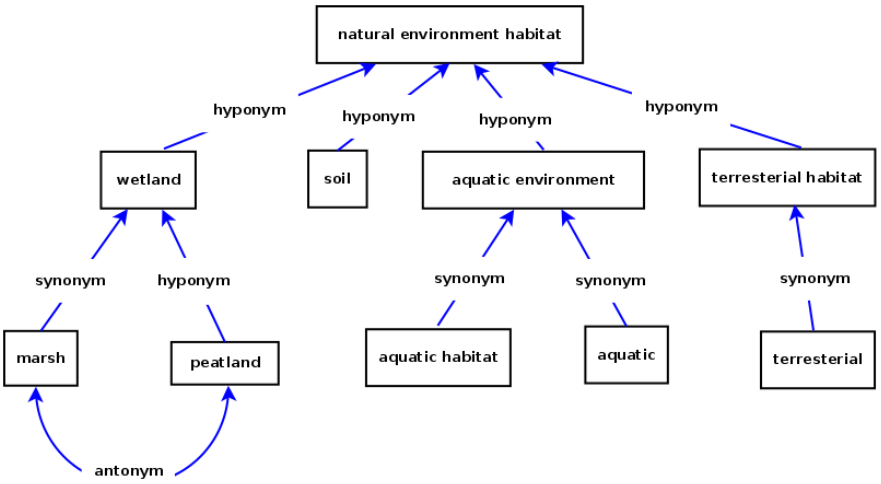


FIGURE 1.2 – Réseau terminologique associé au terme « natural environment habitat » extrait des données fournies par notre partenaire INRA-MIG.

Les descripteurs sont des termes préférentiels retenus pour éliminer toute ambigüité dans un domaine spécifique. Tous les autres termes du domaine qui n'ont pas été retenus sont les non-descripteurs ou les termes non-préférentiels du domaine. Le thésaurus a permis de construire d'autres ressources comme les ontologies [Chrisment *et al.*, 2008] ou les cartes thématiques [3] [Ellouze *et al.*, 2012]. Beaucoup de thésaurus sont disponibles, comme par exemple :

- UNESCO est « une liste de termes contrôlés et structurés pour l'analyse thématique et la recherche de documents et publications dans les domaines de l'éducation, la culture, les sciences naturelles, les sciences sociales et humaines, la communication et l'information » [4]. Il comporte 7 000 termes en anglais et en russe, et 8 600 en français et en espagnol.
- MeSH (Medical Subject Headings) est « le thésaurus de référence dans le domaine biomédical et les domaines connexes. Il est utilisé pour indexer, classer et rechercher des documents, notamment ceux des bases de données de la NLM [5], dont MEDLINE/PubMed » [6]. Il comprend 26 853 descripteurs en 2013 répartis en 16 catégories recouvrant la biologie, la médecine et des domaines connexes.

1.2.3 Dictionnaire

Il existe différentes catégories de dictionnaires à savoir les dictionnaires encyclopédiques, les dictionnaires de langue, les dictionnaires bilingues et les dictionnaires de spécialité. On s'intéresse à la dernière catégorie qui permet de décrire les informations d'un domaine précis. Un dictionnaire spécialisé est défini par [Jacquemin et Ploux, 2006] comme « un ensemble des entrées retenues et des définitions ». Les entrées retenues sont des termes

3. `http://topicmaps.org/`
4. `http://databases.unesco.org/thesfr/`
5. National Library of Medicine
6. `http://mesh.inserm.fr/mesh/`

représentant des unités lexicales dont le sens peut être associé à un domaine spécifique. Les relations sémantiques entre les termes s'inspirent largement des relations lexicales dans une terminologie (synonymie, antonymie, etc). Citons l'exemple du dictionnaire DiCoInfo (Dictionnaire fondamental de l'informatique et de l'Internet) qui a pour objectif de décrire les termes fondamentaux du domaine de l'informatique.

1.2.4 Glossaire

Un glossaire est défini par [Lassila et McGuinness, 2001] comme « un ensemble de termes et leur signification ». Autrement dit, les termes représentent les notions du domaine et les significations qui sont associées aux définitions. Les définitions d'un terme donné ne sont pas nécessairement les mêmes dans un dictionnaire et un glossaire, car le dictionnaire est lié à la langue générale (exemples et illustrations) alors que le glossaire est plus centré sur les termes peu fréquents en décrivant les termes techniques d'un domaine spécifique. Citons par exemple, le glossaire des arbres[7] comportant les termes utilisés pour identifier un arbre à savoir les feuilles, les fleurs, les fruits, etc.

1.2.5 Ontologie

Une des définitions de l'ontologie la plus citée dans l'état de l'art est celle de Gruber [Gruber, 1995] : « une ontologie est une spécification explicite et formelle d'une conceptualisation partagée d'un domaine de connaissances ». Cette définition met en évidence les caractéristiques suivantes d'une ontologie :

- Conceptualisation, cela sert à représenter un modèle abstrait ainsi que les concepts spécifiant le monde réel ;
- Explicite, correspond à la définition précise des concepts et des contraintes de leur utilisation ;
- Formelle, cela fait référence à la compréhension de l'ontologie par les machines ;
- Partagée, cela réfère à la nécessité d'une vision consensuelle du modèle.

Il existe trois types d'ontologies d'après [van Heijst *et al.*, 1997] : (i) les ontologies de haut niveau, contenant des concepts très abstraits indépendants du problème traité (ex. SUMO [Niles et Pease, 2001], Upper Cyc [Microsystems, 2001]), (ii) les core-ontologies (ontologies noyau), comportant des concepts centraux d'un problème donné (ex. LKIF-core [Hoekstra *et al.*, 2007], Dolce [Borgo et Masolo, 2009]), et (iii) les ontologies du domaine pour une application, possédant des concepts spécifiques à un domaine et une application particulière (ex. ONTOLINGUA). On s'intéresse dans notre travail à ce dernier type d'ontologies car nous considérons que ce type d'ontologies est plus spécifique que les autres et aussi il est considéré comme la meilleure représentation des connaissances d'un domaine particulier. L'ontologie comporte un ensemble de concepts et des relations entre eux (relations hiérarchiques et relations associatives).

Dans une perspective centrée autour du texte qui établit les liens entre les ontologies et les ressources textuelles, il est important de lier les deux niveaux lexical et structurel. On parle dans ce cas de « lexicalisation » de l'ontologie [Cimiano, 2006]. Ces ontologies

7. `http://www.lesarbres.fr/glossaire.php#FLE`

comportent des concepts possédant des termes associés. Un concept est dénoté par plusieurs termes. Les termes d'un concept sont généralement reliés par une relation de synonymie. Pour qu'il soit « lexicalisé », un concept doit avoir au minimum un terme associé.

Plusieurs formalismes de représentation des ontologies dans le WS ont été proposés tels que OWL.

1.2.6 Taxonomie

Une taxonomie est définie par [Kefi *et al.*, 2006] comme « un ensemble de concepts reliés par des relations is-a ». Plus précisément, une taxonomie est une hiérarchie de concepts organisés dans le sens de la spécialisation (du général au particulier). On peut dire qu'une taxonomie est comme une ontologie lexicalisée sans les relations associatives (rôles). Citons l'exemple d'une taxonomie des sciences de la vie et de la terre et de l'univers (SVSTU) [8] qui présente une hiérarchie de concepts de la santé, de la terre, etc.

1.3 Défis de la gestion des ressources sémantiques

Notre objectif de recherche se situe au cœur des ressources sémantiques et de leur apport pour construire le Web Sémantique. Les défis portent sur la gestion de ces ressources. Différentes méthodes ont été proposées dans le domaine de l'ingénierie de connaissances.

Dans ce qui suit (voir figure 1.3), nous présentons les méthodes de gestion des ressources sémantiques. Ensuite, nous exposons la problématique de la réutilisation de ces ressources et enfin, nous décrivons une des solutions proposées dans l'état de l'art relatif à l'interopérabilité sémantique à savoir l'alignement.

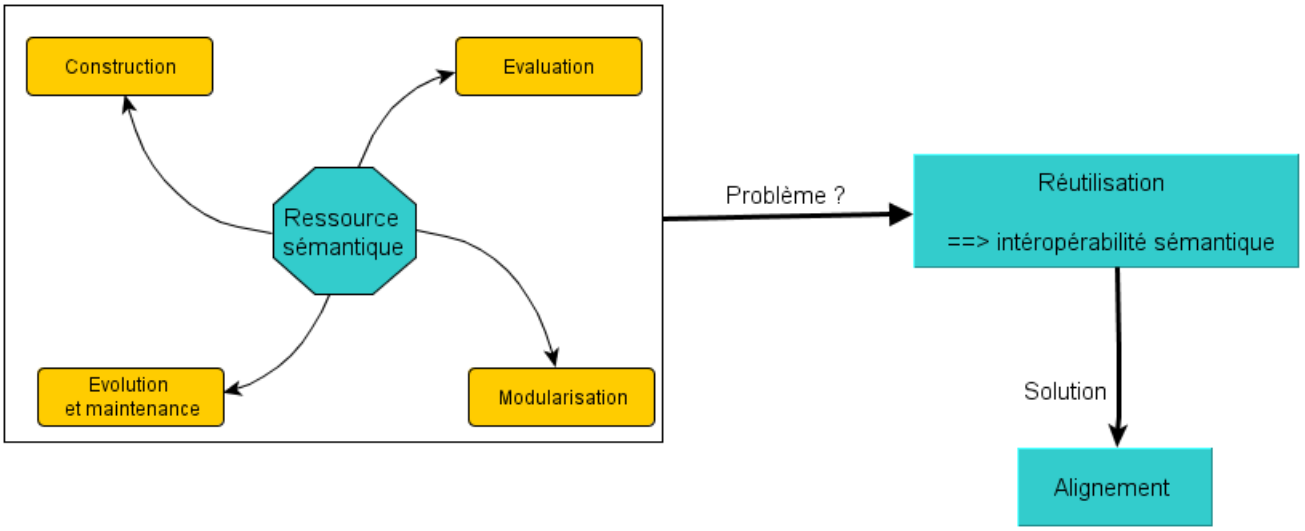


FIGURE 1.3 – Plan de la section 1.3

8. http://pratic.ens-lyon.fr/projets/meta-LOM-CDM/taxo_edu/taxonomies/svstu/

Construction d'une ressource sémantique Construire une ressource sémantique est un processus de modélisation de connaissances d'un domaine particulier. La création d'un modèle de connaissances nécessite soit de spécifier les besoins à partir de rien [Bourigault *et al.*, 2004, Cimiano et Völker, 2005] soit de réutiliser les ressources sémantiques [Abadie et Mustière, 2010]. De nombreuses méthodes de construction de ressources sémantiques ont été proposées, notamment Terminae [Aussenac-Gilles *et al.*, 2000] pour la construction des ontologies à partir de textes.

Évaluation des ressources sémantiques La diversité des motivations de construction des ressources et la complexité des domaines de spécialité rend difficile l'évaluation de la « qualité » des ressources sémantiques. L'évaluation des ressources apparaît cependant comme une problématique incontournable [Maedche *et al.*, 2002] : on cherche à évaluer les résultats fournis par les outils d'acquisition de ces ressources. Il n'existe pourtant pas encore de cadre fédérateur permettant l'évaluation des ressources sémantiques, ce qui s'explique en partie par la complexité de leurs structures. Nous avons généralement une référence (ou *gold standard*) qui permet d'effectuer des comparaisons. Or, la référence elle même est très dépendante du domaine modélisé et de l'utilisation prévue. Pour évaluer, on utilise généralement des mesures classiques telles que la précision et le rappel. Nous avons proposé dans [Ben Abbès *et al.*, 2010] de décomposer le problème de l'évaluation des ontologies en sous-problèmes suivant la nature des données manipulées (concepts, relations) et de calculer la précision et le rappel selon des jugements gradués.

Évolution et maintenance des ressources sémantiques L'évolution des ressources sémantiques, d'après [Haase et Stojanovic, 2005], consiste à gérer des changements qui peuvent intervenir sur une ressource suite à un ajout, une modification ou une suppression d'entités. Les travaux de [Zablith *et al.*, 2008, Baneyx et Charlet, 2006] s'appuient sur une base de connaissances pour enrichir une ressource sémantique ou la modifier. Faire évoluer une ressource repose sur différentes règles et patrons de changements relatifs aux spécificités de chaque ressource. La création de ces règles est importante car on peut par exemple effacer une entité et génèrer de ce fait des problèmes de raisonnement.

[Sellami *et al.*, 2012] propose un outil de construction et de maintenance d'ontologies à partir de textes, appelé DYNAMO (DYNAMic Ontology for information retrieval). Cet outil se fonde sur un système multi-agents comportant un premier agent qui se charge d'extraire la partie terminologique de l'ontologie (extraire les relations lexicales) et un deuxième agent permettant d'extraire la partie conceptuelle de l'ontologie (extraire les relations conceptuelles). A partir des informations textuelles, DYNAMO permet de proposer de nouveaux concepts, de nouvelles relations qui font évoluer l'ontologie.

Modularisation des ressources sémantiques La multitude de ressources sémantiques de grande taille sur le Web donne un large choix de ressources mais cela soulève aussi des problèmes d'exploitation et de raisonnement. La notion de modularisation, récemment apparue, concerne les ontologies dans le but de les décomposer en différentes sous-parties autonomes dites « modules » ayant des relations entre elles. A titre d'exemple, l'ontologie

de gènes et des produits géniques GO (Gene Ontology [9]) peut être fractionnée en plusieurs modules : module pour toutes les espèces, module des produits géniques au milieu intracellulaire, module des gènes, etc [d'Aquin *et al.*, 2009]. Différents critères ont été définis pour découper ces ontologies, par exemple la taille des modules, la distance inter-modules (liens entre les concepts de deux modules) ou la distance intra-modules (liens entre les concepts du module) [d'Aquin *et al.*, 2009]. Le travail de [Ben Abbès *et al.*, 2012] va plus loin pour caractériser les ontologies déjà modulaires et disponibles sur le web, par des patrons décrivant une structure plus claire de ces ontologies.

Notre problématique est centrée sur les aspects de réutilisation des ressources sémantiques. Plus spécifiquement, la diversité des ressources ainsi que l'hétérogénéité de leur contenu nécessite une réflexion approfondie sur la manière de les réutiliser. Cette hétérogénéité est due à la diversité des moyens mis en œuvre pour créer ces ressources et à la complexité des domaines de spécialité. Elle repose sur la spécificité de la ressource et les connaissances qu'elle manipule :

- nature de la ressource : les ressources sémantiques manipulent deux types de connaissances : (i) terminologique (termes dans le thésaurus et le dictionnaire) et conceptuelle (concepts dans les ontologies et taxonomies). A titre d'exemple, la notion de « bâtiment » (voir la figure 1.4) peut être représentée sous forme d'un concept « bâtiment » dans une ressource conceptuelle et sous forme d'un terme « bâtiment » (buildings) dans une ressource terminologique avec des relations lexicales (UF ; Used For, pour les synonymes), des relations associatives (RT ; *related term*) et des relations hiérarchiques (BT : *broader term*, pour un terme générique et NT : *narrower term*, pour un terme spécifique).
- granularité de la description de la ressource : une même notion (connaissance) n'a pas le même niveau de détails d'une ressource à l'autre. On peut avoir une ressource ayant un vocabulaire plus riche qu'une autre, ce qui permet de décrire d'une manière détaillée les notions d'un domaine particulier. Le degré de détails est utile quand il s'agit de définir des liens entre les différents types de ressources. Dans la figure 1.5, deux concepts « réservoir » de deux ontologies différentes sont identiques mais ils ne sont pas représentés avec le même degré de détails dans une ressource que dans l'autre. Il existe des ressources qui contiennent des notions générales sur un domaine (ex. Eurovoc) et d'autres qui sont spécifiques à un sujet donné d'un domaine particulier (ex. l'ontologie Kaon [10] décrivant les plantes : fleurs, couleur, longueur, etc) ;
- couverture d'un domaine de spécialité : la couverture d'un domaine spécifique varie d'une ressource sémantique à l'autre. Les ressources sémantiques couvrent plus ou moins partiellement un domaine de spécialité. La couverture est reliée à la présence des entités et des relations de ressources sémantiques associées aux données d'un domaine particulier. Un des supports où les informations d'un domaine peuvent être circonscrites, est le texte. Plus particulièrement, dans le travail de [Brewster *et al.*, 2004], on parle de la notion de couverture pour évaluer l'adéquation d'une ontologie à un corpus textuel. La couverture dans ce travail est définie par le taux de termes en corpus qui sont associés aux concepts de l'ontologie. Plus

9. `http://www.geneontology.org/`
10. `http://sourceforge.net/projects/kaon/`

FIGURE 1.4 – Exemple de deux représentations (concept et terme) de la même notion « bâtiment » (extrait d'une ontologie fournie par l'IGN)

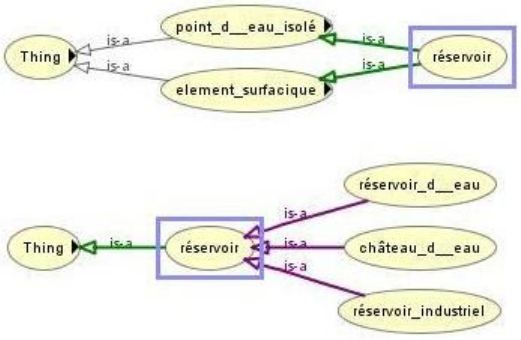


FIGURE 1.5 – Exemple de degré de granularité de la description d'une ressource

l'ontologie possède un nombre élevé de concepts présents dans le corpus, plus elle est considérée comme couvrante et donc de qualité.

Nous constatons que de nombreuses ressources sémantiques couvrant un même domaine ou des domaines connexes sont créées indépendamment les unes des autres. Il est illusoire de chercher à construire une ressource globale qui couvrirait différents domaines et serait adaptée à différentes applications. Il est néanmoins utile de réutiliser les ressources existantes et de les lier entre elles. Pour ce faire, l'hétérogénéité de contenu des ressources doit

être résolue. Cette problématique est reliée à l'interopérabilité sémantique entre ressources.

L'interopérabilité sémantique sert à donner « un sens aux informations échangées et s'assure que ce sens est commun dans tous les systèmes entre lesquels des échanges doivent être mis en œuvre » [Jouanot, 2000]. L'interopérabilité sémantique consiste donc à conserver le sens des connaissances partagées pour faciliter l'exploitation des ressources. A cette fin, une solution de l'interopérabilité sémantique est de créer des liens entre les ressources sémantiques, d'où l'alignement de ces ressources.

Nous détaillons dans ce qui suit les notions d'alignement et de correspondance. Nous présentons ensuite les méthodes d'alignement en nous appuyant sur : (1) les techniques lexicales et structurelles, (2) les ressources externes utilisées, et (3) l'application.

1.4 Alignement de ressources sémantiques

Afin de partager les connaissances entre différentes ressources sémantiques conçues indépendamment les unes des autres, il est nécessaire de dresser des ponts sémantiques entre ces ressources. Dans l'état de l'art, un ensemble d'approches et de techniques ont été proposées pour mettre en correspondance des ressources sémantiques. Ce processus de mise en correspondance est nommé alignement. Un état de l'art détaillé sur l'alignement est présenté dans [Euzenat *et al.*, 2004, Choi *et al.*, 2006] et différents outils ont été rendus disponibles grâce aux campagnes d'évaluation OAEI [11] (Ontology Alignment Evaluation Initiative).

Dans cette section, nous définissons les notions d'alignement et de correspondance et nous présentons ensuite les méthodes d'alignement de ressources sémantiques en détaillant les techniques standars d'alignement (terminologiques et structurelles), l'exploitation d'une ressource externe et l'importance de l'application visée.

1.4.1 Définitions

Deux notions sont utiles à rappeler : correspondance et alignement [12]. Nous reprenons les définitions de [Euzenat, 2004] dans le cadre de ce travail :

Définition 1 (correspondance) [13] La correspondance entre deux entités sémantiques de deux ressources différentes est une relation entre ces deux entités avec un indice de

11. `http://www.ontologymatching.org`

12. An alignment is « a set of mapping elements. The matching operation determines the alignment (A) for a pair of schemas/ontologies (o and o'). There are some other parameters which can extend the definition of the matching process, namely : (i) the use of an input alignment (A) which is to be completed by the process; (ii) the matching parameters, p (e.g., weights, thresholds) ; and (iii) external resources used by the matching process, r (e.g., thesauri) ».

13. A mapping element is « a 5-uple : $< id, e, e', n, R >$ where :
- id is a unique identifier of the given mapping element ;
- e and e' are the entities (e.g., tables, XML elements, properties, classes) of the first and the second schema/ontology respectively ;
- n is a confidence measure in some mathematical structure (typically in the [0,1] range) holding for the correspondence between the entities e and e' ;
- R is a relation (e.g., equivalence (=)) holding between the entities e and e'. »

confiance sur cette relation. Cet indice indique la fiabilité de la relation entre les entités à rapprocher. Soient deux ressources sémantiques RS_1 et RS_2, une correspondance m est composée d'un 5-uplet $< id, e, e', n, Rel >$ où :

- id est l'identifiant unique d'une correspondance donnée entre entités.
- e et e' sont deux entités respectivement des ressources RS_1 et RS_2.
- n est un indice de confiance exprimé dans l'intervalle [0,1].
- Rel est une relation entre les deux entités rapprochées e et e' (ex. relation d'équivalence).

Définition 2 (alignement) L'alignement entre deux ressources sémantiques RS_1 et RS_2 est un ensemble de correspondances entre ces deux ressources. C'est aussi un processus qui est appliqué sur deux ressources sémantiques, et qui fournit un ensemble de correspondances entre les deux ressources RS_1 et RS_2. Ce processus comporte différentes étapes définies par [Ehrig et Staab, 2004] (voir figure 1.6) et qu'on trouve généralement dans les méthodes d'alignement. La première étape consiste à définir les entrées du processus d'alignement. Dans cette même étape, les ressources sémantiques sont transformées en un formalisme commun. La deuxième étape permet d'identifier les couples d'entités de ressources sémantiques sur lesquelles l'alignement est appliqué, ce qu'on appelle les entités candidates. Dans cette étape, les méthodes d'alignement peuvent considérer un sous-ensemble d'entités à mettre en correspondance. La troisième étape repose sur le calcul de la similarité entre les couples d'entités rapprochés à l'étape 2. La quatrième étape permet d'agréger ou de combiner les mesures utilisées dans l'étape 3 pour obtenir une valeur globale de similarité. La cinquième étape consiste à interpréter les valeurs de similarité de l'étape précédente entre les entités mises en relation. Pour ce faire, l'interprétation repose sur soit un seuil préalablement fixé soit sur les critères de techniques d'alignement (ex. lexicale, structurelle). La dernière étape montre que ce processus d'alignement peut être itératif.

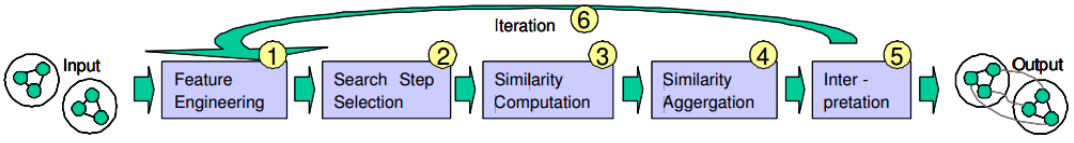


FIGURE 1.6 – Processus d'alignement selon [Ehrig et Staab, 2004]

En sortie de ce processus, nous obtenons un alignement qui peut être :

- alignement de type 1:1 : une entité de la première ressource est en relation avec une unique entité de la deuxième ressource ;
- alignement de type 1:m : une entité de la première ressource peut correspondre à plusieurs entités de la deuxième ressource ;
- alignement de type n:1 : plusieurs entités de la première ressource correspondent à une même entité de la deuxième ressource ;
- alignement de type n:m : plusieurs entités de la première ressource correspondent à plusieurs entités de la deuxième ressource.

Dans le but de mettre en correspondance des entités de ressources, des techniques d'alignement ont été proposées dans l'état de l'art. Nous détaillons dans la section suivante les techniques lexicales et structurelles sur lesquelles nous nous appuyons dans notre travail de thèse.

1.4.2 Techniques lexicales et structurelles

Le processus d'alignement s'appuie sur des techniques qui exploitent (i) les attributs des entités de ressources sémantiques, et (ii) la position de ces entités dans ces ressources. Ces techniques ont été détaillées dans [Euzenat et Shvaiko, 2007]. Dans cette section, nous exposons deux techniques d'alignement respectivement lexicales et structurelles.

1.4.2.1 Techniques lexicales

Ces techniques sont fondées sur la comparaison des chaînes de caractères [Euzenat et Shvaiko, 2007]. Elles sont appliquées pour mesurer la similarité entre les entités des ressources sémantiques ayant un volet lexical (les termes, les relations et les commentaires). Plus deux entités possèdent de caractères en commun, plus elles sont proches. Cette comparaison revient à calculer une similarité entre deux termes composés de chaînes de caractères. Différentes mesures ont été proposées, telles que :

- la distance de Levenshtein [Levenshtein, 1966] : c'est une distance d'édition qui permet de calculer la similarité entre deux chaînes de caractères ch_1 et ch_2. La distance $d(ch_1, ch_2)$ est le coût minimal des opérations élémentaires pour modifier ch_1 en ch_2. Cette distance est calculée en fonction des opérations suivantes qui ont chacune un coût : modifier un caractère de ch_1 en un caractère de ch_2, insérer dans ch_1 un caractère de ch_2 et supprimer un caractère de ch_1.
- la distance de Jaro-Winkler [Winkler, 1999] : c'est une distance de similarité qui étend entre deux chaînes de caractères la mesure proposée par [Jaro, 1989].
  La mesure de Jaro-Winkler est définie comme suit :

$$d_{jw} = d_j + (lp(1 - d_j))$$

  où :
  l : la longueur du préfixe commun, p : le cœfficient permettant de privilégier les chaînes de caractères ayant un $l > 0$, d_j : distance de Jaro entre les deux chaînes de caractères telle que :

$$d_j = \frac{1}{3}\left(\frac{m}{|ch_1|} + \frac{m}{|ch_2|} + \frac{m-t}{m}\right)$$

  $|ch_i|$: longueur de la chaîne de caractères ch_i,
  m : nombre de caractères identiques et
  t : nombre de caractères différents.

Afin de détecter la similarité entre deux termes, ces techniques s'appuient aussi sur la lemmatisation et l'analyse morpho-syntaxique.

1.4.2.2 Techniques structurelles

Ces techniques se fondent sur des informations structurelles et notamment sur la position des entités dans la hiérarchie des ressources sémantiques [Euzenat et Shvaiko, 2007] et les attributs de ces entités. Dans l'état de l'art, deux types de techniques structurelles ont été proposées :

1. **techniques liées à la structure interne des ressources** : ces techniques permettent d'exploiter la structure interne des entités des ressources sémantiques, composée du nom, du type, de la multiplicité, des restrictions [Valtchev, 1999, Rahm et Bernstein, 2001]. Les entités similaires de deux différentes ressources sont les entités qui partagent le plus de ces propriétés.
2. **techniques liées à la structure externe des ressources** : ces techniques permettent d'exploiter la structure externe des entités de ressources (leurs relations avec les entités de ressources) [Euzenat et Shvaiko, 2007]. Concrètement, les ressources sémantiques sont considérées comme des graphes orientés [Fürst et Trichet, 2006] où elles possèdent des nœuds et des arcs étiquetés avec les relations. La comparaison entre ressources est faite entre une ressource de référence (préalablement utilisée) et une autre ressource sémantique. Pour déduire la similarité entre deux entités, il suffit de s'appuyer sur leur voisinage comportant les ancêtres, les descendants et les frères. Différentes mesures, pour ces techniques, ont été proposées :
   - la mesure de Wu-Palmer [Wu et Palmer, 1994] : c'est une mesure qui calcule la similarité entre deux entités dans une hiérarchie d'entités sémantiques. La mesure est définie comme suit :

     $$Sim_{W\&P}(e_i, e_j) = \frac{2*depth(e_i,e_j)}{depthE(e_i)+depthE(e_j)}$$

     où :
     Les fonctions depth(X,Y) et depthE(X) retournent respectivement le nombre d'arcs séparant X de la racine et le nombre d'arcs séparant X de la racine en passant par E (E est le plus petit ancêtre commun de deux entités).
   - la mesure de Rada [Rada *et al.*, 1989] : c'est une distance entre deux entités dans une hiérarchie. Cette distance correspond au nombre minimum d'arcs à parcourir entre deux entités.

Ces deux techniques permettent de déterminer la similarité entre deux entités de ressources en s'appuyant sur leurs caractéristiques ainsi que leur position dans la ressource. La valeur de confiance attribuée à chaque correspondance correspond généralement à cette similarité.

D'autres méthodes d'alignement reposent sur des ressources externes pour mettre en relation les entités de ressources. Nous présentons dans la section suivante ces travaux.

1.4.3 Exploitation de ressources externes

Différents travaux permettent l'alignement de ressources sémantiques en utilisant une ressource externe. On distingue deux types de ressources externes : (1) structurées : une ressource lexicale telle que WordNet et une ou plusieurs ontologies, et (2) non-structurées : le texte.

Utilisation de WordNet WordNet[14] est une ressource lexicale généraliste composée d'ensembles de termes synonymes lemmatisés dits « synsets ». Ces synsets sont reliés par des relations hiérarchiques et des relations de composition.

Le travail de [Bach *et al.*, 2004] permet d'exploiter WordNet pour aligner deux ontologies. La méthode proposée dans ce travail permet de calculer la similarité entre deux concepts d'ontologies en comparant leurs étiquettes associées aux synonymes des synsets dans WordNet. Deux concepts sont d'autant plus similaires qu'ils partagent plus de synonymes. Un autre travail de [Kwak et Yong, 2010] propose pour calculer la similarité entre deux concepts d'ontologies, une méthode fondée sur « the Super Word Set Similarity ». Ce dernier représente un ensemble agrégé de relations sémantiques entre les synsets dans WordNet (ex. hyperonymie, hyponymie, méronymie).

Le travail de [Kefi *et al.*, 2006] s'appuie sur la hiérarchie de WordNet pour aligner deux taxonomies ayant des concepts et des étiquettes associées. L'alignement proposé est un alignement orienté. La taxonomie source est très peu structurée et la taxonomie cible est plus structurée. L'objectif de ce travail est d'unifier l'accès et l'interrogation des bases de données hétérogènes d'un domaine donné. Le processus d'alignement de TaxoMap comporte trois étapes [Kefi *et al.*, 2006] : (1) normalisation des étiquettes des concepts : cette étape consiste à unifier les étiquettes des concepts en les associant aux racines et aux lemmes, (2) calcul de similarité : cette étape consiste à identifier pour un concept de la taxonomie source, les concepts candidats de la taxonomie cible et cela en se fondant sur la mesure de similarité de Lin [Lin, 1998] calculée sur les étiquettes, (3) application des techniques d'alignement : cette étape repose sur l'exploitation des correspondances fournies par la mesure de similarité de l'étape précédente et révise les concepts des taxonomies qui n'ont pas été mis en correspondance [Safar et Reynaud, 2009]. Cette dernière étape repose sur l'utilisation de WordNet. Il s'agit dans un premier temps d'extraire un sous-arbre de WordNet en identifiant un concept racine de WordNet, ce concept permet de générer par la suite tous les concepts de la taxonomie cible pertinents pour le domaine. La deuxième étape permet de rechercher les entrées de WordNet correspondant aux concepts identifiés dans l'étape précédente et le concept racine en suivant les hyperonymes de concepts jusqu'à atteindre la racine. Les correspondances qui sont identifiées sont de type « isA ». L'alignement fourni est de type 1:m.

Utilisation d'une ou plusieurs ontologies Dans le travail de [Quix *et al.*, 2011] la méthode d'alignement GeRoMeSuite, permet de rapprocher les concepts d'une ontologie source (S) des concepts d'une ontologie cible (T), en s'appuyant sur une ontologie « support » O (voir figure 1.7).

Un alignement direct est réalisé A_{dir} entre les deux ontologies O et S en utilisant la similarité des chaînes de caractères des étiquettes de concepts ou la similarité des propriétés des concepts. Deux alignements sont établis dans cette méthode : (1) $A_{O,S}$, l'alignement entre les deux ontologies O et S, et (2) $A_{O,T}$, l'alignement entre les deux ontologies O et T. Ces deux alignements reposent sur des techniques structurelles et lexicales. En raisonnant sur la sortie de $A_{O,S}$ et $A_{O,T}$ et plus spécifiquement sur les relations de subsomption et d'équivalence, les concepts des deux ontologies O et S sont rapprochés avec des relations

14. `http://wordnet.princeton.edu/`

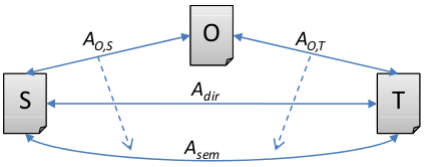


FIGURE 1.7 – Méthode d'alignement de [Quix *et al.*, 2011]

sémantiques A_{sem}.

Le travail de [Sabou et Motta, 2006] a pour but d'aligner des ontologies en s'appuyant sur plusieurs ontologies extraites du moteur de recherche Swoogle [15]. Deux méthodes sont proposées pour aligner deux ontologies source et cible (voir figure 1.8). La première consiste à aligner, en utilisant une méthode lexicale, les concepts candidats de chacune des ontologies (source et cible) et les concepts des ontologies considérées comme une ressource externe. Les correspondances qui sont identifiées sont de type équivalence. La deuxième méthode permet de déduire, tout d'abord, des relations sémantiques (hiérarchiques et équivalence) entre les ontologies et les ressources externes et une méthode lexicale est aussi appliquée pour rechercher les relations d'équivalence entre les concepts d'ontologies source et cible et les concepts de la ressource externe. Enfin, l'alignement entre les deux ontologies est déduit en appliquant un ensemble de règles d'inférence sur les correspondances obtenues dans les étapes précédentes.

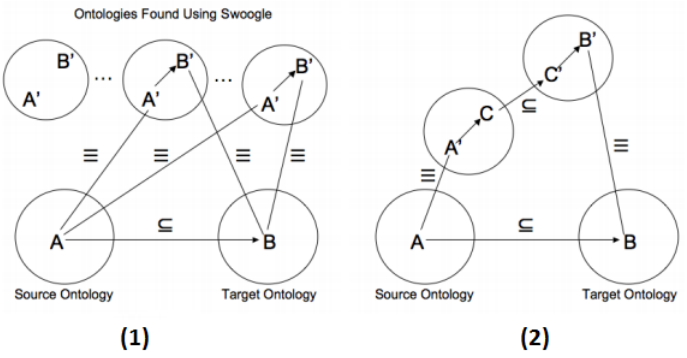


FIGURE 1.8 – Alignement en utilisant plusieurs ontologies [Sabou et Motta, 2006]

Utilisation de textes Le corpus textuel est composé d'un ensemble de documents spécifiques à un domaine. Ce corpus est construit soit d'une manière manuelle soit d'une manière automatique en utilisant un moteur de recherche existant. A notre connaissance, il existe très peu de travaux qui cherchent à aligner des ressources sémantiques en se fondant sur le texte. Nous avons répertorié quatre travaux d'alignement qui exploitent le texte comme une base de connaissances externe, comportant les instances associées aux concepts d'ontologies. Le calcul de similarité dentre deux concepts est effectué en fonction de la fréquence d'apparition des mots associés aux termes dénotant les concepts dans leur contexte (ex.

15. `http://swoogle.umbc.edu/`

phrase, paragraphe).

Il existe deux techniques des comparaisons de la distribution des mots dans un contexte donné. La première est statistique. Elle a été proposée par [Cheng *et al.*, 2008] et elle consiste à appliquer deux mesures de similarité : (1) mesure du cosinus, qui permet de chercher la similarité entre deux concepts en comparant les vecteurs de contextes, et (2) mesure de Jaccard, qui permet de comparer l'intersection des instances des concepts à aligner dans le texte, par rapport à l'union des instances de chaque concept. Dans son travail, [Kassaie *et al.*, 2012] utilise la mesure de Jaccard calculée en utilisant la deuxième technique d'apprentissage fondée sur la distribution des termes. Cette mesure est exprimée en fonction du nombre de documents où les concepts à aligner apparaissent ensemble par rapport au nombre de documents où chaque concept apparait seul. Dans le même contexte, le travail proposé par [Isaac *et al.*, 2007] utilise la même définition de la mesure Jaccard que dans [Cheng *et al.*, 2008] pour aligner deux thésaurus. Une première étape de la méthode de [Isaac *et al.*, 2007] repose sur des occurrences de descripteurs (préférentiels et non-préférentiels) dans le corpus textuel. La deuxième étape consiste à calculer les annotations communes entre les descripteurs des deux thésaurus.

Le travail de [Kingkaew, 2012] permet d'utiliser les techniques lexicales pour aligner deux ontologies. La méthode d'alignement proposée dans ce travail est composée de trois étapes : (1) étape de recherche d'information : le but de cette étape est de chercher des documents sur le Web, (2) étape de traitement automatique du langage naturel (TAL) : cette étape consiste à découper les textes en phrases et à établir l'étiquetage morpho-syntaxique de ces phrases. Un graphe de similarité est créé où chaque mot est associé à la phrase analysée grammaticalement, (3) étape d'alignement : cette étape consiste à appliquer la mesure de similarité lexicale de Jaro-Winkler entre le graphe de similarité et chacune des ontologies (source et cible) qui sont elles-mêmes considérées comme deux graphes.

Ces travaux montrent l'intérêt d'utiliser le texte pour chercher des correspondances entre les ressources sémantiques. Cependant, le texte est utilisé comme une base « d'instances » associées aux termes de concepts d'ontologies ou aux descripteurs de thésaurus. Les travaux d'alignement permettent souvent d'obtenir des relations d'équivalence sans pour autant se concentrer ni sur : (i) la richesse des relations entre les mots du texte, (ii) les relations que peuvent entretenir les entités de ressources, et (iii) l'interprétation des relations entre une paire d'entités associées à des unités textuelles.

Les méthodes d'alignement proposées visent une application donnée. Dans la section suivante, nous présentons le rôle de l'application dans l'alignement.

1.4.4 Rôle de l'application

L'application joue un rôle essentiel dans le processus d'alignement. Elle permet d'influencer la sortie de ce processus pour qu'elle soit adaptée le plus possible aux besoins.

Le travail de [Zhdanova et Shvaiko, 2006] permet d'éliminer manuellement des correspondances fournies pour construire un service Web comportant les ontologies intéressantes pour une communauté ayant un intérêt commun.

L'application visée dans le travail de [Kefi *et al.*, 2006] où ils proposent l'outil Taxo-

Map, est d'interroger des bases de données distribuées. Cette application s'appuie sur une taxonomie support (taxonomie cible) et une taxonomie source. La taxonomie support est créée pour unifier l'accès aux différentes bases de données. Dans d'autres travaux, l'alignement permet de faire communiquer et échanger des informations entre deux agents dont chacun possède ses spécificités [Sampson, 2005] à savoir sa capacité d'agir, son autonomie, etc. Afin d'obtenir une communication fiable entre les deux agents, la méthode d'alignement utilisée prend en compte les caractéristiques de ces agents pour rapprocher les concepts des ontologies. L'alignement est aussi intéressant pour répondre aux requêtes des utilisateurs d'une manière efficace dans les services Web des entreprises [Jin *et al.*, 2009]. Ces services sont exprimés par des ontologies hétérogènes. Pour échanger les données entre ces services, il est nécessaire d'établir des correspondances entre les ontologies de chaque service. Cet alignement sera influencé par les besoins des utilisateurs des ces services Web.

On peut conclure que l'alignement de ressources sémantiques est souvent dépendant de l'application visée. Or, les travaux d'alignement ne font appel à l'application visée qu'à la fin du processus d'alignement et cela ne devient intéressant que si l'application est considérée dès le début du processus.

Une des manières de représenter les notions d'un domaine d'intérêt est la « cartographie ». Un état de l'art sur la cartographie et des notions proches est présenté dans la section suivante.

1.5 La cartographie dans la littérature

L'objectif de notre travail ne se résume pas à aligner des ressources sémantiques disponibles mais est plutôt de caractériser le domaine d'intérêt et de représenter l'information utile pour l'ingénieur de la connaissance. Pour cela, nous avons besoin de réviser la sortie de l'alignement dans le but de présenter les liens nécessaires à l'ingénieur de la connaissance.

Dans l'état de l'art, les informations utiles d'un domaine d'intérêt sont présentées sous forme de deux notions similaires « cartographie » ou « carte ». Nous empruntons la terminologie du comité français de cartographie [16] pour la définition générale de ces deux notions, comme suit :

- la cartographie est « l'ensemble des études et des opérations scientifiques, artistiques et techniques, intervenant à partir des résultats d'opérations directes ou de l'exploitation d'une documentation, en vue de l'élaboration et de l'établissement de cartes, plans et autres modes d'expression, ainsi que dans leur utilisation ».
- la carte est « une représentation géométrique conventionnelle, généralement plane, en positions relatives, de phénomènes concrets ou abstraits, localisables dans l'espace ; c'est aussi un document portant cette représentation ou une partie de cette représentation sous forme d'une figure manuscrite, imprimée ou réalisée par tout autre moyen ».

La cartographie est donc considérée comme un processus permettant de réaliser une carte selon les deux définitions précédentes. [Tricot et Roche, 2004] proposent un type nouveau de cartographie nommé une cartographie sémantique. Elle est définie comme « un espace informationnel d'une organisation répondant aux besoins de naviguer selon la sémantique

16. `http://www.lecfc.freesurf.fr`

du domaine, de proposer une vision à plusieurs échelles et de proposer une carte adaptée à l'utilisateur. Ainsi, elle devient une activité essentielle à la gestion des connaissances, permettant de tirer partie de toute la richesse des informations de l'organisation ». La cartographie sémantique a donc pour but de comprendre et de faciliter aux organisations l'étude et la conception de leurs informations. Le processus de construction de la cartographie sémantique, défini par [Tricot et Roche, 2004], est composé d'un ensemble d'opérations dont le but est de donner accès à des documents d'une manière structurée :

- structurer l'espace informationnel brut en utilisant des méthodes statistiques d'analyse des données telles que le clustering ;
- représenter l'espace informationnel structuré avec un modèle de connaissances (ex. ontologie) ;
- visualiser l'espace informationnel représenté avec un outil donné (ex. EyeTree) sous forme d'une carte.

Après la modélisation des informations du domaine d'intérêt, les données de la cartographie sémantique peuvent être visualisées dans une carte. La figure 1.9 montre la carte résultant de la cartographie des informations autour de la notion « condition de fonctionnement » (28 documents ont été retrouvés).

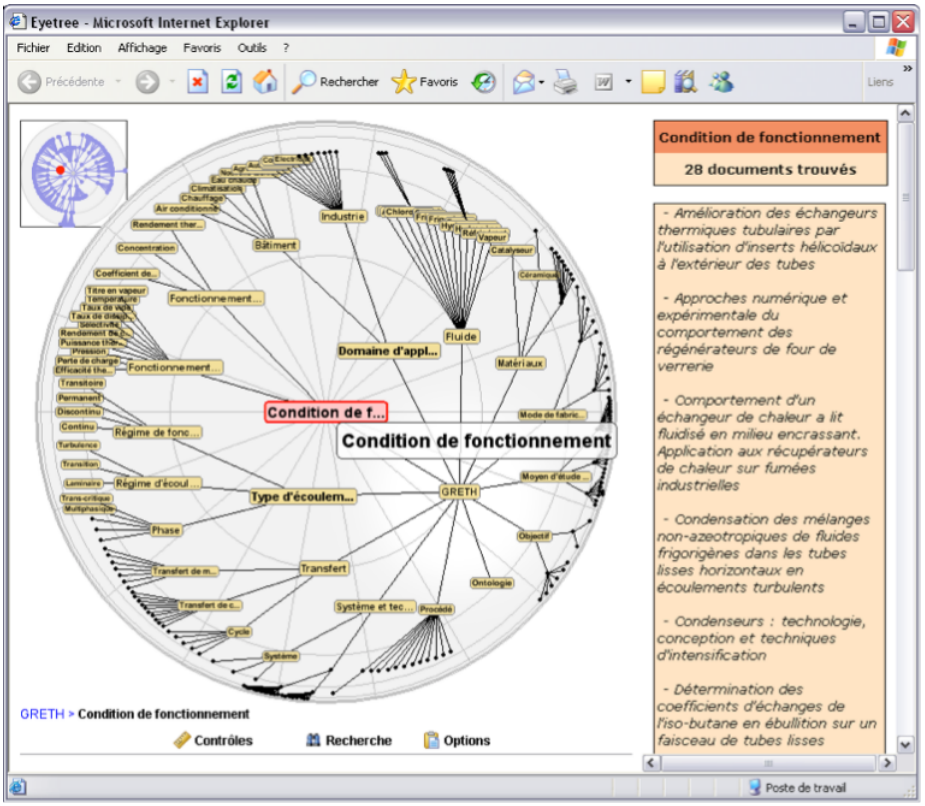


FIGURE 1.9 – Visualisation sous forme d'une carte les informations reliées à la notion « condition de fonctionnement » [Tricot et Roche, 2004]

Dans le même contexte de la construction d'une cartographie, LOV [17] (Linked Open Vocabularies) a été défini par [Vandenbussche *et al.*, 2011] comme « une initiative qui recense les vocabulaires utilisables et communément utilisés pour décrire les données ouvertes sur le Web et les relations qui les lient ». LOV (voir figure 1.10) est donc une base de vocabulaires qui sont indépendamment publiés et reliés entre eux. L'alimentation de LOV est effectuée suite à une mise à jour manuelle du répertoire de vocabulaires par les utilisateurs. Ce travail a pour but de faciliter la recherche des vocabulaires décrivant les données d'un domaine.

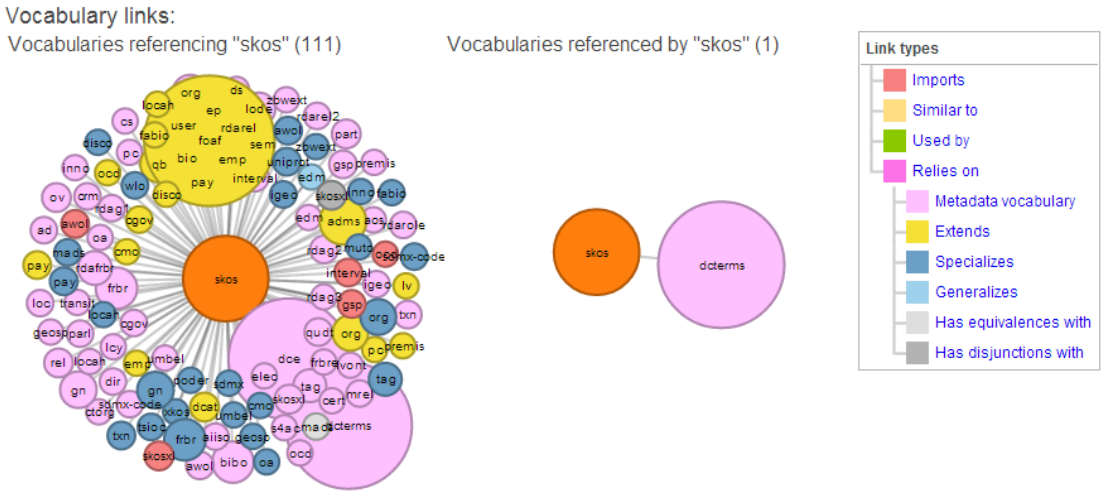


FIGURE 1.10 – Copie d'écran de LOV montrant le langage formel « skos » et ses relations

De nombreux types de cartes ont été proposés dans l'état de l'art à savoir la carte conceptuelle, la carte thématique et la carte heuristique ou mentale. La carte conceptuelle est définie par [Novak et Cañas, 2006] comme « une représentation graphique des connaissances ». Elle est composée de concepts et des liens orientés et étiquetés entre eux. Cette carte a pour but d'organiser les connaissances d'un sujet particulier. De nombreux outils ont été proposés pour créer ce type de cartes à savoir Lucidchart [18], Visual Understanding Environment [19]. Quant à la carte thématique, elle est définie selon la norme ISO/IEC 13250 [20] comme « une technologie de codage de connaissances permettant de relier ces connaissances encodées aux ressources d'information pertinentes. Les *Topics Maps* sont organisées autour de Topics, qui représentent des sujets du monde du discours, des associations représentant les relations entre les sujets et des occurrences qui permettent de relier les sujets aux ressources d'informations pertinentes ». Une carte thématique permet donc de structurer les connaissances d'un domaine selon des sujets et des relations entre eux. Pour construire une carte thématique, [Ellouze *et al.*, 2012] propose une approche fondée

17. `http://lov.okfn.org/dataset/lov/`
18. `https://www.lucidchart.com/`
19. `http://vue.tufts.edu/`
20. `http://www.isotopicmaps.org/sam/sam-model/` : « Topic Maps is a technology for encoding knowledge and connecting this encoded knowledge to relevant information resources. Topic maps are organized around topics, which represent subjects of discourse; associations, representing relationships between the subjects; and occurrences, which connect the subjects to pertinent information resources. »

sur plusieurs documents textuels multilingues et prend en compte l'évolution de la carte selon les requêtes posées par les utilisateurs. Parmi les outils qui permettent de créer ces cartes, on trouve Topic Map Designer [21] et K42 [22]. La carte heuristique ou mentale est définie dans [Deladrière et Kilian, 2009] comme « une représentation graphique d'idées ». Elle a pour but d'analyser un problème donné. Freeplane [23] est l'un des outils qui permettent de créer les cartes heuristiques.

La cartographie que nous proposons n'a pas pour vocation à être une nouvelle ressource sémantiquement cohérente mais elle doit préserver les ressources sur lesquelles elle s'appuie, qui ont été publiées et qui ont *a priori* chacune leur cohérence propre : elle doit montrer à la fois ce qui rapproche les ressources disponibles et ce qui les distingue.

1.6 Positionnement et conclusion

Nous avons présenté, dans ce chapitre, un aperçu des différents champs de recherche sur lesquels s'appuie notre travail. Nous avons mis l'accent, tout d'abord, sur la définition des différents types de ressources sémantiques en soulignant leur hétérogénéité. Nous avons notamment souligné le fait qu'il y a des différences de nature dans les connaissances qu'elles contiennent (terme *vs.* concept), une granularité de description variable et souvent une couverture différente des domaines modélisés.

Dans cette thèse, nous avons choisi de mettre l'accent sur un seul type de ressources, les ontologies lexicalisées, que nous considérons comme la clef de voûte en Web sémantique parce qu'elles représentent d'une manière explicite les informations du Web et qu'elles permettent d'annoter les documents textuels. Dans ses grandes lignes, notre approche peut toutefois s'appliquer à des ressources moins riches comme des thésaurus ou des terminologies.

L'alignement de ressources a été proposé comme une solution à l'interopérabilité sémantique et il occupe une place centrale dans la gestion des ontologies. Il permet de calculer des correspondances (relation d'équivalence ou hiérarchique, par ex.) entre des ressources *a priori* disjointes. Ce sont généralement des techniques lexicales et structurelles qui sont utilisées : elles reposent sur le calcul de la similarité entre les étiquettes ou les positions des concepts des ontologies à aligner. L'alignement est également souvent guidé par une ressource extérieure : un thésaurus, une autre ontologie et plus rarement des textes.

Pour la construction de cartographies de domaine, nous souhaitons adapter l'alignement au domaine visé et nous considérons des textes comme ressources externes. Nous supposons en effet que l'ingénieur de la connaissance peut fournir un texte représentatif du domaine et de la perspective auxquels il s'intéresse et que celui-ci va influer sur la cartographie produite. Le texte joue un triple rôle dans le processus de construction d'une cartographie de domaine :

1. il peut servir en premier lieu à sélectionner les ressources candidates mais nous n'abordons pas dans cette thèse cette première étape de sélection des ressources ;

21. `http://www.topicmap-design.com/`
22. `http://k42.empolis.co.uk`
23. `http://freeplane.sourceforge.net`

2. il sert également à pondérer les correspondances trouvées par des méthodes lexicales et structurelles d'alignement, ce qui permet de les adapter au domaine visé ;
3. il permet enfin de détecter de nouvelles correspondances – textuelles ou distributionnelles – que les méthodes d'alignement traditionnelles ne permettent pas de retrouver.

Pour exploiter le texte, nous avons besoin d'ontologies lexicalisées qui articulent les deux niveaux lexical et conceptuel. Le processus d'alignement que nous proposons est présenté dans le chapitre 3.

Dans un contexte où l'interopérabilité sémantique devient un enjeu majeur, nous proposons en réalité un nouvel outil pour la gestion des ressources sémantiques. Les cartographies de domaine que nous cherchons à construire doivent permettre à un ingénieur de prendre connaissance de l'éventail de ressources disponibles pour un domaine particulier, de leurs différences et de leurs parentés. Cette thèse propose une méthode pour construire d'une manière automatique de telles cartographies de domaine à partir d'un corpus textuel.

Le chapitre suivant présente une vue d'ensemble de cette méthodologie de construction de cartographies de domaine.

Chapitre 2

Méthodologie de construction de la cartographie de domaine

Sommaire

2.1 Introduction

Avant de choisir la ressource à utiliser dans une nouvelle application ou de décider de construire une nouvelle ressource sémantique, il faut prendre connaissance des ressources existantes pour le domaine auquel on s'intéresse. C'est en effet souvent en combinant des « bouts » d'ontologies, de terminologies et de thesaurus que l'on peut avoir une ressource adaptée à la nouvelle application cible.

L'objectif de ce travail n'est donc pas de produire de nouvelles connaissances mais plutôt de fournir un support qui permette à l'utilisateur de prendre connaissance des ressources existantes sur le domaine qui l'intéresse.

L'hétérogénéité des ressources sémantiques signalée dans le chapitre 1 tient à la nature des connaissances qu'elles comportent (connaissances terminologiques dans les thesaurus et dictionnaires ou conceptuelles dans les ontologies et taxonomies), au fait qu'elles couvrent plus ou moins largement le domaine de spécialité visé, qu'elles en donnent une description générale pour certaines (ex. Eurovoc) et spécialisée pour d'autres (ex. l'ontologie Kaon décrivant les plantes : fleurs, couleur, longueur, etc). Cette hétérogénéité nuit à l'objectif d'interopérabilité du web sémantique.

L'approche que nous proposons tente de compenser cet handicap. L'objectif est de tirer parti de la richesse et de la diversité des ontologies existantes pour un domaine mais aussi de l'organiser en proposant une « cartographie de domaine » qui donne une vue d'ensemble des ressources disponibles en les articulant les unes par rapport aux autres. Étant donné un ensemble de ressources relatives au domaine visé, la cartographie de domaine que l'on cherche à construire se présente comme un alignement entre ces ressources – soit

un ensemble de correspondances entre les entités qui les composent – et un ensemble des zones remarquables qui montrent les écarts de conceptualisation et les points de jonction entre les ressources alignées.

La méthodologie que nous proposons permet de construire de telles cartographies de domaine à partir d'un ensemble de ressources préexistantes et d'un texte représentatif du domaine. Le processus de construction se décompose en plusieurs phases :

- une première phase d'annotation sémantique permet de lier les ressources au texte en y projetant les entités ontologiques, ce qui permet aussi de repérer celles qui sont « ancrées » (présentes dans le texte) ;
- une deuxième phase d'alignement guidé par le texte permet alors de rapprocher les entités des différentes ressources tout en privilégiant les correspondances qui sont corroborées dans le texte : le texte sert alors de support pour l'alignement et l'on tient compte des contextes (en pratique, les phrases) dans lesquelles figurent les entités à aligner ;
- la troisième phase construit à proprement parler la cartographie du domaine en identifiant dans les sorties d'alignement les configurations remarquables qui montrent comment les ressources alignées se positionnent l'une par rapport à l'autre.

Ce chapitre présente une vue d'ensemble de la méthodologie proposée. Il comporte 1) une revue sur les ressources sémantiques à exploiter, (2) une présentation de la méthodologie de construction la cartographie de domaine, et (3) un exemple montrant ce qu'on obtient à chaque phase de notre méthodologie.

2.2 Ressources sémantiques à exploiter

Rappelons qu'une ressource sémantique est un modèle de connaissances défini par des entités (ex. concept, terme, descripteur, instance, propriété) et des relations qu'elles entretiennent entre elles (ex. relation hiérarchique, relation associative). Nous distinguons deux types de ressources en fonction de la nature des connaissances (*cf.* chapitre 1). Les ressources conceptuelles (ex. ontologie, taxonomie) décrivent le domaine sur lequel elles portent sous la forme de concepts et de relations conceptuelles. D'autres ressources décrivent le domaine sous la forme d'unités lexicales, nous parlons alors de ressources terminologiques (ex. terminologie, glossaire). Il existe aussi des ressources qui font l'articulation entre les niveaux lexical et conceptuel, ces ressources sont appelées ontologies lexicalisées. Dans notre travail, nous nous intéressons à la dernière catégorie de ressources.

Nous avons choisi dans notre travail le texte comme un support décrivant le domaine et la perspective qui intéresse l'ingénieur de la connaissance. Pour faciliter l'exploitation du texte, il est nécessaire de faire le lien entre le volet conceptuel et lexical. C'est le rôle des ontologies lexicalisées ou termino-ontologies O_{lex}. Une ontologie lexicalisée est définie dans notre travail par $O_{lex} = \{C, RC\}$, où :

- C : ensemble de concepts décrivant un domaine donné. Un concept est défini par un identifiant unique, un ensemble d'étiquettes désignant des termes qui partagent des relations lexicales (ex. synonymie).
  Un concept doit avoir au minimum une étiquette (ou label) pour qu'il soit considéré comme étant « lexicalisé » ;

– *RC* : ensemble de relations entre les concepts. Ces relations sont de deux types : hiérarchiques et non-hiérarchiques (rôles) ;

La taxonomie dans ce contexte est considérée comme étant une ontologie allégée.

La section 2.3 présente la méthodologie que nous proposons. Elle présente les problématiques auxquelles la méthodologie doit répondre. Nous décrivons ensuite les phases de la méthodologie qui permettent d'obtenir une cartographie de domaine (que nous détaillons par la suite). La cartographie a pour objectif d'aider l'ingénieur de la connaissance à analyser le domaine d'intérêt, à s'y positionner mais également de filtrer les connaissances inutiles pour éviter de surcharger la représentation et pour garder une vision d'ensemble.

2.3 Présentation de la méthodologie

Notre but est d'assister l'ingénieur de la connaissance pour analyser les ontologies existantes les unes par rapport aux autres en exploitant la richesse et la diversité de ces ontologies pour les articuler entre elles. Il s'agit de faciliter la réutilisation des ontologies pour mieux présenter les notions d'un domaine d'intérêt. Ces ontologies sont hétérogènes. Afin de capturer les connaissances partagées entre ces ontologies et de viser l'intéropérabilité sémantique, notre méthodologie repose sur les informations textuelles relatives au domaine de spécialité. Le texte est choisi comme un support décrivant le centre d'intérêt de l'ingénieur et sert à sélectionner des ontologies et à les aligner les unes par rapport aux autres.

Il existe différentes façons de constituer un corpus textuel [Lame, 2002] soit en interrogeant le Web par des requêtes spécifiant le domaine [Koo *et al.*, 2003] soit en recueillant les connaissances dans des interviews mais dans notre travail nous supposons que le texte est déjà construit et possède de bonnes propriétés comme par exemple de couvrir les notions importantes du domaine à modéliser.

La méthodologie proposée repose sur l'exploitation de la richesse et de la diversité des ontologies en préservant la cohérence propre à chacune et en les articulant entre elles. Les phases de notre méthodologie sont au nombre de trois (voir figure 2.1 pour le cas de deux ontologies). La première phase est l'annotation. Elle vise à établir des liens entre les ontologies et le texte.

La deuxième phase est celle de l'alignement guidé par le texte. Cette phase permet de rapprocher les entités de plusieurs ontologies en s'appuyant sur le texte. La méthode proposée dans cette phase est automatique. Nous récupérons toutes les correspondances possibles entre les entités qui sont suffisamment validées par le texte. Cet alignement est donc de type n:m. La sortie d'alignement comporte deux types de correspondances : association sémantique et équivalence sémantique. Ces relations sont déduites à partir des relations entre les termes dans le texte. La dernière phase est celle de la construction de la cartographie. Cette phase consiste à analyser et réviser les liens entre les ontologies dans le but de présenter un ensemble de liens cohérent à l'ingénieur de la connaissance. Dans cette phase, un certain nombre de problèmes et de correspondances remarquables sont repérées. L'objectif de cette phase est de guider l'exploitation des correspondances entre les entités d'ontologies.

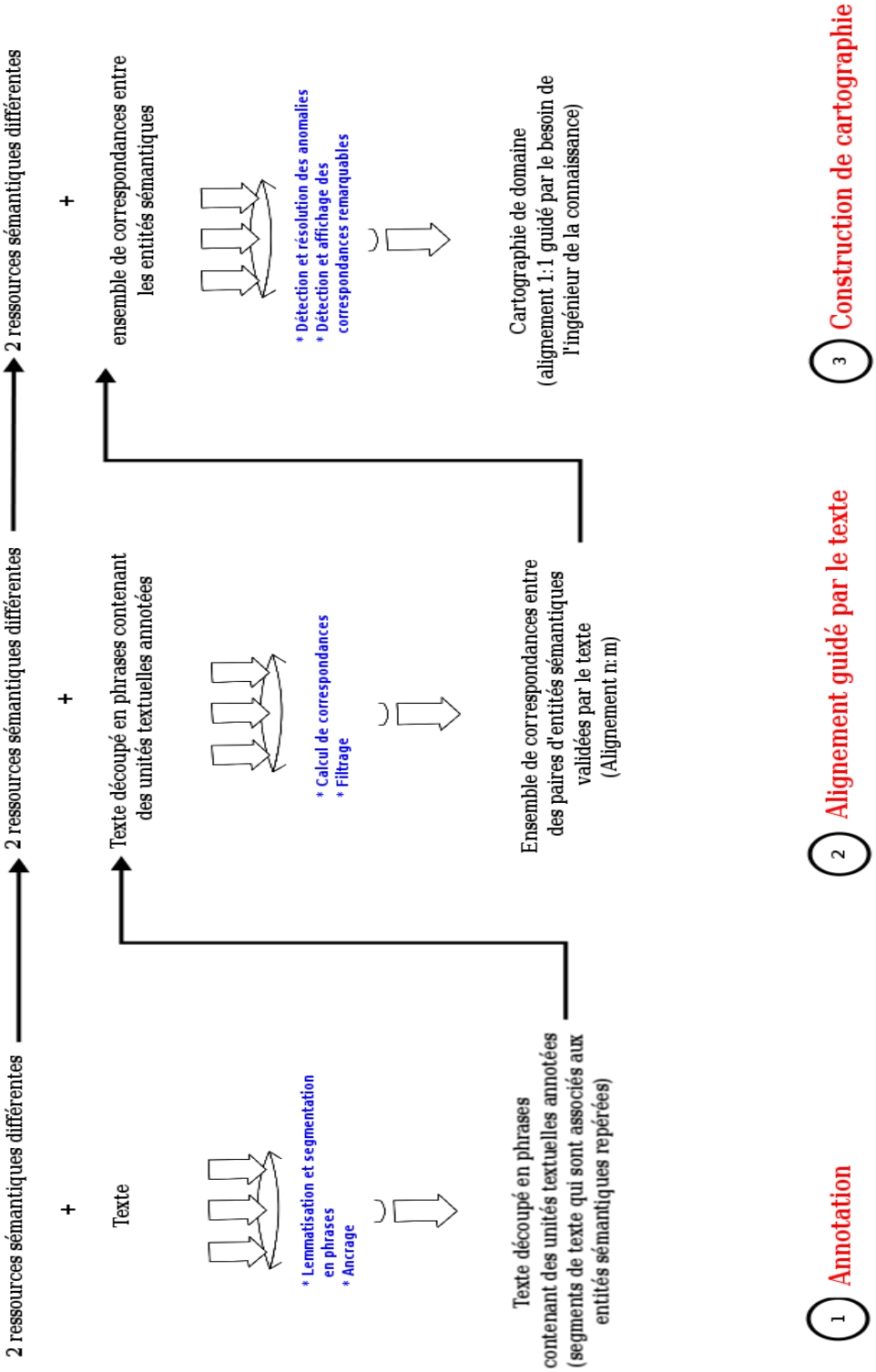


FIGURE 2.1 – Méthodologie de construction d'une cartographie à partir de ressources sémantiques

2.3.1 Phase d'annotation

Cette phase permet de lier le texte à une ontologie. Un certain nombre de travaux se sont penchés sur ce problème d'annotation de texte. L'annotation sémantique est définie dans [Amardeilh et Francart, 2006] comme « une représentation formelle d'un contenu exprimé à l'aide de concepts, relations et instances décrits dans une ontologie et reliés au document ». Le processus d'annotation comporte trois phases dans [Desmontils et Jacquin, 2002]. Une première phase sert à repérer des références de concepts de l'ontologie dans le document. Une deuxième phase consiste à instancier les attributs de concepts en les caractérisant par les informations du document. Une troisième phase permet d'ajouter aux références de concepts des informations non explicitement présentes dans le document et cela à travers les attributs des concepts. L'annotation dans l'état de l'art est utilisée dans le but d'(i) améliorer les systèmes de recherche d'information [Kiryakov *et al.*, 2004], (ii) extraire des informations pour enrichir ou peupler une ontologie [Aussenac-Gilles *et al.*, 2013] et (iii) explorer la sémantique des documents comme les textes médicaux [Ben-Abacha et Zweigenbaum, 2010], les textes décrivant des réglementations métier [Ma *et al.*, 2013].

Le processus d'annotation peut être réalisé d'une manière manuelle [Handschuh et Staab, 2003] suite à la lecture d'un document. L'utilisateur sélectionne un document ou un passage du document et précise l'annotation (ex. commentaire, correction, méta-donnée). Le processus d'annotation peut être aussi effectué d'une manière semi-automatique [Kahan et Koivunen, 2001] en apprenant des annotations définies manuellement. L'autre manière d'établir l'annotation est automatique par l'intermédiaire par exemple des patrons reposant sur des expressions régulières préalablement définies [Dingli *et al.*, 2003].

Nous mettons en place, dans un premier temps, une annotation triviale où le principe est de comparer deux chaînes de caractères (terme relatif à un concept avec un mot lemmatisé du texte). Nous supposons que le problème d'ambiguïté est résolu. La phase d'annotation dans notre travail, consiste à projeter les concepts des ontologies sur le texte. En d'autres termes, cette phase consiste à lier les entités sémantiques d'ontologies avec les unités textuelles correspondantes. Une entité sémantique est représentée par deux liens : son lien vers le texte et sa représentation dans l'ontologie. En d'autres termes, une entité sémantique est un concept d'ontologie présent dans le texte. Son lien vers l'ontologie est exprimé par un identifiant uniforme de ressource (URI) qui la lie à une ontologie et son lien vers le texte est exprimé par un identifiant textuel (offset) qui la lie au texte.

Une unité textuelle est exprimée dans notre travail par une unité lexicale (simple ou composée) constituant le texte.

La figure 2.2 représente 15 concepts dans une ontologie, parmi lesquels 5 entités sémantiques sont utilisées dans le texte grâce aux termes qui les dénotent. Dans le texte, il existe 300 unités textuelles dont 80 unités textuelles correspondent aux concepts. Nous obtenons donc 80 couples qui sont constitués de l'entité sémantique ES et l'unité sémantique correspondante UT.

Une phase d'annotation prend en entrée les deux ontologies lexicalisées et le texte de référence et fournit en sortie un texte découpé en phrases contenant des unités textuelles

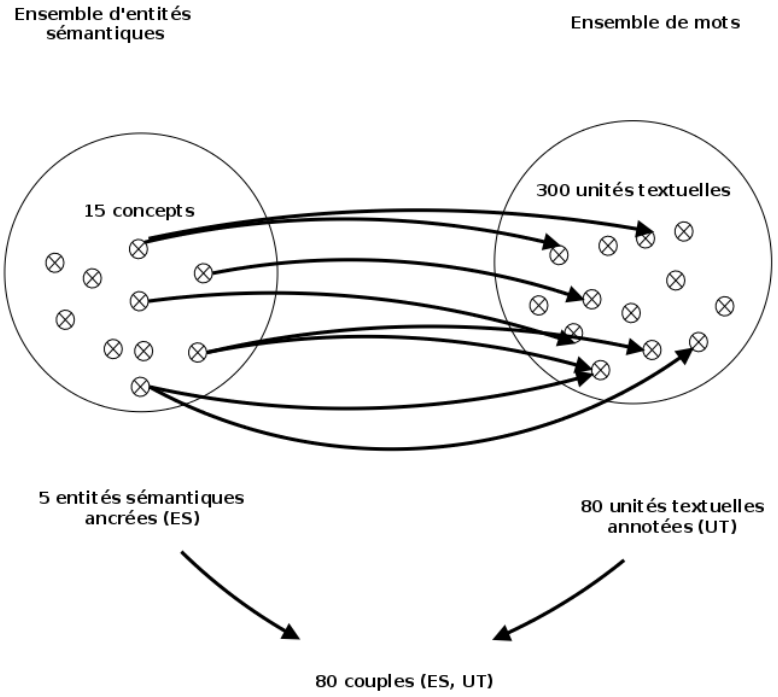


FIGURE 2.2 – Principe d'annotation dans notre méthodologie de construction de la cartographie de domaine

qui sont associées aux entités sémantiques repérées. Cette phase d'annotation comporte deux étapes : (1) la lemmatisation et la segmentation en phrases, et (2) l'ancrage. La première étape consiste à segmenter le texte en phrases et à appliquer l'étiqueteur morpho-syntaxique Treetagger [1] pour associer un lemme à chaque unité textuelle. Tous les mots du texte sont donc lemmatisés dans notre travail. Prenons la phrase suivante « Endothelial cells are also stimulated to grow and divide by direct contact with bacterial cells », en appliquant Treetagger, nous obtenons la lemmatisation de la phrase présentée dans la figure 2.3.

Endothelial	JJ	endothelial
cells	NNS	cell
are	VBP	be
also	RB	also
stimulated	VBN	stimulate
to	TO	to
grow	VB	grow
and	CC	and
divide	VB	divide
by	IN	by
direct	JJ	direct
contact	NN	contact
with	IN	with
bacterial	JJ	bacterial
cells	NNS	cell
.	SENT	.

FIGURE 2.3 – Exemple d'une phrase lemmatisée par Treetagger

1. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

La deuxième étape consiste à comparer les lemmes des mots aux lemmes des étiquettes associées aux concepts des ontologies. La figure 2.4 montre un exemple de texte annoté. Deux couleurs différentes (rouge et bleu) sont présentées dans le texte ; elles représentent les entités ancrées de deux ontologies différentes.

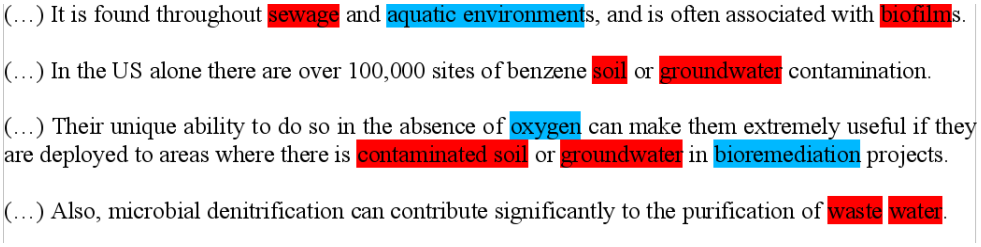


FIGURE 2.4 – Exemple de texte annoté par les ontologies de la figure 2.5

Le processus d'annotation que nous mettons en place peut être remplacé par un outil d'annotation plus performant.

Une fois les liens entre les ontologies et le texte établis, nous cherchons à aligner ces ontologies en nous appuyant sur les informations textuelles, ce que nous exposons dans la section suivante.

2.3.2 Phase d'alignement guidé par le texte

L'alignement de deux ontologies consiste à retrouver des correspondances entre leurs entités. Dans le chapitre 1, nous avons décrit des méthodes d'alignement d'ontologies qui proposent des techniques terminologiques et structurelles. D'autres s'appuient sur des ressources externes comme WordNet, une ontologie ou le texte pour rapprocher les concepts des ontologies. Notre travail s'inscrit dans la dernière catégorie de travaux où le but est d'identifier des liens entre différentes ontologies d'une manière automatique en se fondant sur la richesse de la langue naturelle exprimée par le texte. Notre travail se focalise sur les ontologies lexicalisées.

Les méthodes d'alignement sont rarement utilisées en dehors de toute application. Les travaux d'alignement des ontologies font une hypothèse assez forte en supposant que les ontologies décrivent toutes les informations utiles à leur exploitation. En effet, l'alignement est rarement une fin en soi, il est devrait de ce fait être guidé par l'application visée. Nous proposons une approche complémentaire qui permet d'exploiter les informations qui sont relatives au domaine de spécialité autres que celles véhiculées par les ontologies et qui permet d'orienter l'alignement selon l'application. Dans la phase d'alignement, notre but n'est pas seulement de mettre en correspondance des ontologies mais aussi d'utiliser un support permettant de caractériser les notions du domaine auquel l'ingénieur de la connaissance s'intéresse. Il s'agit dans notre travail du texte comme un support de travail.

Cette phase prend en entrée les deux ressources sémantiques et le texte découpé en phrases avec des unités textuelles annotées et fournit en sortie toutes les correspondances possibles entre les entités sémantiques. Cet alignement est de type n:m. Les types de correspondances entre entités sont généralement de type équivalence et hiérarchique. Dans cette

phase d'alignement, nous avons caractérisé les relations entre entités en nous fondant sur le texte et les relations existant entre les unités textuelles. Nous nous intéressons à deux types de liens : l'association et l'équivalence sémantiques. Ces deux relations sont détaillées dans le chapitre 3. Dans le but d'extraire ces deux types de relations, nous nous appuyons sur la notion de cooccurrence des entités sémantiques dans le texte. Cette phase s'appuie sur deux étapes : le calcul de correspondances entre les ontologies et le filtrage. La première étape consiste à identifier les relations entre les entités sémantiques des ontologies. La deuxième étape permet de réduire le nombre de correspondances obtenues selon un seuil fixé. Ces deux étapes sont détaillées dans le chapitre suivant.

La figure 2.5 montre un exemple d'alignement entre deux ontologies lexicalisées fondé sur le texte.

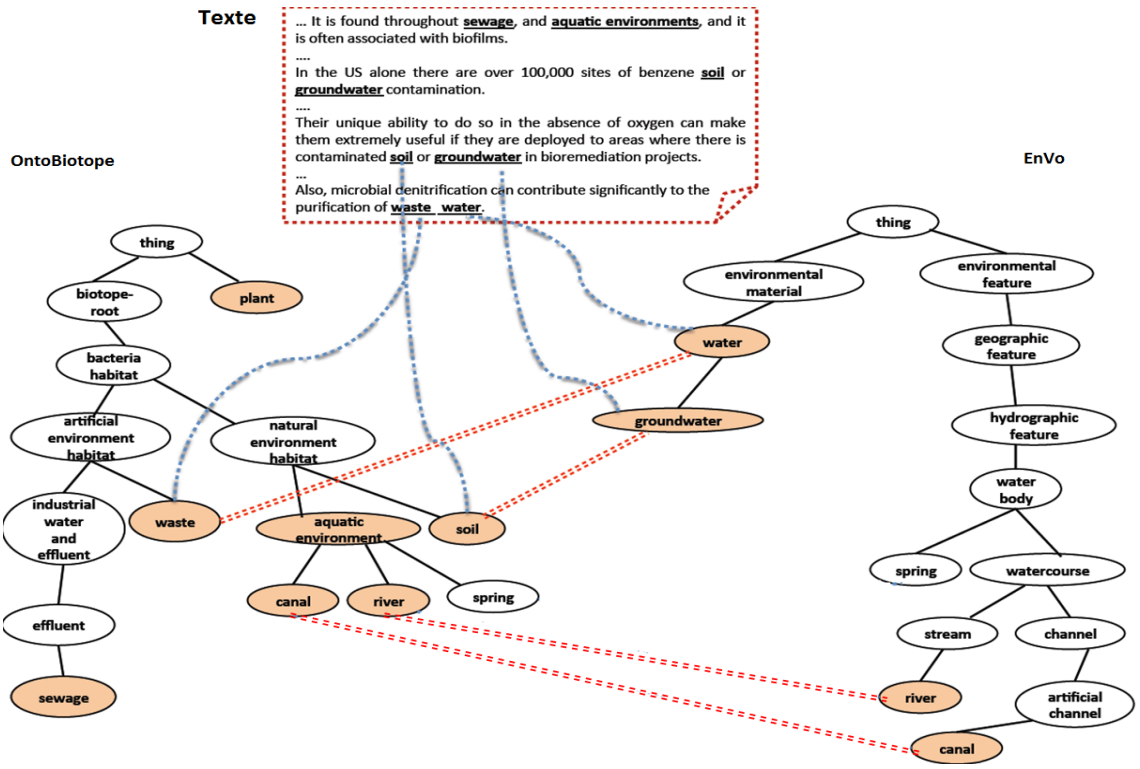


FIGURE 2.5 – Exemple d'alignement entre deux ontologies

La sortie de la phase d'alignement est un ensemble de correspondances. Une correspondance est définie par un 5-uplet : $< idR, e_1, e'_1, score, type\_relation >$ où idR : identifiant de la correspondance, $e_1 \in O_1$ (première ontologie) et $e'_1 \in O_2$ (deuxième ontologie), $score$: le taux de similarité entre deux concepts par l'intermédiaire du texte, $type\_relation$: le type de relation.

La figure 2.6 montre des exemples de correspondances entre les ontologies présentées dans la figure 2.5.

Une fois que nous avons obtenu les liens entre les ontologies en nous laissant guider par le texte, nous nous focalisons sur la révision de la sortie d'alignement.

$< idE_0, river, river, 1.0, equiv >$
$< idE_1, canal, canal, 1.0, equiv >$
$< idA_0, waste, water, 0.56, assoc >$
$< idA_1, soil, groundwater, 0.72, assoc >$

FIGURE 2.6 – Exemple de la sortie de l'alignement de deux ontologies

2.3.3 Phase de construction de la cartographie

L'objectif de cette phase est de revoir l'ensemble de correspondances produit en s'appuyant sur la structure des ontologies. Cette phase prend en entrée l'ensemble de correspondances entre entités sémantiques ainsi que les ontologies et fournit en sortie une cartographie de domaine représentant l'ensemble révisé de relations entre les ontologies. Cette phase comporte deux étapes : (1) la détection et la résolution des anomalies, et (2) la détection et l'affichage des correspondances remarquables. La première étape consiste à identifier les problèmes dans les liens entre les entités mises en correspondance et à les corriger soit d'une manière semi-automatique (intervention de l'ingénieur de la connaissance) soit automatique en précisant l'application visée. Les problèmes que nous avons repérés en raisonnant sur la structure d'ontologies sont liés à l'incompatibilité des liens d'équivalence et à l'ambiguïté avec une entité ou avec des relations d'équivalence et d'association. Tous ces problèmes sont détaillés dans le chapitre 4. La deuxième étape permet d'identifier les liens remarquables entre les ontologies et de les présenter à l'ingénieur de la connaissance. L'objectif est de présenter à l'ingénieur de la connaissance des relations ou des entités qui semblent particulièrement pertinentes pour son domaine d'application.

La sortie de cette étape est une cartographie de domaine représentée sous la forme d'un ensemble de correspondances ainsi que des configurations qui facilitent l'analyse à l'ingénieur de la connaissance (voir chapitre 4 pour plus de détails).

2.4 Exemple

Nous prenons deux extraits des ontologies OntoBiotope et EnVo fournies par l'INRA [2] et un extrait des textes de test de BioNLP-ST 2011 [3] spécifique à la localisation des bactéries nommé BB pour « Bacteria Biotope » (voir chapitre 5 pour une description détaillée). L'extrait de OntoBiotope est constitué de 27 concepts et l'extrait de EnVo comprend 15 concepts. Nous montrons, dans la figure 2.7, les deux ontologies et leurs entités repérées dans le texte. On remarque qu'il existe des entités partagées par les deux ressources (couleur rouge) à savoir « soil », « sediment » et « groundwater ». D'autres entités sont propres à chaque ressource : (i) dans OntoBiotope (couleur bleue) : « plant », « host plant », « human », « cell » et « root » , et (ii) dans EnVo (couleur verte) : « surface » et « habitat ».

2. http://www.inra.fr/
3. http://weaver.nlplab.org/~bionlp-st/BioNLP-ST/

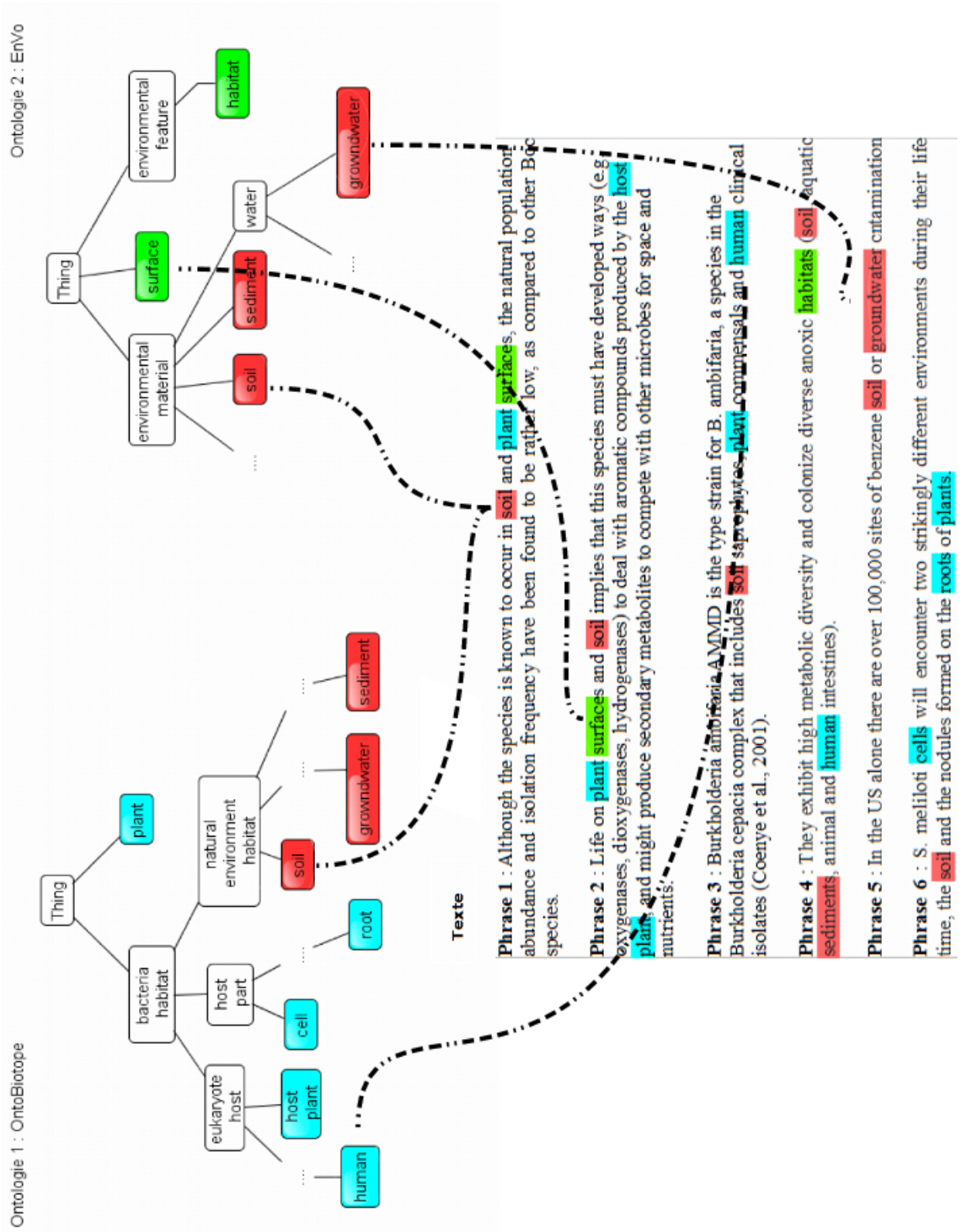


FIGURE 2.7 – Annotation d'un texte par deux ontologies (OntoBiotope et EnVo)

Ce même exemple va être utilisé pour illustrer les autres phases (l'alignement et la construction d'une cartographie de domaine) de la méthodologie proposée. Nous les détaillons dans les chapitres suivants.

2.5 Conclusion

Dans ce chapitre, nous avons présenté la méthodologie que nous proposons pour aider l'ingénieur à prendre connaissance des ressources disponibles sur le domaine auquel il s'intéresse, pour les analyser et les confronter. Cette méthodologie permet de construire une cartographie de domaine en s'appuyant sur un texte représentatif du domaine visé et sur un ensemble de ressources potentiellement pertinentes pour le domaine en question. La construction de cette cartographie se fait en trois étapes : 1) l'annotation du texte au regard des ontologies à aligner, 2) l'alignement de ces ressources en s'appuyant sur le texte pivot ou, plus précisément, sur les propriétés distributionnelles des termes associés aux entités ontologiques et 3) la construction de la cartographie de domaine qui consiste à identifier, dans les sorties d'alignement, des configurations de liens à analyser en priorité pour l'ingénieur de la connaissances.

Nous définissons la cartographie comme une structure de connaissance qui donne une représentation cohérente des correspondances entres les ressources qui la composent et met en évidence les configurations correspondant à des zones ou configurations d'intérêt.

Nous détaillons dans le chapitre 3 la phase d'alignement des ontologies lexicalisées à partir des informations textuelles, puis nous présentons, dans le chapitre 4, la méthode de construction d'une cartographie de domaine en aval du processus d'alignement.

Chapitre 3

Alignement guidé par le texte : *TOM*

Sommaire

3.1 Introduction

De nombreuses méthodes d'alignement des ontologies ont été proposées au cours de la dernière décennie, dans l'objectif de fusionner des ontologies [de Bruijn *et al.*, 2006] ou de développer des connaissances [Huza *et al.*, 2006]. La diversité des types de ressources et leur hétérogénéité sémantique imposent en effet d'établir des ponts entre les différentes ressources que l'on cherche à exploiter.

Le processus d'alignement repose généralement sur deux phases : 1) la transformation des ontologies en un format facile à exploiter (ex. OWL) et 2) la recherche de correspondances entre les entités des ontologies à aligner. Notre approche est complémentaire de celles de l'état de l'art en ce qu'elle s'appuie sur des sources d'informations externes liées au domaine de spécialité et à l'application visée pour guider le processus d'alignement, mais elle s'en distingue par le fait que cette ressource externe est textuelle, une approche qui a encore été peu explorée.

Le texte n'est pas considéré comme une base de connaissances mais plutôt comme un support de travail : on peut exploiter les propriétés distributionnelles des étiquettes des entités ontologiques pour proposer des correspondances entre ces dernières et pour corroborer ou invalider les correspondances détectées par d'autres méthodes d'alignement. Exploiter une source textuelle impose en contrepartie de travailler sur des ontologies lexicalisées où les étiquettes des entités sont des mots de la langue considérée, permettant de lier les textes et les ontologies.

Nous proposons donc une méthode d'alignement guidé par le texte qui prend en entrée deux ontologies lexicalisées et un texte découpé en phrases contenant des unités textuelles annotées et qui fournit en sortie un ensemble de correspondances entre des paires d'entités sémantiques appartenant aux deux ontologies sources (voir figure 3.1). Nous nous appuyons

sur la distribution des entités sémantiques repérées dans le texte pour extraire deux types de relations, des relations d'association et d'équivalence sémantique.

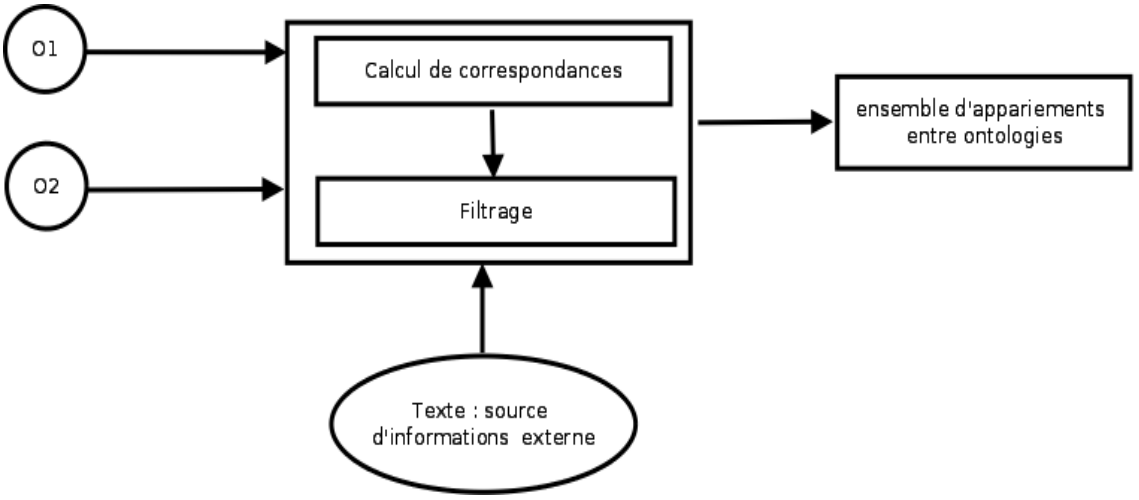


FIGURE 3.1 – Processus d'alignement guidé par le texte

Le reste du chapitre est organisé comme suit : nous caractérisons les relations que nous cherchons à établir entre les entités sémantiques des ontologies. La section 3.3 présente les deux étapes de notre méthode d'alignement *TOM* (Text-based Ontology Mapping) : le calcul des correspondances et leur filtrage. La section 3.4 explique comment cette méthode a été implémentée.

3.2 Types de correspondances recherchés

Il existe une grande richesse des relations existantes entre les mots dans un texte (ex. synonymie). Ces relations vont jouer un rôle important pour extraire des correspondances entre les entités de ressources. Une correspondance entre entités sémantiques est une relation binaire entre deux entités de deux ressources différentes. A partir du texte, nous repérons les entités sémantiques pertinentes du domaine et nous repérons la distribution des entités associées aux unités textuelles. Ces deux éléments (la présence des entités et leur distribution dans le texte) permettent de repérer les deux relations suivantes :

- la relation d'association sémantique se définit par la liaison qui existe entre deux entités qui ont tendance à être souvent contiguës. Ces entités tendent à se combiner l'une avec l'autre ou à apparaître ensemble. Cette relation indique une proximité sémantique entre les entités. Prenons l'exemple de « étudiant » et « université » qui sont deux entités sont souvent liées dans le domaine académique, « sol » et « bactérie » sont aussi deux entités qui sont souvent liées dans le domaine biologique. La nature de la relation d'association est différente dans les deux exemples mais ce sont des termes qui apparaissent souvent combinés l'un à l'autre ; dans le premier exemple, l'étudiant « est inscrit dans » une université. Dans le deuxième exemple, dans le sol, les bactéries se fixent et se multiplient. Cette relation d'association

sémantique peut correspondre aux rôles dans les ontologies.
- la relation d'équivalence sémantique est un lien entre deux entités qui renvoient à la même notion. Ces entités sont sémantiquement identiques et recouvrent le même sens (ex. dans une terminologie, cette relation correspond à deux termes synonymes).

Dans ce qui suit, nous décrivons la méthode proposée pour chercher les deux types de relations présentées entre deux ontologies lexicalisées.

3.3 Calcul d'alignement

Pour mettre en relation sémantiquement les entités d'ontologies repérées dans le texte sous forme d'unités textuelles, nous nous appuyons sur leur répartition dans le texte. Nous exploitons les relations que les unités textuelles entretiennent pour proposer des relations entre les entités sémantiques associées. Elles peuvent apparaître de deux manières ; certaines tendent à apparaître ensemble on s'intéresse alors à leur cooccurrence ; d'autres n'apparaissent pas ensemble mais sont substituables l'une à l'autre, on s'intéresse à leur cooccurrence avec les mêmes unités textuelles.

Dans cette section, nous présentons les deux étapes qui nous permettent d'extraire ces deux relations : (1) le calcul de correspondances, et (2) le filtrage guidé par la cooccurrence.

3.3.1 Calcul de correspondances

Le but du calcul de correspondances est de repérer les entités qui sont suffisamment liées. Nous tenons compte de la force d'association lors de la correspondance. [Grefenstette, 1994] donne trois niveaux d'affinités de mots : (1) le premier niveau : les mots qui tendent à apparaître ensemble, (2) le deuxième niveau : les mots qui partagent les mêmes contextes (similarité), et (3) le troisième niveau, permet la distinction de sens des mots.

Dans ce travail, nous optons pour l'utilisation des deux premiers niveaux (cooccurrence et similarité). Notre approche est simple. Nous procédons comme suit : (1) définition du contexte, (2) calcul d'associations, et (3) calcul de similarités.

Définition du contexte Le contexte d'apparition d'une entité repérée dans le texte est défini par rapport aux segments de texte. Dans les analyses distributionnelles, le contexte de ces entités peut être une fenêtre de mots, un paragraphe ou une phrase. Nous choisissons, dans un premier temps, la phrase comme contexte.

Calcul d'associations La cooccurrence de deux entités repérées dans le texte est le fait que deux entités apparaissent simultanément dans un même contexte. Le traitement des cooccurrences permet de considérer les entités sémantiques dans leur contexte et d'extraire les relations qui peuvent exister.

La cooccurrence est exprimée par un score de fréquence de cooccurrences, ceci n'est pas suffisamment expressif. Pour avoir une force d'association, nous avons donc besoin de

plus d'informations sur la répartition des cooccurrences d'entités dans le texte. Autrement dit, nous étudions la répartition des paires d'entités à rapprocher dans tous leurs contextes ; le fait d'apparaître ensemble et avec toutes les autres entités sémantiques.

Nous voulons une mesure de cooccurrences qui tienne compte non seulement du nombre de cooccurrence entre les entités à rapprocher mais aussi du nombre de cooccurrence avec les autres entités et leur fréquence dans le texte.

Nous prenons en compte les deux critères suivants :

- *Lien entre deux entités* le fait que la présence d'une entité dans un contexte entraîne la présence de l'autre entité de la paire dans le même contexte. Ce lien est représentée par le nombre de fois où les deux entités sémantiques, apparaissent ensemble. On parle de la fréquence absolue de cooccurrence.
- *Lien de chaque entité avec d'autres entités* le fait que l'apparition d'une entité de la paire à rapprocher entraîne l'apparition des autres entités sémantiques dans les mêmes contextes. Ce lien est représenté par : (i) le nombre de fois où la première entité est présente avec d'autres entités et toute seule, et (ii) le nombre de fois où la deuxième entité est présente avec d'autres entités et toute seule.

Plusieurs méthodes ont été proposées pour attribuer une force d'association à une paire d'entités sémantiques. Parmi ces mesures, nous proposons d'adopter celle de [Jaccard, 1901] :

$$S_{Jaccard} = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$$

où :

- E_1 est l'ensemble d'entités de la première ontologie à rapprocher et E_2 est le nombre d'entités de la deuxième ontologie à rapprocher.
- $E_1 \cap E_2$ est le nombre de fois de cooccurrence entre la paire d'entités à rapprocher ;
- $E_1 \cup E_2$ donne le nombre d'occurrences des entités des ontologies ainsi que les cooccurrences avec les autres entités.

Nous construisons la matrice d'associations en nous fondant sur la distribution des couples d'entités dans le texte. Cette matrice est symétrique. Elle contient en lignes et en colonnes les entités sémantiques des deux ontologies dont des mentions figurent dans le texte (voir la figure 3.2). Le score d'associations correspond au calcul de la mesure Jaccard.

Une fois la matrice construite, nous utilisons une partie de cette matrice pour extraire les relations d'association sémantique (voir figure 3.3) entre les concepts. En pratique, l'ensemble de la matrice fournit des relations d'association des deux ressources.

Calcul de similarités La matrice de cooccurrences de la figure 3.2 nous sert aussi à calculer la similarité entre entités. Ce calcul repose sur l'étude des deux vecteurs de scores de cooccurrences des paires d'entités rapprochées. Autrement dit, nous exploitons la cooccurrence de chaque entité avec toutes les entités des deux ontologies (voir figure 3.4).

Différentes mesures de similarité ont été proposées en recherche d'information pour quantifier les similarités entre documents. Une mesure possible est le cosinus qui mesure

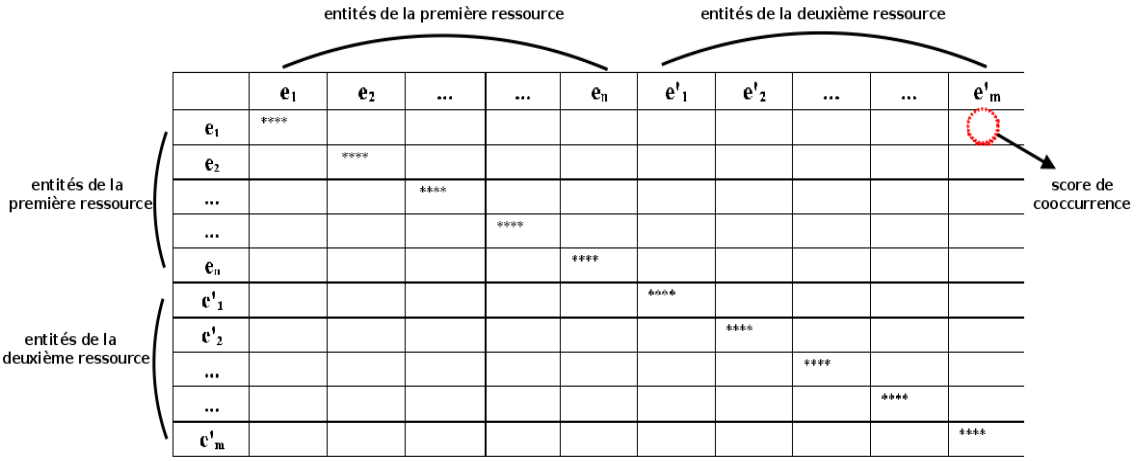


FIGURE 3.2 – Matrice de cooccurrences entre entités sémantiques

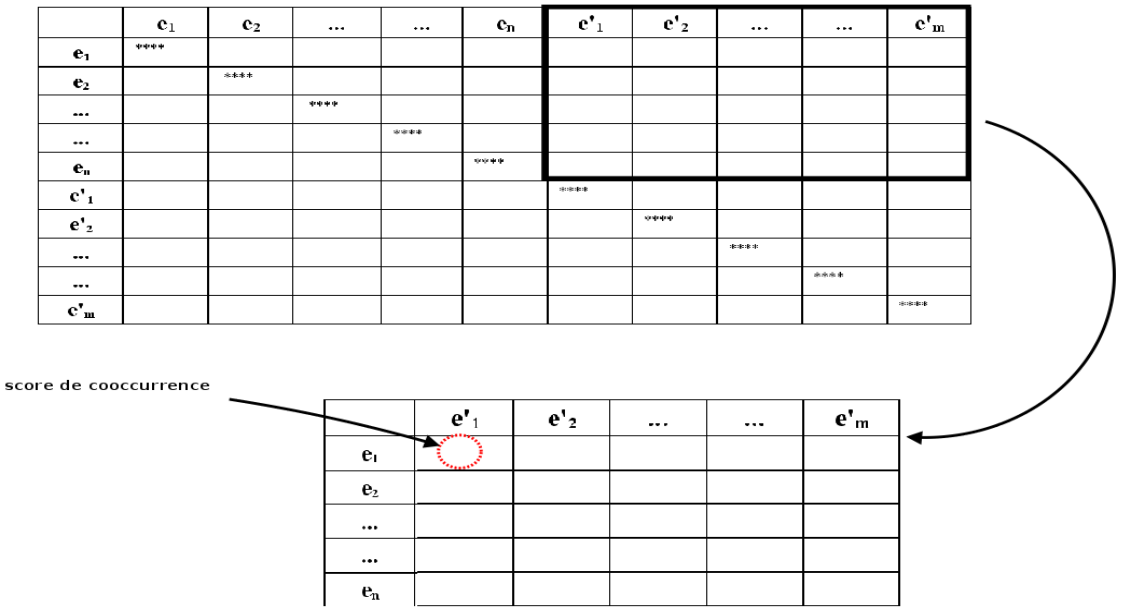


FIGURE 3.3 – Partie de la matrice de cooccurrences pour extraire les relations d'association

la ressemblance entre deux entités sémantiques e_1 et e'_1. Le score du cosinus est calculé à partir des vecteurs de cooccurrences des entités e_1 et e'_1. Le score de similarité correspond au score de cosinus entre entités. Le résultat de cette étape est une matrice de scores de similarité entre entités qui est utilisée pour extraire les relations d'équivalence. Soient e_1 : $(x_1, x_2, ..., x_{n+m})$ et e'_1 : $(y_1, y_2, ..., y_{n+m})$ des vecteurs de cooccurrences de e_1 et e'_1. La mesure du cosinus est exprimée comme suit :

$$cosinus(e_1, e'_1) = \frac{\sum_{i=1}^{n+m} x_i y_i}{\sqrt{\sum_{i=1}^{n+m} x_i^2}\sqrt{\sum_{i=1}^{n+m} y_i^2}}$$

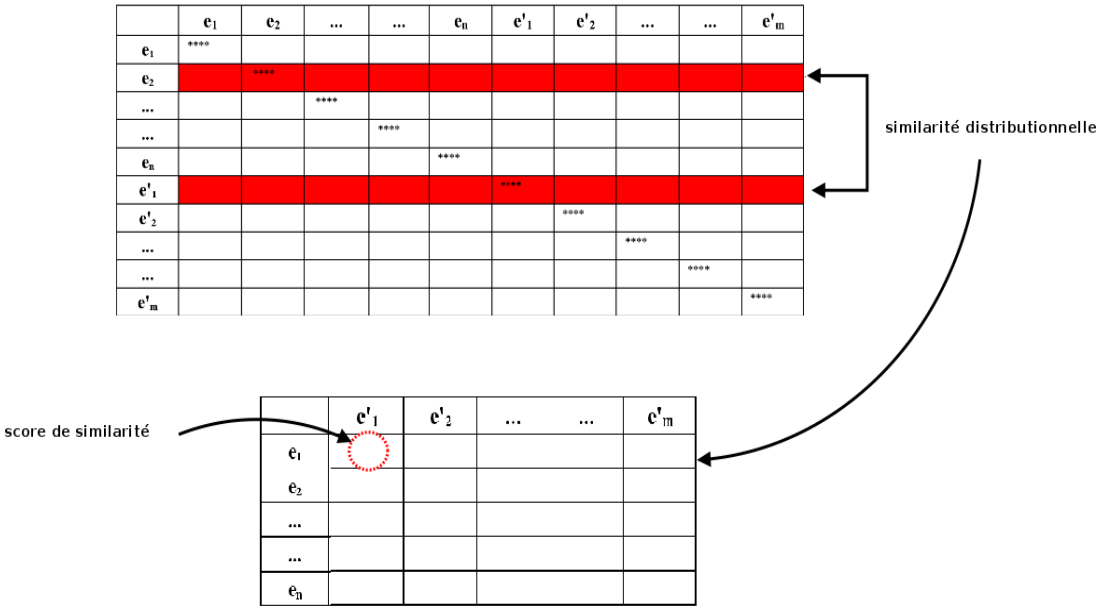


FIGURE 3.4 – Matrice de calcul de similarité pour dériver les relations d'équivalence

La matrice de cooccurrences $Matrice_C$ contenant SC_{ij} représente les scores de co-occurrences SC des entités i et j. La matrice de similarité $Matrice_S$ comportant SS_{ij} représente les scores de similarités SS des entités i et j (voir figure 3.5).

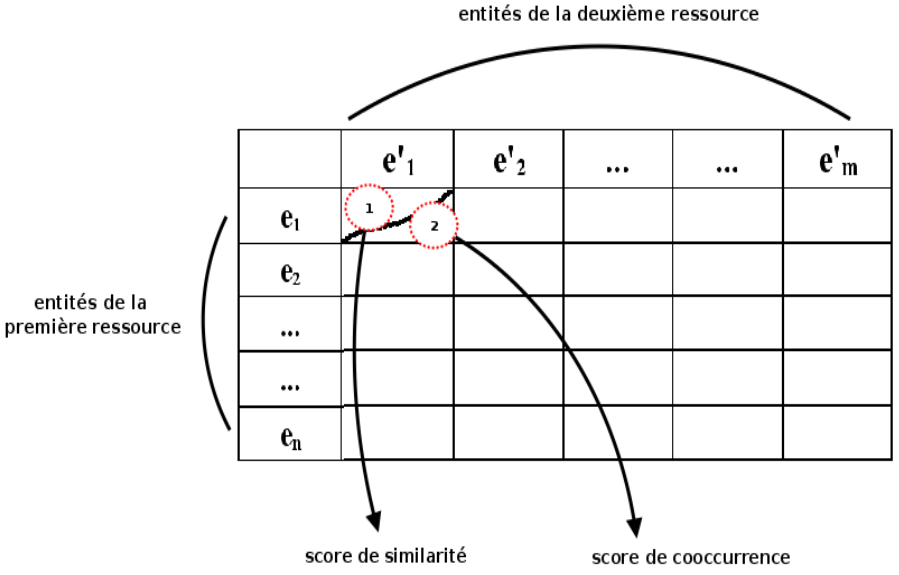
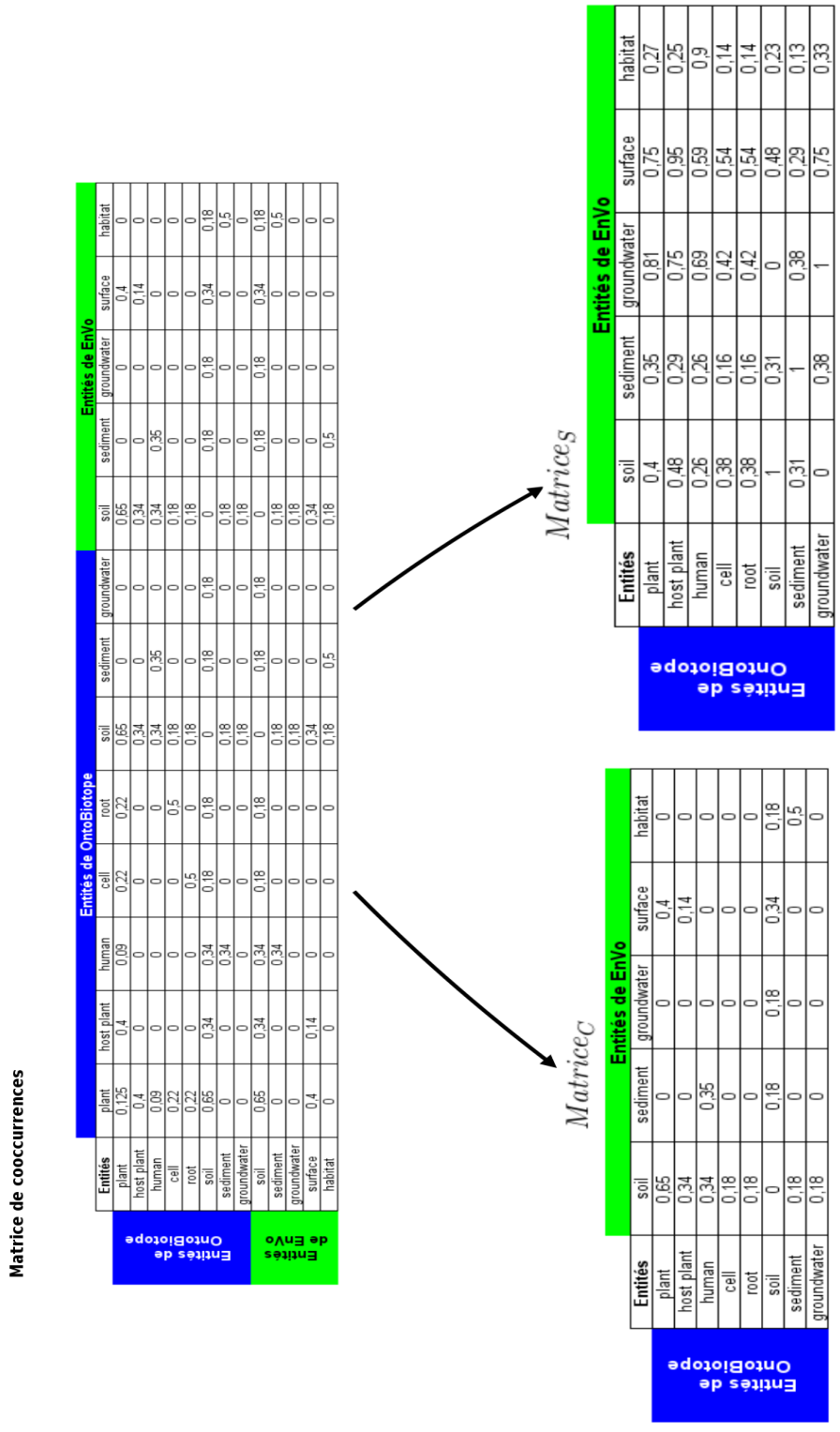


FIGURE 3.5 – Matrice globale contenant les deux matrices : $Matrice_C$ et $Matrice_S$

Etant donné qu'il existe plusieurs mesures de cooccurrences et de similarité, notre méthode d'alignement peut être paramétrée par d'autres mesures.

Exemple Référons-nous à l'exemple présenté dans le chapitre précédent, dans le but d'extraire des relations d'association et d'équivalence sémantiques entre les entités d'ontologies, nous considérons le contexte des entités sémantiques comme étant la phrase. La matrice de cooccurrence dans la figure 3.6 présente le résultat de la force d'association entre entités.

Nous utilisons une partie de cette matrice pour extraire les relations d'association sémantique. La matrice de cooccurrences de la figure 3.6 nous permet aussi de calculer la similarité entre les entités sémantiques. Le calcul de similarité est fondé sur les vecteurs de scores d'association des paires d'entités à rapprocher. La matrice de similarité montre les scores obtenus entre entités suite à l'application du cosinus entre les vecteurs d'entités. Il est à préciser que dans la matrice de cooccurrences ($Matrice_C$), la valeur « 0 » signifie que les deux termes n'apparaissent pas ensemble dans le texte.



Entités	plant	host plant	human	cell	root	soil	sediment	groundwater	soil	sediment	groundwater	surface	habitat
plant	0,125	0,4	0,09	0,22	0,22	0,65	0	0	0,65	0	0	0,4	0
host plant	0,4	0	0	0	0	0,34	0	0	0,34	0	0	0,14	0
human	0,09	0	0	0	0	0,34	0,35	0	0,34	0,35	0	0	0
cell	0,22	0	0	0	0,5	0,18	0	0	0,18	0	0	0	0
root	0,22	0	0	0,5	0	0,18	0	0	0,18	0	0	0	0
soil	0,65	0,34	0,34	0,18	0,18	0	0,18	0,18	0	0,18	0,18	0,34	0,18
sediment	0	0	0,34	0	0	0,18	0	0	0,18	0	0	0	0,5
groundwater	0	0	0	0	0	0,18	0	0	0,18	0	0	0	0
soil	0,65	0,34	0,34	0,18	0,18	0	0,18	0,18	0	0,18	0,18	0,34	0,18
sediment	0	0	0,34	0	0	0,18	0	0	0,18	0	0	0	0,5
groundwater	0	0	0	0	0	0,18	0	0	0,18	0	0	0	0
surface	0,4	0,14	0	0	0	0,34	0	0	0,34	0	0	0	0
habitat	0	0	0	0	0	0,18	0,5	0	0,18	0,5	0	0	0

Entités	soil	sediment	groundwater	surface	habitat
plant	0,65	0	0	0,4	0
host plant	0,34	0	0	0,14	0
human	0,34	0,35	0	0	0
cell	0,18	0	0	0	0
root	0,18	0	0	0	0
soil	0	0,18	0,18	0,34	0,18
sediment	0,18	0	0	0	0,5
groundwater	0,18	0	0	0	0

Entités	soil	sediment	groundwater	surface	habitat
plant	0,4	0,35	0,81	0,75	0,27
host plant	0,48	0,29	0,75	0,95	0,25
human	0,26	0,26	0,69	0,59	0,9
cell	0,38	0,16	0,42	0,54	0,14
root	0,38	0,16	0,42	0,54	0,14
soil	1	0,31	0	0,48	0,23
sediment	0,31	1	0,38	0,29	0,13
groundwater	0	0,38	1	0,75	0,33

FIGURE 3.6 – Matrice d'association $Matrice_C$ et de similarités $Matrices$

3.3.2 Filtrage

A partir des deux matrices précédentes, nous disposons de (n x m) x 2 relations entre entités avec un score associé.

Le filtrage est une étape qui permet d'éliminer les correspondances périphériques possédant des scores très bas. Cette étape a pour but de faciliter l'exploitation des correspondances entre entités sans pour autant se noyer avec un flot de correspondances périphériques qui sont ingérables.

Pour ce faire, plusieurs méthodes de filtrage sont appliquées. La plus intuitive est de fixer un seuil en tenant compte de tous les scores dans chaque matrice (matrice de scores de cooccurrences et de scores de similarités). Nous avons choisi comme seuil la moyenne entre la valeur minimale et la valeur maximale des scores (score de cooccurrence ou score de similarité). A partir de ce seuil, nous estimons que les correspondances pertinentes sont celles qui ont un score associé supérieur au seuil fixé. Les scores des correspondances retenues

indiquent la fiabilité de la correspondance entre entités et cela permet de filtrer le résultat de l'alignement.

La sortie de cette étape de calcul des correspondances est un ensemble de 5-uplets contenant l'identifiant de la relation, la paire d'entités mise en correspondance, la relation extraite et un score indiquant la fiabilité de cette relation.

Deux entités peuvent être liées par deux relations différentes. Une entité peut être liée à plus d'une entité.

Exemple Nous reprenons l'exemple du chapitre précédent et nous l'utilisons dans la phase de filtrage. Nous fixons le seuil pour les matrices de cooccurrences et de similarités $Matrice_C$ et $Matrice_S$ respectivement, comme la moyenne du minimum et du maximum des scores de cooccurrences et des scores de similarités qui sont différents de 0. Le seuil de $Matrice_C$ est $seuil_C = 0.39$ $((0.65 + 0.14)/2)$ et le seuil de $Matrice_S$ est $seuil_S = 0.56$ $((1 + 0.13)/2)$. Les relations d'association et d'équivalence sémantiques retenues sont présentées dans le tableau 3.1. Nous générons 7 relations d'associations sémantiques et 15 relations d'équivalence.

Dans la section suivante, nous décrivons l'implémentation de notre méthode d'alignement d'ontologies guidé par le texte TOM.

3.4 Implémentation

Notre méthode d'alignement *TOM* a été implémentée en Java (voir annexe A). L'application *TOM* comporte quatre modules. Le premier module permet de charger les ontologies à aligner qui sont décrites dans le langage OWL ainsi que le texte sous format textuel. Le second module consiste à lemmatiser et segmenter le texte. Le troisième module permet de lier le texte aux ontologies (ancrage) pour pouvoir aligner les deux ontologies. Le quatrième module permet d'aligner les ontologies.

Rel. d'association	Rel. d'équivalence
$< idA_0, plant, soil, 0.65, assoc >$	$< idE_0, soil, soil, 1, equiv >$
$< idA_1, host\ plant, soil, 0.34, assoc >$	$< idE_1, host\ plant, soil, 0.48, equiv >$
$< idA_2, human, soil, 0.34, assoc >$	$< idE_2, sediment, sediment, 1, equiv >$
$< idA_3, human, sediment, 0.35, assoc >$	$< idE_3, plant, groundwater, 0.81, equiv >$
$< idA_4, plant, surface, 0.4, assoc >$	$< idE_4, plant, surface, 0.75, equiv >$
$< idA_5, soil, surface, 0.34, assoc >$	$< idE_5, soil, surface, 0.48, equiv >$
$< idA_6, sediment, habitat, 0.5, assoc >$	$< idE_6, groundwater, groundwater, 1, equiv >$
*******	$< idE_7, host\ plant, groundwater, 0.75, equiv >$
*******	$< idE_8, host\ plant, surface, 0.95, equiv >$
*******	$< idE_9, human, surface, 0.59, equiv >$
*******	$< idE_{10}, cell, surface, 0.54, equiv >$
*******	$< idE_{11}, root, surface, 0.54, equiv >$
*******	$< idE_{12}, human, groundwater, 0.69, equiv >$
*******	$< idE_{13}, groundwater, surface, 0.75, equiv >$
*******	$< idE_{14}, human, habitat, 0.90, equiv >$

Tableau 3.1 – Tableau de relations retenues d'association et d'équivalence sémantiques

La figure 3.7 présente l'architecture de l'application *TOM*. La sortie de cette application est soit un fichier texte contenant les correspondances qui sera simple à gérer par l'ingénieur de la connaissance, soit un fichier RDF permettant la comparaison avec les outils d'alignement existants et qui permet également une exploitation aisée via des requêtes SPARQL.

Module de chargement Dans ce module, nous avons chargé les deux ontologies OWL à aligner ainsi que le texte choisi comme référence sous format textuel. Nous avons utilisé Jena comme une bibliothèque Java permettant de faciliter la manipulation ainsi que le parcours des ontologies. Jena permet de lire, écrire et interroger une base de fichiers OWL.

Module de lemmatisation et segmentation du texte Ce module permet de découper le texte en phrases et de trouver les lemmes des mots qui le composent et en faisant appel à l'étiqueteur morpho-syntaxique TreeTagger.

Module d'annotation Dans ce module, nous comparons les lemmes du texte aux concepts et à leurs termes associés. Ce module renvoie un texte annoté.

Module d'alignement Ce module permet de rapprocher les entités sémantiques suite à leur présence dans le texte. Nous avons implémenté pour cela la mesure d'association ainsi que celle de similarité. Le résultat de l'alignement est sauvegardé sous format textuel (pour être facile à exploiter par l'ingénieur de la connaissance) ou sous le format de la campagne d'évaluation OAEI. La figure 3.8 présente un exemple du résultat d'alignement.

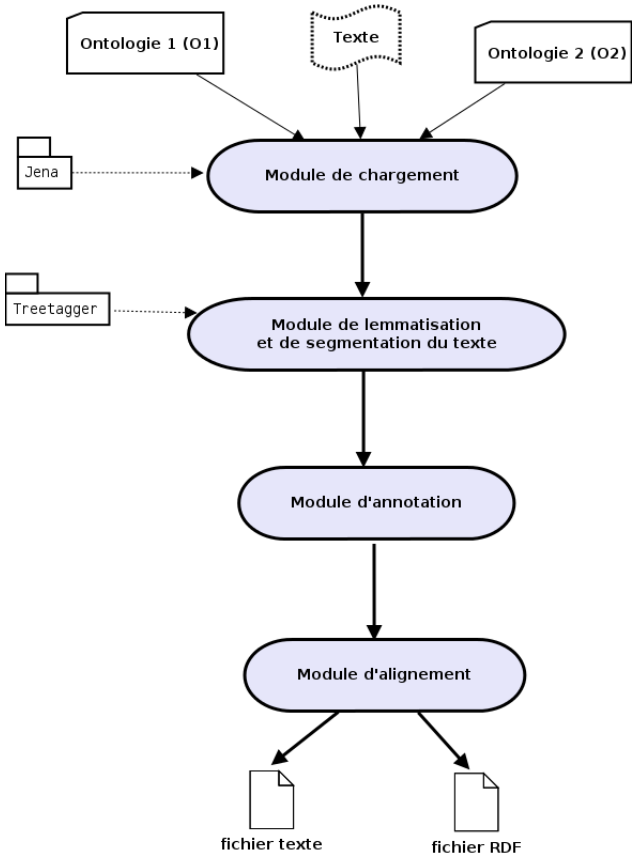


FIGURE 3.7 – Architecture de l'application TOM

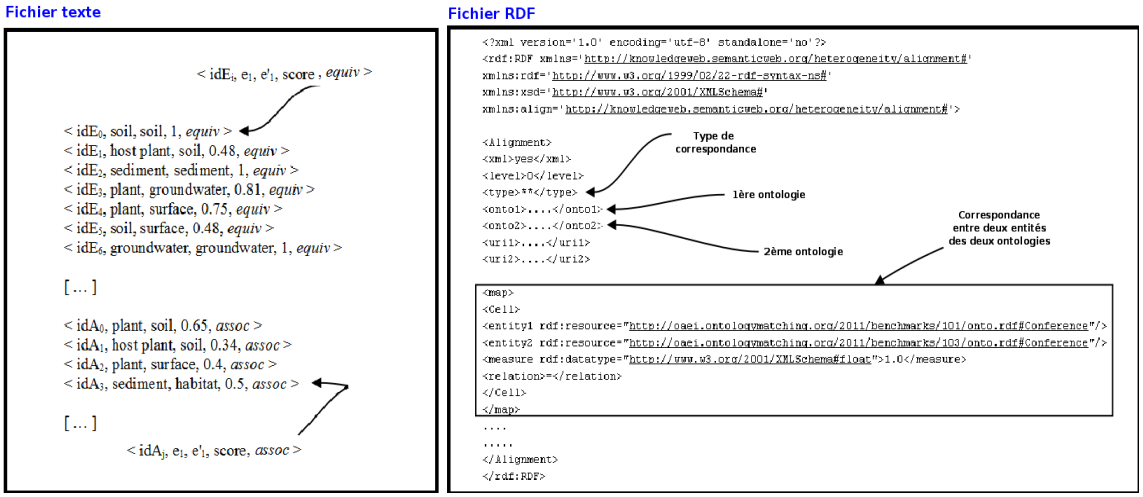


FIGURE 3.8 – Exemple de résultats obtenus par l'alignement des deux ontologies OntoBiotope et EnvO

3.5 Conclusion

Dans ce chapitre, nous avons présenté notre méthode d'alignement guidé par le texte. Cette méthode *TOM* comporte une étape de calcul de correspondances et une étape de filtrage. En sortie, nous obtenons un alignement de type n:m. Cette méthode s'appuie sur la cooccurrence des étiquettes des entités sémantiques dans le texte pour dériver des relations d'équivalence et d'association.

La figure 3.9 présente une maquette d'interface pour montrer comment concrètement les résultats de l'alignement pourraient être présentés à un utilisateur.

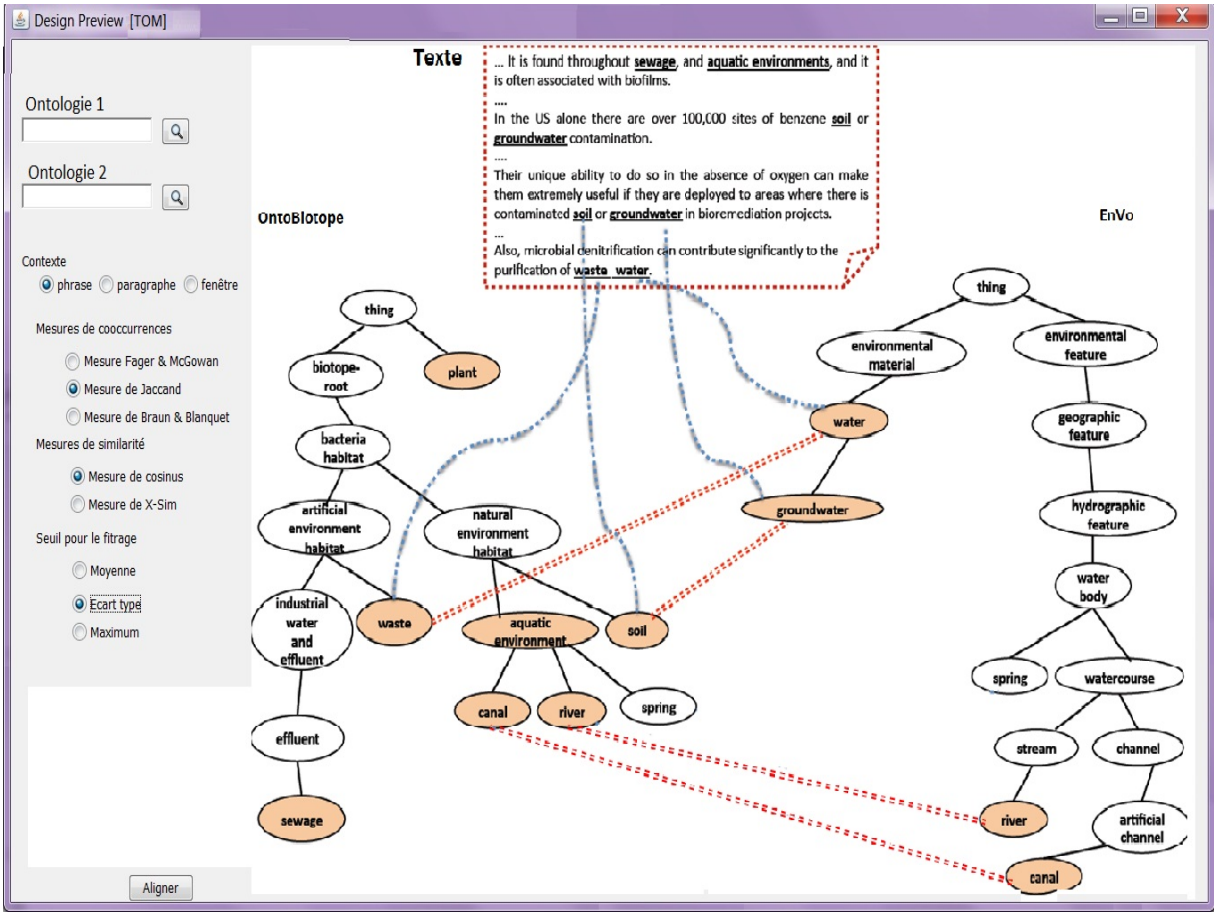


FIGURE 3.9 – Maquette de l'interface de notre méthode d'alignement TOM

Lorsque les ressources sont de taille importante, l'ensemble des correspondances peut être volumineux et difficile à analyser. C'est pourquoi les cartographies ne contiennent pas uniquement le résultat des alignements ; elles comportent également un ensemble de zones d'intérêt qui correspondent à des configurations de correspondances anormales et remarquables et qui méritent d'être analysées par l'ingénieur qui souhaite prendre connaissance de l'état des connaissances sur un domaine donné.

Dans le chapitre suivant, nous montrons donc comment exploiter les correspondances

obtenues en sortie d'alignement, en mettant en évidence les zones dans lesquelles les correspondances se trouvent dans des configurations intéressantes *a priori*, qu'elles soient anormales ou au contraire remarquables. C'est la dernière étape de construction des cartographies de domaine.

Chapitre 4

Construction de la cartographie de domaine

Sommaire

4.1 Introduction

Une fois les ressources alignées et un ensemble de correspondances n:m établi entre les entités qui les composent, il reste à exploiter le résultat obtenu. On peut bien entendu chercher à fusionner ces ressources pour en construire une nouvelle mais c'est un cas extrême ; le plus souvent on s'appuie sur l'alignement pour isoler une sous-partie intéressante ou pour analyser les différents choix de modélisation proposés. Dans tous les cas, il faut commencer par prendre connaissance des ressources disponibles. C'est à cet effet que nous cherchons à construire des cartographies de domaines qui présentent les ressources disponibles en les articulant les unes aux autres (liens de correspondance) ainsi que par rapport au texte de référence (liens d'annotation) et en faisant ressortir les zones d'intérêt dans l'ensemble de liens de correspondance. La cartographie peut être exploitée dans le cadre de nombreuses applications telles que la construction d'une ressource sémantique, le découpage d'ontologies en blocs cohérents (la modularisation) et la fusion de ressources hétérogènes. Ces applications peuvent influencer la définition de ce qui est considéré comme zone d'intérêt et pourrait être paramétré par l'ingénieur

L'analyse de ces correspondances met en effet en évidence des configurations intéressantes. Certaines sont « anormales » et font apparaître que les ressources alignées reposent sur des choix de modélisation très différents, voire incompatibles. D'autres au contraire sont « remarquables » et font apparaître des points de jonction entre les ressources. Notre objectif est de doter les cartographies de domaine d'un outil d'interrogation qui permette à l'utilisateur d'explorer ces configurations pour l'aider à prendre connaissance des ressources à sa disposition lorsque ces dernières sont de grande taille et donc difficiles à analyser. Ces

configurations sont également utiles pour valider les correspondances fournies par notre algorithme d'alignement, si besoin est. Ce type d'outil d'exploration des alignements est d'autant plus utile qu'on peut envisager à terme d'aligner non pas deux ontologies mais plusieurs ressources sémantiques les unes par rapport aux autres.

D'autres chercheurs se sont intéressés à la révision des sorties d'alignement. [Hanif *et al.*, 2006] proposent par exemple une méthode permettant d'éliminer des correspondances erronées en utilisant deux techniques d'alignement (terminologique et structurelle) et en privilégiant les correspondances obtenues par les deux techniques. C'est naturellement une approche très sélective. [Stuckenschmidt *et al.*, 2005] proposent d'évaluer une sortie d'alignement à partir de ses propriétés de cohérence et de minimalité. [Wang et Xu, 2008] montrent comment repérer les erreurs d'alignement comme par exemple les correspondances redondantes qui sont inutiles à présenter à la fin du processus d'alignement. Notre construction de cartographies de domaine s'inscrit dans le prolongement de ces travaux qui analysent et retraitent les sorties d'alignement.

Nous nous intéressons dans ce chapitre non plus à la cooccurrence des entités dans le texte mais plutôt à leur position dans les ontologies. Notre méthode permet de détecter des anomalies en raisonnant sur la structure formée par chacune des ontologies et l'ensemble des correspondances, l'objectif à terme étant d'assurer la cohérence de l'ensemble (éliminer les correspondances erronées) et de mettre en évidence les points de jonction les plus intéressants. Nous définissons pour cela un premier ensemble de configurations intéressantes, anormales ou remarquables. Cet ensemble peut être enrichi en fonction de la finalité de la cartographie. Nous proposons deux méthodes pour les repérer dans les sorties d'alignement : la première est algorithmique et la seconde repose sur le moteur de recherche sémantique Corese [Corby *et al.*, 2006] et des requêtes SPARQL. Nous proposons enfin une méthode d'élimination automatique des configurations anormales même si celles-ci peuvent aussi être analysées manuellement.

Ce chapitre est organisé autour de deux sections centrales qui portent sur les configurations anormales et leur élimination (section 4.2) et sur les configurations remarquables et leur affichage (section 4.3).

4.2 Détection et élimination des anomalies

Nous avons repéré trois configurations, que nous considérons anormales qui sont relatives à des relations de type équivalence et association. Une configuration dite anormale est détectée en raisonnant sur la structure des deux ontologies alignées. Une configuration anormale est une configuration qui regroupe des correspondances qui génèrent des incohérences (ou inconsistances) dans l'une ou l'autre des ontologies.

4.2.1 Configurations anormales

Trois configurations anormales sont détectées. Ces configurations ne couvrent évidemment pas tous les problèmes possibles. Elles permettent, néanmoins, de caractériser les cas des problèmes avec :

– une inversion de hiérarchie ;

– une entité ambiguë ;
– une ambiguïté de relations.

Configuration avec inversion de hiérarchie (C_{hi})

Définition 1 *Une configuration avec inversion de hiérarchie est une anomalie où les entités, qui font partie des deux relations de correspondance de type équivalence, sont structurées d'une manière hiérarchique inversée dans l'une et l'autre des ontologies (voir figure 4.1). On parle d'une incompatibilité des liens d'équivalence. On note une telle configuration par* $C_{hi}(eq_{ij}, eq_{uv})$ *où* $eq_{ij} < idE_x, e_i^1, e_j^2, score, equiv >$, $eq_{uv} < idE_y, e_u^1, e_v^2, score, equiv >$ *sont deux relations de correspondance de type équivalence incompatibles, idE : l'identifiant de la correspondance de type équivalence et* $x \neq y$. *On dit que* eq_{ij} *et* eq_{uv} *sont incompatibles si et seulement si* $[(e_u^1 \sqsubset e_i^1) \wedge (e_j^2 \sqsubset e_v^2)] \vee [(e_i^1 \sqsubset e_u^1) \wedge (e_v^2 \sqsubset e_j^2)]$.

Ce cas peut provenir de (i) une erreur de calcul dans les correspondances, (ii) une erreur d'identification des entités sémantiques, et (iii) d'une erreur de typage de la correspondance. Même si une telle inversion de hiérarchie peut refléter un choix de modélisation différent, nous considérons que c'est une configuration anormale parce qu'elle reflète des choix de modélisation incompatibles.

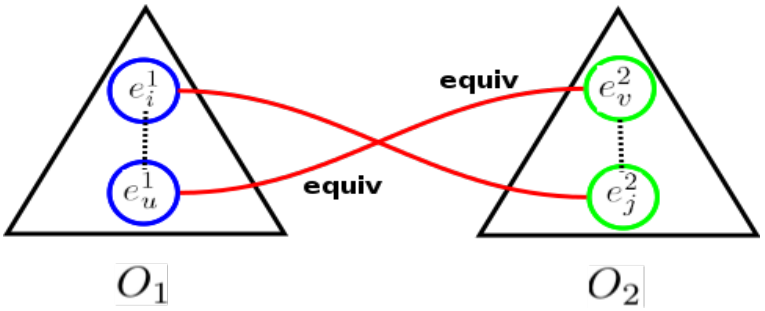


FIGURE 4.1 – Schéma de la hiérarchie inversée $C_{hi}(eq_{ij}, eq_{uv})$

Exemple 1 *La figure 4.2 montre une hiérarchie inversée. Il existe un lien hiérarchique entre les deux entités « pont » et « passerelle » de la première ressource BDTopo et un lien hiérarchique inversé entre les deux entités « chemin » et « sentier » de la deuxième ressource BDCarto.*

Méthodes de détection de l'anomalie Nous proposons deux méthodes pour détecter l'anomalie de la hiérarchie inversée. La première est algorithmique et consiste à tester si les correspondances de type équivalence sont incompatibles. Cette méthode prend en entrée un ensemble de liens d'équivalence $E_{12} = \{< idE, e_x^1, e_y^2, score, equiv > / e_x^1 \in R_1$ et $e_y^2 \in R_2\}$, et donne en sortie une liste de couples de correspondances de type équivalence incompatibles.

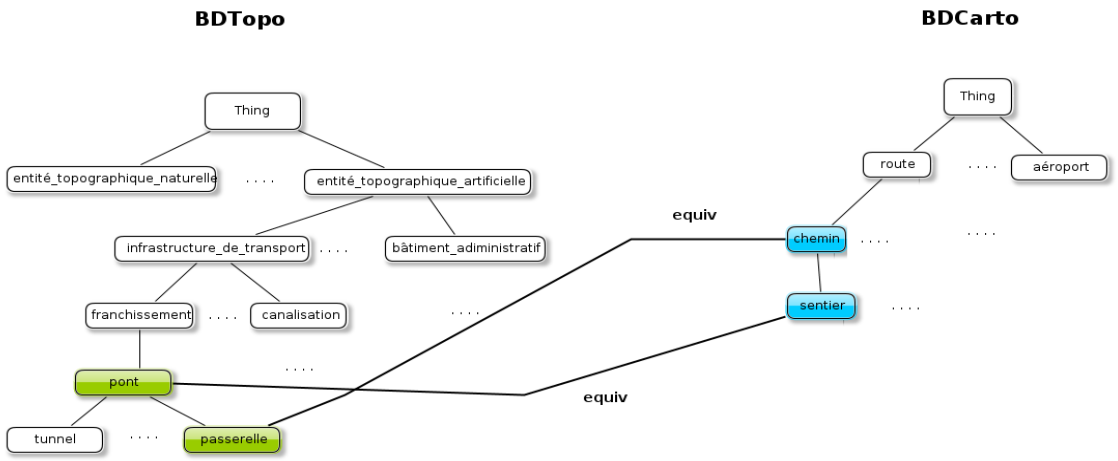


FIGURE 4.2 – Exemple de l'anomalie : hiérarchie inversée

Algorithme 1 *InconsistentLinks*() : liste de couples de correspondances de type équivalence incompatibles

Entrée : $E_{12} = \{< idE, e_x^1, e_y^2, score, equiv >\}$
Sortie : L_{incomp} : liste de couples de correspondances de type équivalence incompatibles
Variable : $incomp$: booléen
Début
pour tout $eq_{ij} \in E_{12}$ *ET* $eq_{uv} \in E_{12}$ **faire**
  pour tout ij allant de 1 à $taille(E_{12})$ **faire**
    pour tout uv allant de $ij + 1$ à $taille(E_{12})$ **faire**
      %% Tester si eq_{ij} et eq_{uv} sont incompatibles
      $incomp = \underline{IncompCheck}(eq_{ij}, eq_{uv})$
      si $incomp = vrai$ **alors**
        inserer$(eq_{ij}, eq_{uv}, L_{incomp})$
      fin si
    fin pour
  fin pour
fin pour
retourne L_{incomp}
Fin

Algorithme 2 $IncompCheck(eq_{ij} : < idE_x, e_i^1, e_j^2, score, equiv >, eq_{uv} < idE_y, e_u^1, e_v^2, score, equiv >)$: booléen

1: **Sortie :** *incomp* : booléen
2: **Début**
3: **si** $[(e_u^1 \sqsubset e_i^1) \wedge (e_j^2 \sqsubset e_v^2)] \vee [(e_i^1 \sqsubset e_u^1) \wedge (e_v^2 \sqsubset e_j^2)]$ **alors**
4: $incomp = vrai$
5: **sinon**
6: $incomp = faux$
7: **fin si**
8: retourne *incomp*
9: **Fin**

La deuxième méthode consiste à appliquer une requête formelle structurée (SPARQL) dans un moteur de recherche sémantique, comme Corese [1]. La requête s'appuie sur une base de correspondances sous le format RDF de la campagne d'évaluation OAEI et les deux ontologies lexicalisées sous format OWL. La requête SPARQL qui permet de détecter le problème de la hiérarchie inversée est la suivante :

```
PREFIX align:<http://knowledgeweb.semanticweb.org/heterogeneity/alignment#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
SELECT ?x1 ?x2 ?y1 ?y2 ?cell1 ?cell2 ?t ?v1 ?v2 WHERE
        {?cell1 rdf:type align:Cell
        ?cell1 align:entity1 ?x1
        ?cell1 align:entity2 ?y1
        ?cell1 align:relation ?t
        ?cell2 rdf:type align:Cell
        ?cell2 align:entity1 ?x2
        ?cell2 align:entity2 ?y2
        ?cell2 align:relation ?t
        ?v1 align:onto1 align:uri1
        ?x1 owl:Class ?v1
        ?y2 owl:Class ?v1
        ?v2 align:onto2 align:uri2
        ?x2 owl:Class ?v2
        ?y1 owl:Class ?v2
        ?y2 rdfs:subClassOf ?x1
        ?y1 rdfs:subClassOf ?x2
        Filter regex(?t, "=")
        }
```

Où : Cell est une balise ($< Cell > ... < /Cell >$) dans le format RDF proposé par la campagne d'évaluation OAEI, qui regroupe les entités des deux ontologies mises en

1. http://www-sop.inria.fr/edelweiss/software/corese/

correspondance par un type de relation, et = : pour la correspondance de type équivalence sémantique.

Configuration avec une entité ambigüe (C_{AmEq})

Définition 2 *Une configuration avec une entité ambigüe est une anomalie où une même entité est associée à deux entités distinctes avec deux relations de correspondance de type équivalence (voir figure 4.3). On parle d'une ambiguïté des deux liens d'équivalence. On note une telle configuration par* $C_{AmEq}(eq_{ij}, eq_{uv})$ *où* $eq_{ij} < idE_x, e_i^1, e_j^2, score, equiv >$ *et* $eq_{uv} < idE_y, e_u^1, e_v^2, score, equiv >$*, sont deux relations de correspondance de type équivalence ambigües, idE : identifiant de la correspondance de type équivalence et* $x \neq y$*. On dit que* eq_{ij} *et* eq_{uv} *sont ambigües si et seulement si il existe une même entité* $(e_i^1 = e_u^1)$ *qui est concernée par ces deux correspondances.*

Ce cas peut provenir d'une erreur d'identification des entités sémantiques aussi si ce type de configuration reflète un choix de modélisations.

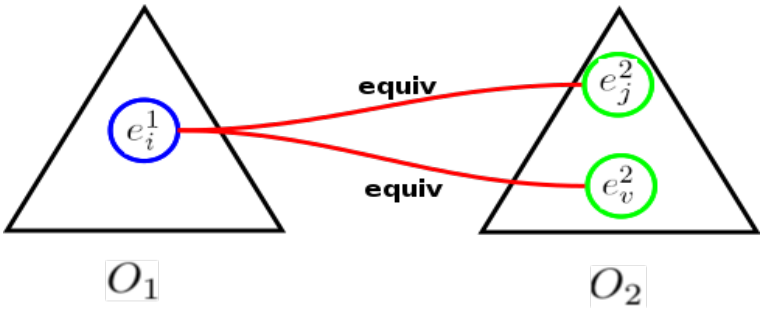


FIGURE 4.3 – Schéma du problème d'ambiguïté avec une entité sémantique

Exemple 2 *La figure 4.4 montre une ambiguïté avec l'entité sémantique « river ». Cette dernière (« river ») de la première ressource OntoBiotope est mise en relation d'équivalence avec deux entités « sludge » et « soil » de la deuxième ressource EnvO.*

Méthodes de détection de l'anomalie On peut détecter une anomalie avec une entité ambigüe de deux façons. La méthode algorithmique teste si les correspondances de type équivalence sont ambigües. Elle prend en entrée un ensemble de liens d'équivalence $E_{12} = \{< idE, e_x^1, e_y^2, score, equiv > / e_x^1 \in R_1 \text{ et } e_y^2 \in R_2\}$, et donne en sortie une liste de couples de correspondances de type équivalence ambigües.

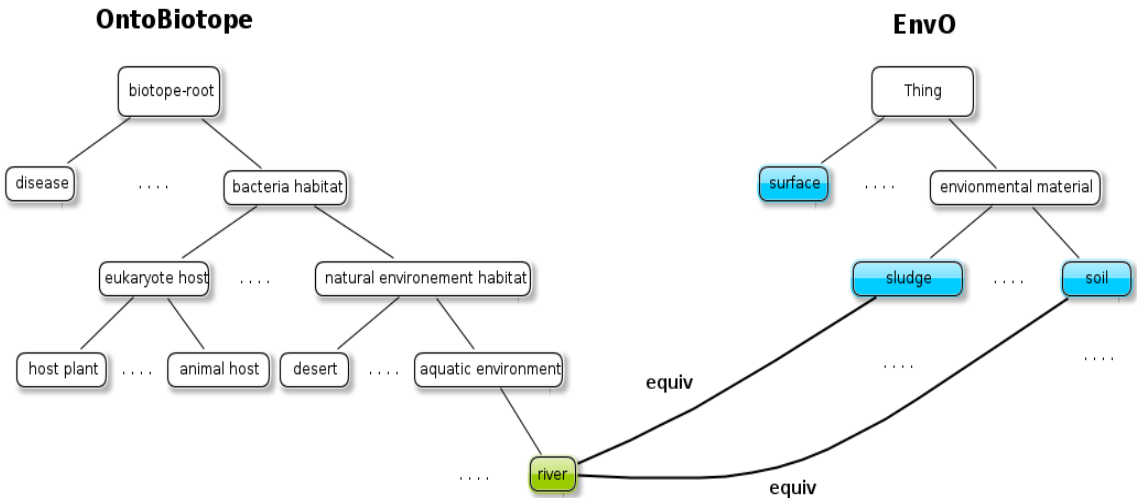


FIGURE 4.4 – Exemple de la configuration avec une entité ambigûe

Algorithme 3 *AmbiguousLinks*() : liste de couples de correspondances de type équivalence ambigüs

Entrée : $E_{12} = \{< idE, e_x^1, e_y^2, score, equiv >\}$
Sortie : L_{amb} : liste de couples de correspondances de type équivalence ambigüs
Variables : amb : booléen
Début
pour tout $eq_{ij} \in E_{12}$ *ET* $eq_{uv} \in E_{12}$ **faire**
  pour tout ij allant de 1 à $taille(E_{12})$ **faire**
    pour tout uv allant de $ij + 1$ à $taille(E_{12})$ **faire**
      %% Tester si eq_{ij} et eq_{uv} sont ambigus
      $amb = \underline{AmbiguityCheck}(eq_{ij}, eq_{uv})$
      si $amb = vrai$ **alors**
        inserer$(eq_{ij}, eq_{uv}, L_{amb})$
      fin si
    fin pour
  fin pour
fin pour
retourne L_{amb}
Fin

Algorithme 4 *AmbiguityCheck*(eq_{ij} : < $idE_x, e_i^1, e_j^2, score, equiv$ >, eq_{uv} < $idE_y, e_u^1, e_v^2, score, equiv$ >) : booléen

1: **Sortie :** *amb* : booléen
2: **Début**
3: **si** ($e_i^1 == e_u^1$) $\wedge$ (e_j^2 et e_v^2) sont distinctes) **alors**
4: $amb = vrai$
5: **sinon**
6: $amb = faux$
7: **fin si**
8: retourne *amb*
9: **Fin**

La requête formelle sur la base de connaissances des correspondances de type équivalence qui permet de détecter ces correspondances est la suivante :

```
PREFIX align:<http://knowledgeweb.semanticweb.org/heterogeneity/alignment#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
SELECT ?x1 ?y1 ?y2 ?cell1 ?t ?v1 ?v2 WHERE
        {?cell1 rdf:type align:Cell
        ?cell1 align:entity1 ?x1
        ?cell1 align:entity2 ?y1
        ?cell1 align:entity2 ?y2
        ?cell1 align:relation ?t
        ?v1 align:onto1 align:uri1
        ?x1 owl:Class ?v1
        ?v2 align:onto2 align:uri2
        ?y1 owl:Class ?v2
        ?y2 owl:Class ?v2
        Filter regex(?t, "=")
        }
```

Configuration avec une ambiguïté de relations ($C_{AmEqAss}$)

Définition 3 *Une configuration avec une ambiguïté de relations est une anomalie où une même entité correspond à une autre entité avec deux relations de correspondance de type équivalence et association (voir figure 4.5). On parle d'une ambiguïté des deux liens d'équivalence et d'association.*

On note une telle configuration par $C_{AmEqAss}(eq_{ij}, ass_{uv})$ *où* eq_{ij} < $idE_x, e_i^1, e_j^2, score, equiv$ > *et* ass_{uv} < $idA_y, e_u^1, e_v^2, score, assoc$ >, *sont deux relations de correspondance de type équivalence et association ambigües, idE : identifiant de la correspondance de type équivalence et idA : identifiant de la correspondance de type association. On dit que* eq_{ij} *et* ass_{uv} *sont ambigües si et seulement si* $e_i^1 = e_u^1$ *et* $e_j^2 = e_v^2$.

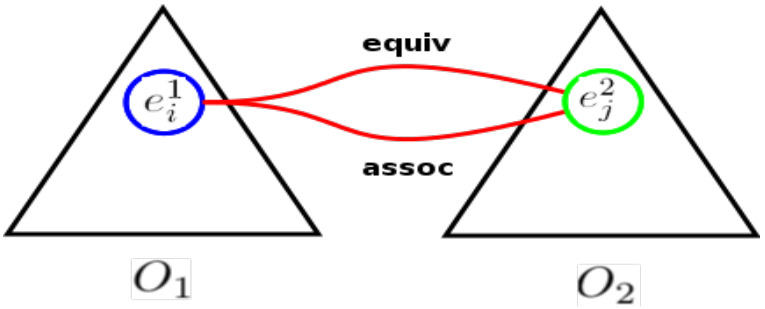


FIGURE 4.5 – Schéma de la configuration avec une ambiguïté de relations

Exemple 3 *La figure 4.6 montre une ambiguïté entre deux relations d'équivalence et d'association entre les deux entités « China » et « city ». L'entité « China » de la première ressource OntoBiotope est mise en correspondance avec l'entité « city » de la deuxième ressource EnvO, avec deux types de liens différents.*

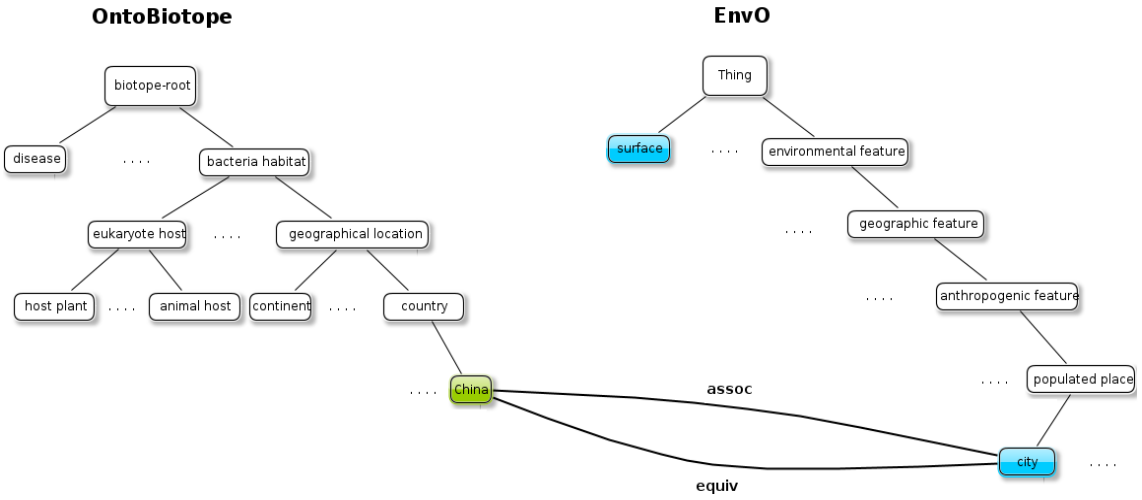


FIGURE 4.6 – Exemple d'une ambiguïté de relations

Méthodes de détection de l'anomalie La méthode algorithmique prend en entrée un ensemble de liens d'équivalence et d'association $E_{12} = \{< idE, e_x^1, e_y^2, score, equiv > \cup < idA, e_z^1, e_t^2, score, assoc > / e_x^1, e_z^1 \in R_1 \text{ et } e_y^2, e_t^2 \in R_2\}$, et donne en sortie une liste de couples de correspondances de type équivalence et association ambigus.

Algorithme 5 *AmbiguousLinksEqAss*() : liste de couples de correspondances de type équivalence et association ambigus

Entrée : $E_{12} = \{< idE, e_x^1, e_y^2, score, equiv > \cup < idA, e_z^1, e_t^2, score, assoc >\}$
Sortie : $L_{ambEqAss}$: liste de couples de correspondances de type équivalence et association ambigus
Variables : $ambEqAss$: booléen
Début
pour tout $(eq_{ij} \in E_{12}) \wedge (ass_{uv} \in EAR_{12})$ **faire**
  pour tout ij allant de 1 à $taille(E_{12})$ **faire**
    pour tout uv allant de $ij + 1$ à $taille(E_{12})$ **faire**
      %% Tester si eq_{ij} et ass_{uv} sont ambigus
      $ambEqAss = \underline{AmbiguityCheckEqAss}(eq_{ij}, ass_{uv})$
      si $ambEqAss = vrai$ **alors**
        inserer$(eq_{ij}, ass_{uv}, L_{ambEqAss})$
      fin si
    fin pour
  fin pour
fin pour
retourne $L_{ambEqAss}$
Fin

Algorithme 6 *AmbiguityCheckEqAss*$(eq_{ij} : < idE_x, e_i^1, e_j^2, score, equiv >, ass_{uv} < idA_y, e_u^1, e_v^2, score, assoc >)$: booléen

Sortie : $ambEqAss$: booléen
Début
si $(e_i^1 == e_u^1) \wedge (e_j^2 == e_v^2)$ **alors**
  $ambEqAss = vrai$
sinon
  $ambEqAss = faux$
fin si
retourne $ambEqAss$
Fin

La requête SPARQL sur la base de connaissances des correspondances de type association et équivalence est la suivante :

```
PREFIX align:<http://knowledgeweb.semanticweb.org/heterogeneity/alignment#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
SELECT ?x1 ?y1 ?cell1 ?t1 ?t2 ?v1 ?v2 WHERE
        {?cell1 rdf:type align:Cell
        ?cell1 align:entity1 ?x1
        ?cell1 align:entity2 ?y1
        ?cell1  align:relation ?t1
```

```
?cell1  align:relation ?t2
?v1 align:onto1 align:uri1
?x1 owl:Class ?v1
?v2 align:onto2 align:uri2
?y1 owl:Class ?v2
Filter regex(?t1, "=") && regex(?t2, "<>")
}
```

Où : <> : pour la correspondance de type association sémantique.

4.2.2 Élimination des anomalies

L'élimination des anomalies consiste à éliminer des correspondances qui sont : (i) incompatibles, et (ii) ambigües. Cette étape peut être interactive si l'ingénieur de la connaissance veut avoir la main et s'assurer qu'on ne supprime pas des correspondances qui ont été jugées anormales alors qu'elles proviennent de choix de modélisation différents. En effet, notre calcul d'incohérence repose sur une vision unifiée des deux ontologies et suppose un raisonnement uniforme. Pour cela, nous pouvons d'abord proposer d'afficher ces configurations à l'instar des configurations remarquables (cf. section 4.3) pour ensuite activer ou non l'élimination automatique.

Notre intuition, dans cette étape, est qu'une correspondance peut poser à la fois des problèmes d'incompatibilité et d'ambiguïté. Le fait de supprimer cette correspondance diminue plusieurs problèmes simultanément. A titre d'exemple, une relation de correspondance de type équivalence peut faire partie de la configuration avec inversion de hiérarchie et de la configuration avec une entité ambigüe. Quand on élimine cette relation, on résout deux anomalies en même temps au lieu d'agir deux fois pour les résoudre.

Nous proposons donc de raisonner globalement sur le nombre d'incompatibilités et d'ambiguïtés par relation. Si le nombre d'incompatibilités et d'ambiguïtés entre deux relations de correspondance ayant le même type de relation (équivalence ou association) est le même, nous raisonnons localement et par ordre de fiabilité des relations. Autrement dit, nous retenons la correspondance ayant le score le plus élevé. Dans le cas où le nombre d'incompatibilités et d'ambiguïtés est le même pour deux types différents (équivalence et association) de relations de correspondance, nous retenons la relation de correspondance de type équivalence. Le choix peut également être donné à l'ingénieur de la connaissance.

L'objectif du raisonnement global est de supprimer les liens, fournis par l'alignement guidé par le texte, qui posent des problèmes. Nous prenons pour cela tous les liens d'équivalence et d'association et nous supprimons le(s) lien(s) qui posent le plus de problèmes. Nous avons trois listes de couples de liens qui correspondent aux différentes configurations :

- $L_{incomp} = \{(eq_{ij}, eq_{i'j'})\}$: liste de couples de correspondances de type « équivalence incompatible » ;
- $L_{amb} = \{(eq_{ij}, eq_{i'j'})\}$: liste de couples de correspondances de type « équivalence ambiguë » ;
- $L_{ambEqAss} = \{(eq_{ij}, ass_{i'j'})\}$: liste de couples de correspondances de type « équivalence et association ambiguës ».

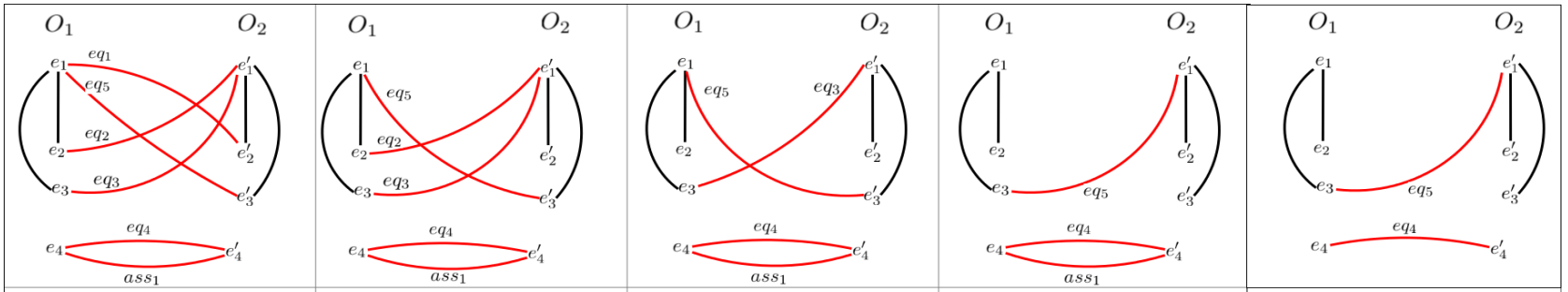
Dans le but de résoudre ces anomalies, nous procédons globalement comme suit :

1) calculer le nombre d'incompatibilités et d'ambiguïtés par type de relation (équivalence et association) ;

Nous répétons le traitement ci-dessous jusqu'à ce que $L_{incomp} = \varnothing$, $L_{amb} = \varnothing$ et $L_{ambEqAss} = \varnothing$

2) détecter le lien qui génère le plus d'incompatibilités et d'ambiguïtés (extraction du maximum) ; si les relations ayant le même type (équivalence ou association) possèdent le même nombre élevé d'incompatibilités et d'ambiguïtés alors on retient la relation de correspondance ayant le score le plus élevé et on rejette l'autre. Si les relations ayant deux types différents (équivalence et association) possèdent le même nombre élevé d'incompatibilités et d'ambiguïtés alors nous retenons la relation d'équivalence ou nous proposons à l'ingénieur de la connaissance d'intervenir pour choisir l'une des deux.
3) mise à jour de L_{incomp}, L_{amb} et $L_{ambEqAss}$ (suppression des couples de relations qui sont dépendants de la relation supprimée) ;
4) mise à jour du nombre d'incompatibilités et d'ambiguïtés des liens qui sont dépendants de la relation supprimée.

L'exemple 4.7 montre l'application de notre algorithme. Dans cet exemple, nous disposons de 4 entités dans O_1 (e_1, e_2, e_3 et e_4) et 4 entités dans O_2 (e'_1, e'_2, e'_3 et e'_4). L'alignement de ces ontologies a permis d'obtenir 6 relations qui posent problème (eq_1, eq_2, eq_3, eq_4, eq_5 et ass_1). Dans l'itération 1, nous obtenons quatre relations de même type ayant le même nombre d'incompatibilités et d'ambiguïtés (3). Nous avons supposé que eq_1 a un score le plus faible pour supprimer toutes les correspondances qui sont dépendantes à cette eq_1. Le même raisonnement dans les autres itérations (ex. dans l'itération 2, eq_2 est supprimée car nous avons supposé qu'elle possède un score faible).



Itération 1	Itération 2	Itération 3	Itération 4	Itération 5
$L_{incomp} = \{(eq_1, eq_2); (eq_1, eq_3); (eq_2, eq_5); (eq_3, eq_5)\}$ $L_{amb} = \{(eq_2, eq_3); (eq_1, eq_5)\}$ $L_{ambEqAss} = \{(eq_4, ass_1)\}$	$L_{incomp} = \{(eq_2, eq_5); (eq_3, eq_5)\}$ $L_{amb} = \{(eq_2, eq_3)\}$ $L_{ambEqAss} = \{(eq_4, ass_1)\}$	$L_{incomp} = \{(eq_3, eq_5)\}$ $L_{amb} = \varnothing$ $L_{ambEqAss} = \{(eq_4, ass_1)\}$	$L_{incomp} = \varnothing$ $L_{amb} = \varnothing$ $L_{ambEqAss} = \{(eq_4, ass_1)\}$	$L_{incomp} = \varnothing$ $L_{amb} = \varnothing$ $L_{ambEqAss} = \varnothing$
relations | Nb. incomp.& ambig. eq_1 | 3 eq_2 | 3 eq_3 | 3 eq_4 | 1 eq_5 | 3 ass_1 | 1	relations | Nb. incomp.& ambig. eq_2 | 2 eq_3 | 2 eq_4 | 1 eq_5 | 2 ass_1 | 1	relations | Nb. incomp.& ambig. eq_3 | 1 eq_4 | 1 eq_5 | 1 ass_1 | 1	relations | Nb. incomp.& ambig. eq_4 | 1 eq_5 | 0 ass_1 | 1	relations | Nb. incomp.& ambig. eq_4 | 0 eq_5 | 0

FIGURE 4.7 – Phase d'alignement : exemple de l'élimination des configurations anormales

Dans la section suivante, nous décrivons les correspondances remarquables qui permettent de marquer les entités intéressantes à exploiter dans les différentes ontologies.

4.3 Détection et affichage de correspondances remarquables

Notre idée est née du fait qu'on veut présenter à l'ingénieur de la connaissance les résultats d'alignement de manière à l'assister lors du processus de construction d'une ressource sémantique. Après la correction des anomalies, nous avons différentes informations qui facilitent le travail de restitution à l'ingénieur de la connaissance : (1) le nombre de liens qui partent de la même entité, et (2) la distance des entités mises en relation par rapport aux feuilles.

Notre but est donc d'attirer l'attention de l'ingénieur de la connaissance sur des correspondances entre les ontologies qui semblent intéressantes.

Nous proposons pour cela de détecter des configurations dites remarquables. Une configuration remarquable est une configuration qui regroupe des correspondances qui présentent des caractéristiques spécifiques qui mettent en valeur les entités sémantiques mises en relation. Ces caractéristiques portent sur la position des entités par rapport à la racine et aux feuilles ainsi que sur le nombre de liens partagés entre entités.

Nous présentons donc dans la cartographie, des correspondances avec des indications sur des entités pour montrer à l'ingénieur de la connaissance que ces liens sont remarquables.

Cette section est organisée comme suit : nous détaillons dans un premier temps deux configurations remarquables qui présentent : (1) une différence de granularité sémantique, (2) plusieurs liens d'association. Nous présentons ensuite l'affichage de ces configurations.

4.3.1 Configurations remarquables

Dans cette section, nous définissons deux types de configurations remarquables :

1. configuration avec une différence de niveau de généralité : deux entités sont mises en correspondance soit par un lien d'équivalence soit par une association, mais ne sont pas classées au même niveau hiérarchique.
2. configuration avec plusieurs liens d'association : une entité d'une première ontologie est en relation d'association avec plusieurs entités distinctes d'une deuxième ontologie.

Configuration avec une différence de granularité sémantique C_{gs}

Définition 4 *Une configuration avec une différence de niveau de généralité est une configuration où deux entités mises en correspondance, avec un lien de type équivalence ou association, possèdent un niveau de généralité différent dans la description sémantique des deux ontologies (voir figure 4.8). Le niveau de généralité d'une entité d'ontologie est exprimé par sa hauteur par rapport aux feuilles (représentées par le plus lointain descendant).*

On note une telle configuration $C_{gs}(eqAss)$ où $eqAss$ peut être un lien d'équivalence ou un lien d'association avec une différence de niveau de généralité.

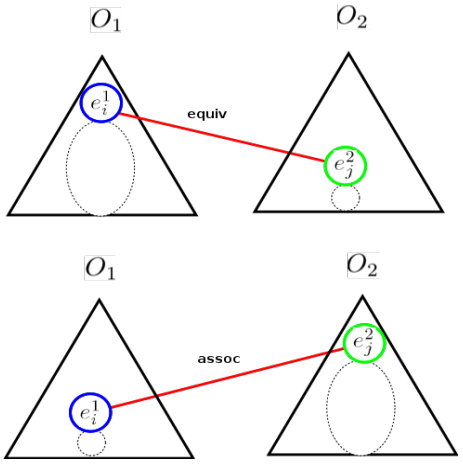


FIGURE 4.8 – Schéma de la configuration avec une différence de granularité sémantique

Exemple 4 *La figure 4.9 montre une différence de granularité sémantique (avec une relation d'équivalence) des deux entités « soil » de OntoBiotope et « soil » de EnvO. La première entité de l'ontologie OntoBiotope possède un niveau de généralité de 5 alors que la même entité dans EnvO est à un niveau de 3.*

Méthode de détection de la configuration remarquable La méthode ci-dessous teste si un ensemble de correspondances de type équivalence ou association se combine avec une différence de granularité sémantique. Cette méthode prend en entrée un ensemble de liens d'équivalence ou d'association $EQG_{12} = \{< idE, e_x^1, e_y^2, score, equiv > / e_x^1 \in R_1$ et $e_y^2 \in R_2\}$, ou $EQG_{12} = \{< idA, e_x^1, e_y^2, score, assoc > / e_x^1 \in R_1$ et $e_y^2 \in R_2\}$ et donne en sortie une liste de couples d'entités mis en correspondance avec leur granularité sémantique. Nous présentons l'algorithme pour les relations d'équivalence.

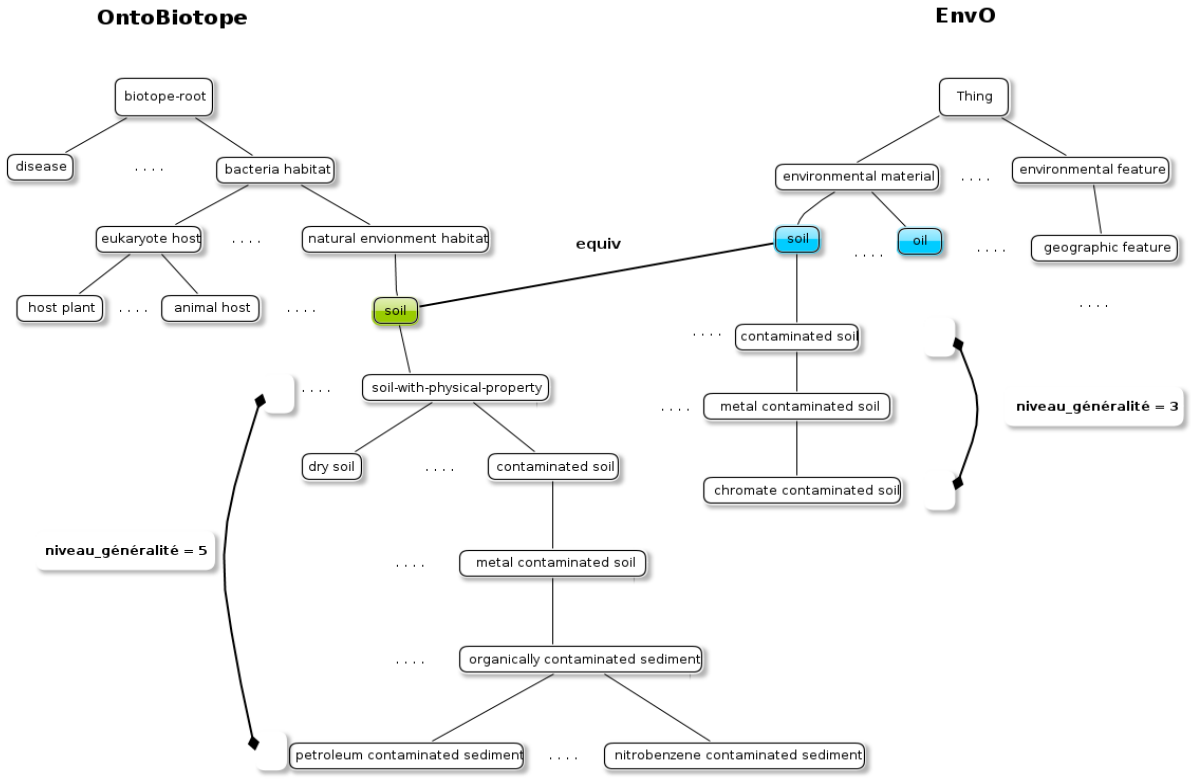


FIGURE 4.9 – Exemple de la configuration avec une différence de granularité sémantique

Algorithme 7 *GranularityEntitityEqAss*() : liste de couples d'entités mis en correspondance avec leur granularité sémantique

Entrée : $EQG_{12} = \{< idE, e_x^1, e_y^2, score, equiv >\}$
Sortie : $L_{GranEntityEq}$: liste de couples d'entités mises en correspondance avec l'équivalence associées à leur granularité sémantique.
Variables : $niveau - generalite(e_i)$, $niveau - generalite(e_j)$: entiers
Début
pour tout $(eq_{ij} \in EQG_{12})$ **faire**
  pour tout ij allant de 1 à $taille(EQG_{12})$ **faire**
    %% Tester si $niveau - generalite(e_i) > niveau - generalite(e_j)$
    $niveau - generalite(e_i) = nbrArcs - entre(e_i, feuille(e_i))$
    $niveau - generalite(e_j) = nbrArcs - entre(e_j, feuille(e_j))$
    si $(niveau - generalite(e_i) > niveau - generalite(e_j))$ **alors**
      inserer$(e_i, e_j, niveau - generalite(e_i), niveau - generalite(e_j), L_{GranEntityEq})$
    fin si
  fin pour
fin pour
retourne $L_{GranEntityEq}$
Fin

Configuration avec plusieurs liens d'association C_{plAss}

Définition 5 *Une configuration avec plusieurs liens d'association est une configuration où une entité est reliée par au minimum deux relations de correspondance de type association, avec d'autres entités. On parle de la centralité d'une entité d'ontologie qui est une valeur d'intérêt portée à cette entité. Cette centralité est décrite par le nombre de liens de type association qui relient une entité à d'autres entités. Plus une entité possède des relations de correspondance de type association vers d'autres entités, plus cette configuration est considérée comme remarquable.*

On note une telle configuration $C_{plAss}(ass_{ij}, ass_{uv})$ *où* $ass_{ij} < idA_x, e_i^1, e_j^2, score, assoc >$ *et* $ass_{uv} < idA_y, e_u^1, e_v^2, score, assoc >$ *et* $x \neq y$, *sont deux liens d'association contenant des entités centrales (voir figure 4.10).*

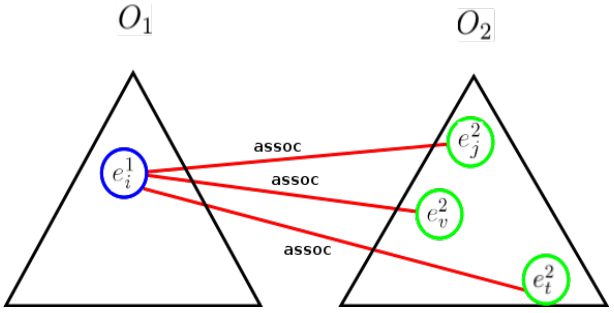


FIGURE 4.10 – Schéma de la configuration avec plusieurs liens d'association

Exemple 5 *La figure 4.11 montre une configuration remarquable où une entité sémantique est reliée à différentes autres entités avec des relations de type association. L'entité « China » est rapprochée de trois entités différentes « petroleum », « cut » et « city ».*

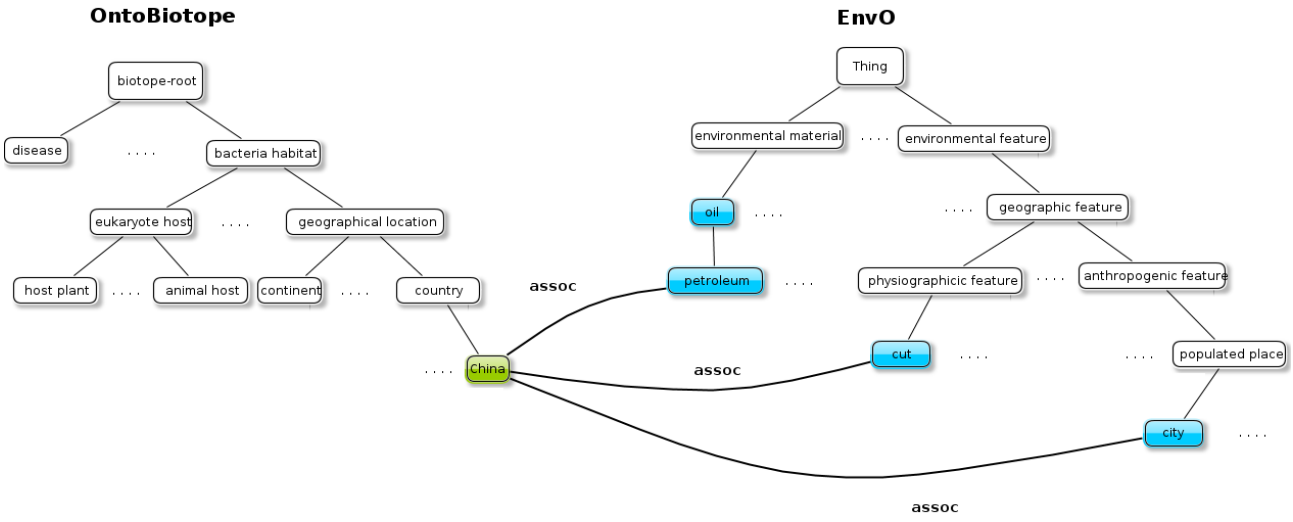


FIGURE 4.11 – Exemple de la configuration avec plusieurs liens d'association

Méthode de détection de la configuration remarquable La méthode ci-dessous teste si les correspondances de type association sont remarquables. Cette méthode prend en entrée un ensemble de liens d'association $EplAss_{12} = \{< idA, e_x^1, e_y^2, score, assoc > / e_x^1 \in R_1 \text{ et } e_y^2 \in R_2\}$, et donne en sortie une liste d'entités mises en relation avec le nombre de liens de type association qu'elles partagent avec les autres entités sémantiques.

Algorithme 8 $MultipleLinksAss()$: liste d'entités mises en relation avec le nombre de liens de type association

1: **Entrée :** $EplAss = \{< idA, e_x^1, e_y^2, score, assoc >\}$
2: **Sortie :** L_{plAss} : liste d'entités mises en relation de type association avec le nombre de liens
3: **Variables :** $plAss$: entier
4: **Début**
5: **pour tout** $ass_{ij} \in EplAss_{12}$ *ET* $ass_{uv} \in EplAss_{12}$ **faire**
6: **pour tout** ij allant de 1 à $taille(EplAss_{12})$ **faire**
7: **pour tout** uv allant de $ij + 1$ à $taille(EplAss_{12})$ **faire**
8: %% Tester si ass_{ij} et ass_{uv} comportent des entités centrales
9: $plAss = \underline{ComputeLinks}(ass_{ij}, ass_{uv})$
10: **si** $plAss >= 2$ **alors**
11: inserer$(e_i^1, plAss, L_{plAss})$
12: **fin si**
13: **fin pour**
14: **fin pour**
15: **fin pour**
16: retourne L_{plAss}
17: **Fin**

Algorithme 9 $ComputeLinks(ass_{ij}, ass_{uv})$

1: **Sortie :** $plAss$: entier
2: **Début**
3: $plAss = 0$
4: **si** $(e_i^1 == e_u^1) \wedge (e_j^2$ et $e_v^2)$ sont distinctes **alors**
5: $plAss = plAss + 1$
6: **fin si**
7: retourne $plAss$
8: **Fin**

4.3.2 Affichage de correspondances remarquables

L'affichage des correspondances remarquables consiste à présenter à l'ingénieur de la connaissance les configurations obtenues des types ci-dessus. Nous proposons donc pour la configuration avec une différence de granularité sémantique et la configuration avec plusieurs liens d'association de présenter des informations complémentaires sur les entités mises en correspondance à savoir : (i) leur position par rapport aux feuilles ($< e_i^1, degre -$

$detail(e_i^1), e_j^2, degre - detail(e_j^2) >$ où $e_i^1 \in O_1$ et $e_j^2 \in O_2$), et (ii) le nombre de liens de type association qu'une entité partage avec les autres entités ($< e_i^x, assoc, nbrLienAss >$ où : $e_i^x \in O_1 \bigvee O_2$ et $nbrLienAss$: nombre de liens que e_i^x partage avec les autres entités).

En terme d'implémentation, le fichier RDF ou textuel qui a été obtenu dans la phase d'alignement va être nettoyé au niveau de l'étape de détection et de l'élimination des anomalies puis utilisé pour présenter les configurations remarquables.

4.4 Conclusion

Dans ce chapitre, nous avons présenté un ensemble de configurations qui révèlent des anomalies et ou des zones remarquables dans les résultats d'alignement. En raisonnant sur la structure des ontologies, nous avons détecté trois configurations liées à l'incompatibilité et l'ambiguïté des liens et nous avons proposé une méthode pour les résoudre. Nous avons également proposé deux configurations remarquables qui font ressortir les entités qui semblent particulièrement intéressantes. Les algorithmes de détection des anomalies et des configurations remarquables ont été implémentés en Java et en utilisant la librairie Jena.

La cartographie de domaine obtenue est donc un ensemble de liens de correspondance établis entre deux ontologies alignées entre elles et ancrées dans un texte choisi comme référence, assorti d'un outil d'exploration qui permet de repérer facilement les anomalies et les configurations remarquables, de les résoudre au besoin et de les afficher. La figure 4.12 montre sous la forme d'une maquette comment une telle cartographie peut être affichée.

A ce stade, nous avons mis l'accent sur l'alignement de deux ontologies lexicalisées mais il faudrait à terme étendre l'approche à plus de ressources et d'autres types de ressources.

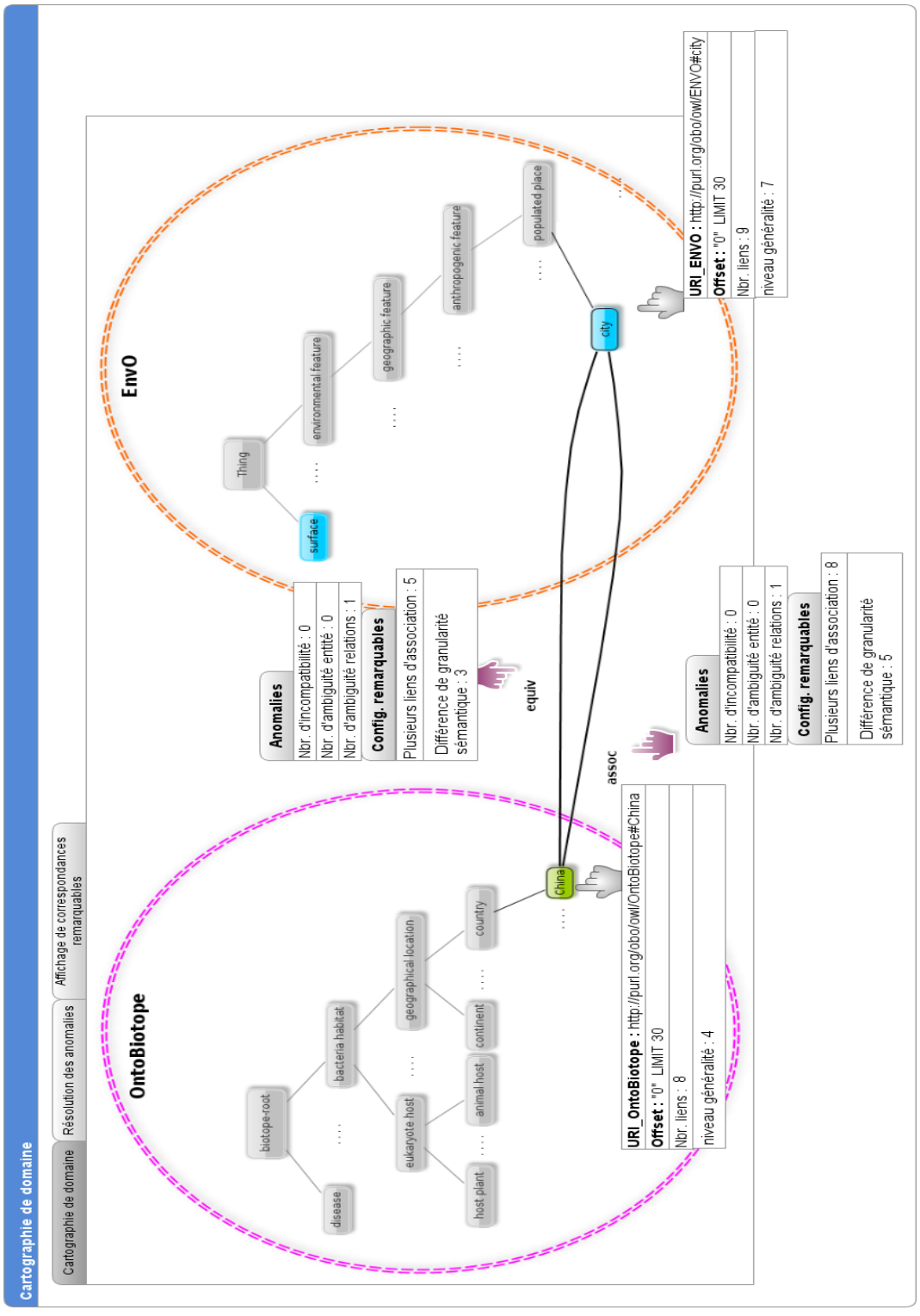


FIGURE 4.12 – Maquette de l'interface de la cartographie de domaine

Chapitre 5

Expériences

Sommaire

5.1 Introduction

Ce chapitre présente les expériences que nous avons menées pour mettre en œuvre et tester la méthodologie proposée dans le chapitreřefchap :methodocarto, notamment les phases d'alignement et de cartographie détaillées dans les chapitres 3 et 4. En pratique nous avons testé notre approche sur trois cas d'usage se rapportant à trois domaines différents : le domaine biologique, alimentaire et géographique. Ce chapitre montre les ressources que nous avons utilisées, la manière dont elles ont été préparées et les résultats que nous avons obtenus pour chaque domaine et à chaque étape (annotation, alignement, construction de la cartographie).

Le chapitre présente, dans un premier temps, les cas d'usage étudiés puis les expérience réalisées pour tester tour à tour les différentes phases de la méthodologie proposée : nous montrons les résultats obtenus dans la phase d'annotation pour les trois domaines étudiés, puis nous détaillons les résultats obtenus pour l'alignement des ontologies et nous présentons les différentes configurations détectées.

5.2 Cas d'usage

Nos expériences ont été faites sur des cas d'usage de trois domaines différents : biologique, alimentaire et géographique. Nous avons récupéré les données du domaine biologique du laboratoire INRA-MIG. Les données alimentaires ont été fournies par le laboratoire INRA Méta@risk. Quant aux données du domaine géographique, nous les avons obtenues du laboratoire COGIT de l'IGN.

Dans ce contexte applicatif et pour le traitement automatique du langage naturel, nos partenaires de recherche se sont intéressés à (i) construire une ressource sémantique de domaine et (ii) annoter le texte décrivant un domaine particulier.

Dans cette section, nous décrivons le processus de collecte des ressources dans les trois domaines de spécialité à savoir les textes d'acquisition et les ontologies de domaine.

5.2.1 Domaine biologique

Le domaine biologique est assez large, nous nous focalisons plus particulièrement sur les biotopes bactériens. Ce domaine comporte l'ensemble des facteurs et des variables environnementales qui hébergent les bactéries. En d'autres termes, ce domaine permet d'étudier la localisation et la mobilité des bactéries dans : (i) les milieux naturels comme l'eau, le sol, l'intestin, etc, et (ii) dans des milieux artificiels comme les produits surgelés.

Nous disposons dans ce domaine d'un texte de référence de BioNLP-ST 2011 [1], nommé *Bacteria Biotope* (BB). Ce texte a été construit manuellement à partir de plusieurs pages Web comme les articles de MicrobeWiki [2] et les pages sur les génomes [3]. Ce texte décrit plus précisément divers types de bactéries avec différentes caractéristiques. Il contient 8 520 mots et 541 phrases.

Dans ce même domaine, notre partenaire INRA-MIG, dans le programme Quaero [4], nous a fourni une ontologie nommée OntoBiotope. C'est une ontologie lexicalisée qui décrit les différents aspects de la localisation (les habitats) des bactéries. Elle a été construite en décembre 2009 par l'équipe de Claire Nédellec, d'une manière semi-automatique à partir de plusieurs documents textuels (ex. articles JSEM), de bases de données (ex. GenBAnk, GOLD) et en réutilisant des classifications existantes (ATCC, EnvO, Metagenome au JGI). Cette ontologie a été créée dans le but d'extraire et de normaliser les informations sur les biotopes [Golick *et al.*, 2011].

La deuxième ontologie du domaine est EnvO (Environment Ontology). Elle résulte du projet EnvO [5] qui vise à assister la description sémantique et cohérente des informations environnementales. C'est une ontologie considérée comme standard pour la description de l'environnement biologique des organismes. Elle décrit les biomes, qui sont des unités écologiques caractérisées par plusieurs facteurs tels que les espèces végétales (ex. arbres) et animales. Ces biomes sont de deux types : aquatique et terrestre. EnvO présente aussi les propriétés environnementales (ex. propriétés géographiques, habitat) ainsi que les cadres de vie des organismes à savoir l'air, l'eau, le sol. Elle a été construite par un consortium spécialiste en environnement et en partie par OBO (Open Biomedical Ontologies foundry) [6]. La première version de cette ontologie date de 12/07/2010. EnvO est une ontologie publiée sur le Web. EnvO et OntoBiotope sont lexicalisées en anglais et sont des ontologies de taille moyenne (voir tableau 5.1).

1. `http://weaver.nlplab.org/`
2. `http://microbewiki.kenyon.edu/index.php/MicrobeWiki`
3. `http://www.genoscope.cns.fr/spip/`
4. Le programme franco-allemand Quaero [Quaero, 2008] vise à réaliser des applications pour l'accès à l'information des documents multimédia (ex. textuels, images, vidéo, musique).
5. `http://www.environmentontology.org`
6. `http://obofoundry.org/`

Ontologies	Entités sémantiques		
	#concepts	#*is-a*	#rôles
OntoBiotope	1575	1513	184
EnvO	1557	1590	184

Tableau 5.1 – Nombre des entités sémantiques de EnvO et OntoBiotope

L'équipe de l'INRA-MIG a utilisé ces deux ontologies, pour annoter des textes biologiques avec les entités de ces ontologies. Ce travail vise, à travers l'annotation des habitats et des propriétés environnementales des bactéries, à normaliser et obtenir une base commune et standard des habitats.

5.2.2 Domaine alimentaire

Nous nous sommes intéressés au domaine des risques alimentaires auxquels est exposé un individu suite à la prise d'un aliment. Les dangers sont de deux types : (i) les risques microbiens associés aux aliments, et (ii) les risques chimiques des aliments. Ces deux types de risques peuvent causer des préjudices à la santé humaine. Les risques microbiens proviennent notamment des bactéries et des insectes. On les trouve généralement en boucherie et dans les pâtisseries. Les risques chimiques proviennent des substances chimiques contenant les toxines, des produits chimiques causant une réaction dangereuse telle que l'allergie chez l'homme et des produits chimiques utilisés dans la transformation des aliments.

L'équipe de l'INRA Méta@risk [7] nous a fourni trois ressources en anglais (mais les ontologies sont aussi enrichies par des étiquettes de concepts en français) : (1) un texte de référence nommé Emballage, et (2) deux ontologies lexicalisées : ontoMicrobio et ontoChemical. Le texte de référence est construit manuellement à partir de pages wikipédia. Il décrit le rôle de l'emballage alimentaire pour assurer la protection des aliments. Il comporte 1 657 349 mots et 29 250 phrases.

Dans le contexte de la construction d'entrepôts de données thématiques ouverts sur le Web, INRA Méta@risk a créé un système appelé ONDINE (ONtology based Data INtEgration) qui consiste à aider ceux qui travaillent dans le domaine de l'alimentation à enrichir leur base de connaissances. Dans ce système, nous trouvons les deux ontologies lexicalisées ontoMicrobio et ontoChemical [8] qui ont été construites en 2011 manuellement pour non seulement annoter sémantiquement un ensemble de tableaux [Hignette *et al.*, 2009] mais aussi interroger ces tableaux annotés [Buche *et al.*, 2012].

Le tableau 5.2 montre le contenu des deux ontologies ontoMicrobio et ontoChemical.

7. `http://www7.paris.inra.fr/metarisk/`

8. `http://www7.paris.inra.fr/metarisk/research_unit/knowledge_engineering/softwares/ondine_corpora_and_otr_july_2011`

Ontologies	Entités sémantiques		
	#concepts	#*is-a*	#rôles
ontoChemical	8223	24782	15
ontoMicrobio	2341	7279	15

Tableau 5.2 – Nombre des entités sémantiques de ontoMicrobio et ontoChemical

5.2.3 Domaine géographique

Ce domaine permet de décrire les données topographiques sous forme de métriques. Il spécifie par exemple les éléments de paysage, l'ensemble de départements français, etc.

Nous disposons dans ce domaine de deux taxonomies lexicalisées en français : BDTopo et BDCarto, et un texte de référence. Les deux taxonomies nous ont été fournies par le laboratoire de COGIT de l'IGN [9] dont l'objectif est la localisation de l'information relative aux problématiques d'aménagement du territoire, d'environnement ou d'urbanisme. Elles décrivent, les formes topographiques sur un terrain, notamment les formes naturelles (hydrographie, relief) et artificielles (musée, chemin, etc). Ces deux taxonomies ont été construites en 2006 de manière semi-automatique [Abadie et Mustière, 2010] à partir de documents textuels de spécifications associés à deux bases de données géographiques, respectivement BDTopo et BDCarto. Ces taxonomies ont été utilisées dans le cadre du projet GeOnto [10] pour différentes applications, à savoir l'appariement de schémas de bases de données géographiques et la formalisation, la consultation des spécifications [Abadie, 2012]. Ces deux taxonomies ne sont pas publiées sur le Web.

Dans ce domaine, nous avons deux textes d'acquisition :

- un premier texte que nous avons récupéré du portail documentaire de la cité de l'architecture et du patrimoine [11]. Il a été créé par le centre d'archives d'architecture de l'institut français d'architecture. Ce texte contient 20 990 mots et 1 292 phrases.
- un deuxième texte fourni par le T2i du LIUPPA, qui a été créé dans le cadre du projet GeOnto. Ce texte est nommé « récits de voyage ». Il décrit les anciens voyages dans les Pyrénées et il est issu de documents patrimoniaux disponibles sur l'ensemble du territoire des Pyrénées françaises et espagnoles. Ce texte contient 106 851 mots et 5 160 phrases.

Le tableau 5.3 montre le contenu des deux taxonomies BDTopo et BDCarto.

Taxonomies	Entités sémantiques	
	#concepts	#*is-a*
BDTopo	612	617
BDCarto	505	486

Tableau 5.3 – Nombre des entités sémantiques de BDTopo et BDCarto

9. COGIT : laboratoire de Conception Objet et Généralisation de l'Information Topographique de l'IGN (`http://www.ign.fr/`), Institut Géographique National, Saint Mardé
10. `http://geonto.lri.fr`
11. `http://portaildocumentaire.citechaillot.fr`

5.3 Résultats d'annotation sémantique

Le processus d'annotation prend en entrée 2 ontologies lexicalisées sous format OWL et fournit en sortie un texte annoté par les entités des ontologies. Ce processus comprend trois étapes : (1) conversion : toutes les ontologies des cas d'usage décrits plus haut sont converties à un format unique OWL, (2) lemmatisation : nous utilisons l'étiqueteur morpho-syntaxique TreeTagger [12] qui prend en paramètre la langue (ex. french, english) pour catégoriser (nom, verbe, adverbe) et mettre sous forme canonique (ex. infinitif pour les verbes) les étiquettes associées aux concepts des ontologies ainsi que les unités textuelles des textes, et (3) ancrage : chacune des étiquettes lemmatisées des ontologies est comparée aux lemmes (formes canoniques) des unités textuelles (comparaison de chaînes de caractères). Le principe de cette étape est de projeter les étiquettes lemmatisées des concepts d'ontologies sur le texte afin d'identifier, ce qu'on appelle les entités ancrées (étiquettes présentes dans le texte).

C'est une annotation sémantique triviale et nous avons eu un certain nombre de problèmes liés au rapprochement entre le texte et les ontologies, suite à un nommage hétérogène des entités d'ontologies. Nous traitons ce problème en appliquant quelques règles pour obtenir de meilleurs résultats. Citons par exemple :

- remplacer les tirets "-" ou le trait de soulignement "_" par des espaces blancs. Par exemple, nous avons remplacé l'entité « chemin_d_exploitation » par « chemin d exploitation » ;
- découper les mots contenant des majuscules à l'intérieur en plusieurs mots. Par exemple, nous avons remplacé « SpeedUnit » par « Speed Unit » ;
- enlever les mots inutiles comme T_Fr_ et T_En_ qui sont collés aux étiquettes de concepts pour mentionner la langue utilisée. Par exemple, nous avons remplacé « T_En_Rice_pudding » et « T_Fr_Filet_de_morue » par « Rice pudding » et « Filet de morue ».

Chaque texte d'un domaine donné est décrit par #mots_différents, le nombre de mots différents après élimination des mots vides (ex. préposition), #UT, le nombre d'unités textuelles liées aux concepts des ontologies et #EA, le nombre de concepts d'ontologies qui sont trouvés dans le texte. Dans le domaine biologique, #EA = 125, avec 91 concepts issus de l'ontologie OntoBiotope (associées à 379 #UT) et 34 concepts issus de l'ontologie EnvO (correspondants à 117 #UT). 209 #EA dans le domaine géographique dont 124 issues de BDTopo et 85 issues de BDCarto. On constate que le taux de couverture du texte par les concepts des ontologies varie d'un domaine à l'autre. En effet, le fait de ne s'intéresser qu'aux concepts des ontologies et de mettre en place une comparaison stricte entre les étiquettes des concepts et les mots peut influencer le degré de couverture. De plus, les ontologies présentées dans le tableau 5.4 ne sont pas construites pour le même objectif. La phase d'annotation fait l'hypothèse qu'il y a un sens unique attribué à une même unité et fait donc l'impasse sur la désambiguïsation.

12. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Domaines	Texte	#mots différents	Ressources	#concepts	#UT	#EA
Domaine biologique	Bacteria Biotope	2755	OntoBiotope EnvO	1575 1557	379 117	91 34
Domaine alimentaire	Emballage	69958	OntoChemical OntoMicrobio	8223 2341	1181 1149	398 112
Domaine géographique	Portail documentaire	5725	BDTopo BDCarto	612 505	398 240	46 46
	Récits de voyage	19571	BDTopo BDCarto	612 505	1085 534	124 85

Tableau 5.4 – Nombre des ressources des cas d'usage dans la phase d'annotation sémantique

La figure 5.1 montre un exemple d'annotation sémantique du texte du portail documentaire au regard des ontologies du domaine géographique (BDTopo et BDCarto). Nous faisons la différence entre trois types d'entités : celles appartenant aux deux ontologies (couleur rouge), celles appartenant à l'ontologie BDTopo (couleur bleue) et celles appartenant à l'ontologie BDCarto (couleur verte). Dans cet exemple d'annotation, 5 entités de BDTopo ont été ancrées dans le texte avec 2 entités issues de BDCarto et 4 entités communes ancrées des deux ontologies. A titre d'exemple, le concept commun « commune » a deux occurrences dans le texte.

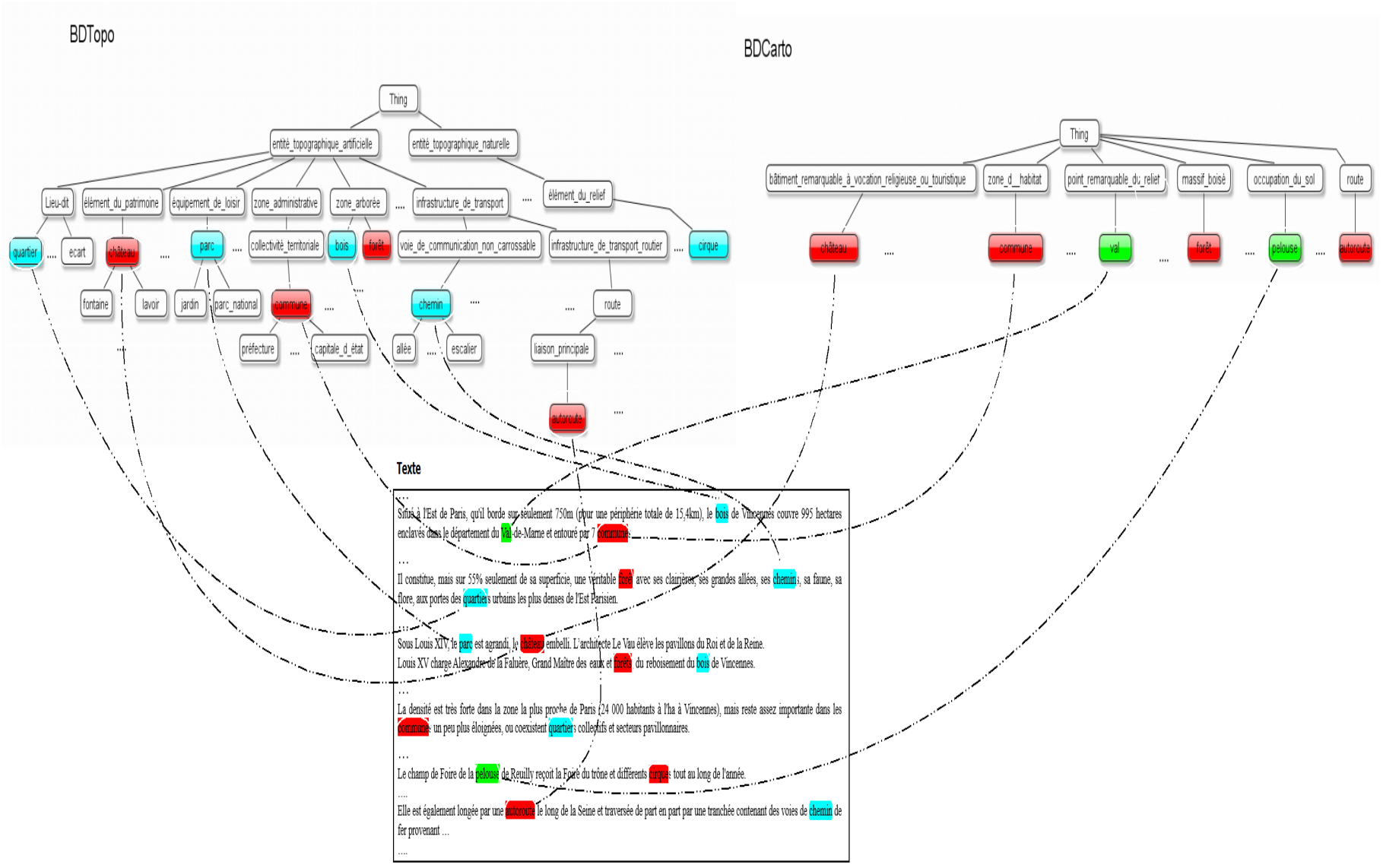


FIGURE 5.1 – Annotation sémantique du texte du portail documentaire avec les deux ontologies BDTopo et BDCarto

5.4 Résultats des alignements des ontologies

Nous avons implémenté un outil d'alignement automatique d'ontologies guidé par le texte appelé *TOM* (Text-based Ontology Mapping). Cet outil prend en entrée deux ontologies sous format OWL et un texte lemmatisé, découpé en phrases et annoté. Il donne en sortie un ensemble de correspondances. Ces correspondances sont générées dans deux formats différents présentés dans le chapitre 3 :

- la syntaxe du format standard [13] utilisé par plusieurs outils dans la campagne d'évaluation OAEI pour pouvoir se comparer à d'autres outils d'alignement ;
- le format textuel pour faciliter la compréhension des correspondances par l'ingénieur de la connaissance. Ce format est sous forme de 5-uplets présenté dans le chapitre 1.

Le tableau 5.5 montre l'alignement lexical entre les ontologies en se fondant sur les textes associés aux trois domaines (biologique, alimentaire et géographique). Nous rappelons que cette phase comporte deux étapes : (1) le calcul de correspondances, et (2) le filtrage (détaillé dans le chapitre 3). On nomme les relations obtenues dans la première étape comme des « relations possibles » et les relations de la deuxième étape comme des « relations candidates ». Le nombre de relations possibles (équivalence ou association) correspond à la combinaison des entités ancrées des 2 ontologies. Les relations candidates sont obtenues suite à une sélection sur les relations précédentes.

Le tableau 5.5 décrit la sortie d'alignement en termes de nombre de relations d'équivalence et d'association. Nous avons appliqué, dans tous les domaines, la mesure de Jaccard pour les scores de co-occurrence et la mesure du cosinus pour les scores de similarités. On constate que dans tous les domaines, le nombre de relations d'équivalence est supérieur à celui d'association. Cela est dû au fait que les correspondances de type association ne sont pas évidentes à retrouver.

La valeur de seuil fixée dans l'étape de filtrage joue un rôle essentiel pour supprimer les relations possibles inutiles. Par exemple, dans le domaine biologique, 185 relations candidates (parmi 3094 relations possibles) ont été retenues dont 33 relations de type association et 152 relations de type équivalence.

5.5 Résultats de l'analyse des configurations

Une fois que l'alignement des ontologies au regard du texte est établi, nous nous intéressons à la dernière phase de construction de la cartographie de domaine. Cette phase contient une étape de détection des anomalies et une étape d'affichage des correspondances remarquables que nous avons présentées dans le chapitre 3.

Le processus de détection des configurations prend en entrée un ensemble de correspondances sous format RDF/XML ou textuel (5-uplets) et la paire d'ontologies sous format OWL, et donne en sortie un ensemble de configurations : anormales et remarquables. Cette sortie est considérée comme un ensemble de suggestions à l'ingénieur de la connaissance qui peut les réviser manuellement ou automatiquement.

Le tableau 5.6 montre le nombre de configurations anormales obtenues dans chaque

13. `http://alignapi.gforge.inria.fr/format.html`

Domaines	**Rel. sémantiques**		**#relations**	
			#rel. possibles	#rel. candidates
Domaine biologique	#rel. association #rel. équivalence		3094	33 152
Domaine alimentaire	#rel. association #rel. équivalence		44352	552 3234
Domaine géographique	Portail documentaire	#rel. association #rel. équivalence	2116	14 110
	Récits de voyage	#rel. association #rel. équivalence	10540	26 1270

Tableau 5.5 – Relations sémantiques (association et équivalence) entre les ressources des cas d'usage

domaine et le nombre des correspondances retenues après la résolution des anomalies de manière automatique. On constate que le nombre d'anomalies de la hiérarchie inversée C_{hi} détectées est faible par rapport aux autres configurations : les ontologies qui ont été alignées possèdent peu d'entités qui font partie de ces correspondances et qui sont structurées d'une manière hiérarchique inversée. Dans le domaine alimentaire, le nombre de configurations avec une entité ambigüe C_{AmEq} est élevé par rapport aux autres domaines : les ontologies alignées comportent plusieurs relations d'équivalence ambigües. Si on procède à une résolution automatique de ces anomalies, presque dans tous les domaines, 50% des relations candidates présentées dans le tableau 5.5, ont été retenues. Ces résultats sont encourageants et montrent que notre méthode conduit à extraire les relations retenues entre les ontologies d'un domaine particulier.

Domaines		**Config. anormales**			**#rel. retenues**	
		$\#C_{hi}$	$\#C_{AmEq}$	$\#C_{AmEqAss}$	#rel. association	#rel. équivalence
Domaine biologique		1	109	30	13	78
Domaine alimentaire		3	2232	507	153	2024
Domaine géographique	Portail documentaire	2	103	12	9	54
	Récits de voyages	2	1053	16	8	527

Tableau 5.6 – Configurations anomales détectées dans les cas d'usage

Le tableau 5.7 montre le nombre de configurations remarquables obtenues lors de la révision de la sortie d'alignement. Le nombre de configurations avec une différence de granularité sémantique C_{gs} est important dans tous les domaines et pour les deux types de relations d'association (assoc) et d'équivalence (equiv). 101 configurations remarquables, par exemple, ont été retrouvées dans le domaine biologique dont 26 configurations sont reliées aux relations d'association et 75 configurations sont reliées aux relations d'équivalence. En revanche, le nombre de configurations avec plusieurs liens d'association C_{plAss}

est relativement petit, ce qui met en évidence peu d'entités centrales.

Domaines		**Config. remarquables**	
		$\#C_{gs}$	$\#C_{plAss}$
Domaine biologique		26 (assoc) et 75 (equiv)	7
Domaine alimentaire		105 (assoc) et 295 (equiv)	232
Domaine géographique	Portail documentaire	10 (aasoc) et 32 (equiv)	4
	Récits de voyages	16 (assoc) et 81 (equiv)	6

Tableau 5.7 – Validation des configurations remarquables affichées dans les cas d'usage

5.6 Conclusion

Dans ce chapitre, nous avons présenté les expériences que nous avons menées pour tester les différentes étapes de notre méthodologie de construction de la cartographie de domaine. Nous nous sommes intéressés à trois cas d'usage qui relèvent de domaines bien différents, les domaines géographique, alimentaire et biologique. En entrée de chaque cas d'usage, nous avons des ontologies et un texte de référence. Nous avons pu montrer que la phase d'annotation est nécessaire pour désambiguïser les mots du texte par les entités ontologiques qui les annotent. Dans la phase d'alignement, nous avons constaté que le nombre de correspondances validées ou corroborées par le texte augmente avec le nombre d'entités ancrées. Enfin, nous avons obtenu des résultats encourageants dans la dernière phase de la méthodologie.

Ces premiers résultats montrent que notre méthodologie a bien le comportement attendu mais il reste important d'évaluer les résultats obtenus. C'est ce que nous présentons dans le chapitre suivant en mettant l'accent sur l'évaluation de notre méthode d'annotation guidée par le texte.

Chapitre 6

Evaluation

Sommaire

6.1 Introduction

Ce chapitre met l'accent sur l'évaluation de notre outil d'alignement TOM (Text-based Ontology Mapping). Nous présentons trois expérimentations. La première compare notre approche TOM et celle de TaxoMap, un outil de l'état de l'art, et montre que l'approche TOM est complémentaire des approches d'alignement existantes. Cette comparaison est réalisée dans le domaine biologique. Nous présentons ensuite le protocole que nous avons défini pour une expérience de validation humaine des correspondances : cette expérience est aujourd'hui prête à être lancée mais nous n'avons pas encore de résultats. La troisième expérimentation compare TOM aux systèmes d'alignement disponibles dans la campagne d'évaluation de l'alignement d'ontologies (OAEI). Nous étudions par ailleurs le choix du texte de référence et son influence sur les résultats d'alignement. Ces expérimentations reposent sur des métriques d'évaluation qui sont détaillées dans la section qui suit.

6.2 Métriques d'évaluation

Nous distinguons trois mesures classiques pour mesurer la pertinence des méthodes d'alignement, la précision, le rappel et la f-mesure qui ont le mérite d'être génériques, faciles à interpréter et utilisées dans différents domaines (ex. recherche d'information [Martin *et al.*, 2004]). Ces mesures ont été appliquées aussi dans la campagne d'évaluation OAEI (Ontology Alignment Evaluation Initiative) [1] pour évaluer la qualité des alignements produits par les outils de participants. Évaluer la qualité de l'alignement s'appuie

1. `http://oaei.ontologymatching.org/`

sur 3 étapes : (1) construction d'un alignement de référence (GS pour *Gold Standard*), (2) comparaison du résultat d'alignement fourni (A) avec celui de la référence, et (3) application des mesures de précision, rappel et f-mesure. La précision mesure l'exactitude des correspondances retournées par une méthode d'alignement. Elle donne le pourcentage des correspondances pertinentes retrouvées rapporté au nombre de correspondances total proposées par la méthode :

$$precision\ (P) = \frac{|GS \cap A|}{|A|} \tag{6.1}$$

Le rappel mesure l'exhaustivité des correspondances retournées par une méthode d'alignement. Il donne le pourcentage de correspondances pertinentes retrouvées rapporté au nombre de correspondances total proposées par la référence :

$$rappel\ (R) = \frac{|GS \cap A|}{|GS|} \tag{6.2}$$

La f-mesure est une combinaison des deux mesures définies plus haut (précision et rappel). Elle est définie comme suit :

$$f - mesure\ (FM) = \frac{2 \times Precision \times Rappel}{Precision + Rappel} \tag{6.3}$$

Les mesures d'évaluation précision, rappel et f-mesure sont appliquées lors de l'évaluation de notre méthode d'alignement TOM.

6.3 Comparaison par rapport à un outil de l'état de l'art

Dans cette section, nous comparons notre méthode d'alignement TOM à un outil de l'état de l'art. Notre choix s'est porté sur la méthode d'alignement TaxoMap pour différentes raisons :

- TaxoMap utilise une technique lexicale qui se fonde sur la mesure de similarité Lin [Lin, 1998] pour l'alignement des ontologies, ce qui permet de le comparer à notre outil d'alignement fondé sur une technique lexicale ;
- TaxoMap fournit des correspondances de type équivalence et proximité, ce qui permet de le comparer à notre outil d'alignement qui donne en sortie deux types de relations : équivalence et association ;
- TaxoMap présente de bons résultats sur le challenge OAEI 2007[Zargayouna *et al.*, 2007] et plus spécifiquement dans les tests qui portent sur le niveau lexical [2] (une précision de 92%). Ce résultat est important pour se comparer à TaxoMap surtout au niveau lexcial.

Comme nous l'avons évoqué dans le chapitre de l'état de l'art, les outils d'alignement d'ontologies sont souvent utilisés pour un objectif particulier. Ils permettent d'aligner deux ontologies en se fondant sur leurs spécificités internes (ex. nom, type) et externes (voisinage d'une entité). C'est une hypothèse assez forte qui suppose que l'alignement se fait en dehors

2. `http://oaei.ontologymatching.org/2007/results/benchmarks/index.html`

de toute application. C'est dans cette optique de prise en compte de l'application que nous avons proposé notre méthode d'alignement TOM.

Rappelons que l'alignement fourni par TaxoMap sert à interroger des bases de données hétérogènes. L'alignement est orienté (d'une taxonomie source vers une taxonomie cible) de type 1:n. L'hypothèse de TaxoMap est que la taxonomie source est très peu structurée et la taxonomie cible est bien structurée. Notre méthode TOM a pour but de créer un alignement destiné à la construction d'une ressource sémantique. Cet alignement est guidé par le domaine et l'application. Il n'est pas orienté et il est de type n:m car nous souhaitons obtenir toutes les correspondances possibles et pertinentes à exploiter entre deux ontologies. La méthode TOM repose sur l'hypothèse que les ontologies sont lexicalisées. Nous avons choisi le texte comme support d'informations d'un domaine d'intérêt et des besoins d'une application visée. L'intérêt de TOM n'est pas simplement de retrouver des correspondances entre les entités d'ontologies en appliquant des techniques terminologiques et structurelles mais plutôt de découvrir les correspondances ignorées par les outils existants alors qu'elles sont intéressantes. C'est le rôle du texte dans le cadre de la construction d'une ressource sémantique. TOM est conçue comme une méthode complémentaire aux autres méthodes d'alignement disponibles.

Notre protocole de comparaison consiste à appliquer TaxoMap et TOM sur les mêmes ontologies du domaine biologique. Rappelons que dans ce domaine, nous possédons deux ontologies OntoBiotope et EnvO, et le texte de référence BB de BioNLP-ST 2011. TOM prend en entrée, en plus des ontologies, le texte BB, alors que TaxoMap ne prend en entrée que les deux ontologies. Le tableau 6.1 montre le nombre de correspondances communes et différentes entre TOM et TaxoMap. En appliquant TaxoMap, nous n'avons obtenu que des relations d'équivalence. 2379 relations d'équivalence sont identifiées par TaxoMap mais non retrouvées par TOM parce que les entités mises en correspondance ne sont pas ancrées dans le texte. Il était prévisible que le résultat de TaxoMap comporte plus de relations d'équivalence car il existe plusieurs entités qui sont lexicalement identiques. Cependant, notre objectif n'est pas de rapprocher le maximum d'entités ontologiques mais d'aligner les ontologies en nous fondant sur les informations textuelles qui reflètent le domaine et l'application visés. Cela veut dire que les entités qui nous intéressent dans un premier temps sont les entités qui se trouvent déjà dans le texte. Pour les autres entités, il suffit d'appliquer un outil d'alignement ou même de comparer des chaînes de caractères pour les récupérer.

Le tableau 6.1 présente 2379 correspondances avec une similarité stricte sur les chaînes de caractères lemmatisées avec un seuil de 99% (des chaînes de caractère identiques). Même si on diminue le seuil, TaxoMap ne va pas plus loin dans la comparaison lexicale entre entités sémantiques car la méthode se fonde sur le fait que les concepts généraux sont inclus dans des concepts spécifiques. La comparaison entre TOM et TaxoMap est réalisée dans le domaine biologique où les étiquettes de concepts peuvent être différentes ce qui explique les résultats de TaxoMap (163 concepts trouvés par TOM et non trouvés par TaxoMap) mais cela est valorisé par TOM car nous nous intéressons au niveau lexical mais surtout à exploiter d'autres informations qui se trouvent dans les textes.

	TOM	¬TOM	Total
TaxoMap	22	2379	2401
¬TaxoMap	163	X	163
	185	2379	

Tableau 6.1 – Nombre de correspondances obtenues par TOM et TaxoMap

Rappelons que la relation d'équivalence dans TOM est définie par le fait qu'une entité semble substituable à une autre soit parce que ces deux entités sont synonymes ou s'il y a un lien hiérarchique sur la base de leurs distributions. Parmi les 163 relations trouvées par TOM et omises par TaxoMap, il existe 130 relations d'équivalence. Par exemple, les correspondances $< idE_0$, China, city, 0.85, *equiv* $>$, $< idE_1$, River, habitat, 0.95, *equiv* $>$, $< idE_2$, sludge, habitat, 0.70, *equiv* $>$ peuvent être modélisées comme des relations hiérarchiques entre les concepts (voir figure 6.1) dans une ontologie.

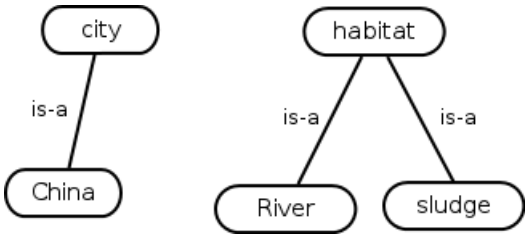


FIGURE 6.1 – Modélisation des exemples de relations d'équivalence trouvées par TOM

Sans surprise, nous avons obtenu 22 liens d'équivalence communs entre TOM et TaxoMap. Pour TOM, ces liens sont trouvés à partir du texte et vont servir à la construction d'ontologies et pour TaxoMap, ces relations sont retrouvées par l'application d'une mesure de similarité de chaînes de caractères. Les entités mises en correspondance dans ces liens sont des entités présentes dans les ontologies mais aussi importantes dans notre domaine au vu des textes.

Imaginons que les entités mises en correspondance dans les 22 correspondances communes entre TOM et TaxoMap sont vraiment fusionnables et que le processus de construction d'ontologies est incrémental, on commence par un noyau de correspondances orienté application et on l'enrichit avec des correspondances où les entités possèdent exactement les mêmes chaînes de caractères.

En focalisant sur les entités ancrées, le tableau 6.2 montre le nombre d'entités ancrées qui ont été mises en correspondance par TOM et TaxoMap. 67% d'entités de OntoBiotope et 100% d'entités de EnvO, ont été mise en relation par notre système TOM. En contre partie, on trouve 34% d'entités de OntoBiotope et 76% d'entités de EnvO obtenues par TaxoMap.

	Entités ancrées	TOM	TaxoMap
OntoBiotope	91	61	31
EnvO	34	34	26

Tableau 6.2 – Nombre d'entités ancrées mises en correspondance avec TOM et TaxoMap

Il faut noter aussi que dans le tableau 6.1, parmi les 163 relations fournies par TOM, 33 correspondances de type association sont identifiées grâce à la présence de leurs entités dans le texte. Elles sont pertinentes pour la construction d'une ressource sémantique quand on veut lier deux entités qui tendent à avoir des rôles. A titre d'exemple, les correspondances suivantes : $< idA_0$, offspring, organ, 0.5, $assoc >$, $< idA_1$, compost, mushroom, 0.5, $assoc >$ et $< idA_2$, Petroleum, city, 0.5, $assoc >$, peuvent être modélisées comme des rôles entre les entités (voir figure 6.2) dans une ontologie.

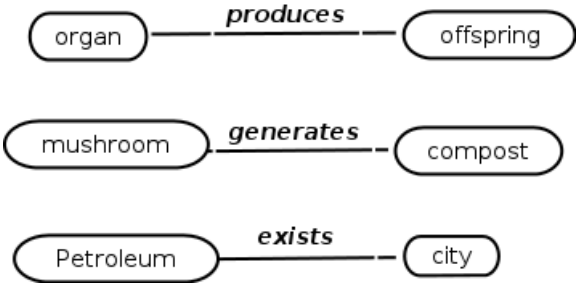


FIGURE 6.2 – Modélisation des exemples de relations d'association trouvées par TOM

Discussion Nous avons comparé deux systèmes qui sont proches en mode de fonctionnement mais chacun avec son objectif d'alignement différent. Nous avons montré l'intérêt d'utiliser le texte dans un processus d'alignement. Dans les comparaisons ci-dessus, on peut conclure que TOM et TaxoMap sont deux systèmes complémentaires ce qui laisse penser que TOM devrait pouvoir enrichir un système d'alignement existant. On observe en effet une augmentation du nombre de relations trouvées par TOM entre les ontologies dans le tableau 6.1. Ces relations ont été ignorées par TaxoMap. Cela montre aussi que le texte est utile pour améliorer la sortie d'alignement surtout quand cet alignement est guidé par une application donnée.

6.4 Évaluation par rapport à un jugement d'expert

Le but de cette expérience est de demander à un ingénieur de la connaissance de mesurer la qualité de la sortie d'alignement. Nous décrivons dans cette section le protocole d'évaluation et les résultats obtenus en comparant notre approche à l'alignement de référence.

Protocole d'évaluation Le protocole de validation propose à l'ingénieur de la connaissance l'ensemble des entités mises en relation et le texte annoté (voir figure 6.3). L'ingénieur de la connaissance doit donner son jugement sur les correspondances. Nous demandons à l'ingénieur de la connaissance de juger chacune des correspondances proposées. Il faut noter que nous ne lui fournissons pas les scores des relations pour qu'il ne soit pas influencé par ces valeurs. Quatre mentions lui sont proposées pour catégoriser les relations entre entités (par défaut, la relation est indéfinie entre deux entités) :

- « relation *is-a* », quand l'ingénieur de la connaissance est d'accord que les deux entités à rapprocher entretiennent une relation de subsomption, sans tenir compte du sens de cette relation ;
- « relation d'équivalence », quand l'ingénieur de la connaissance est d'accord que les deux entités à mettre en correspondance sont équivalentes ;
- « relation associative », quand l'ingénieur de la connaissance est d'accord que les deux entités entretiennent une relation associative, c'est-à-dire une relation du domaine (un rôle) ;
- « rien », quand l'ingénieur de la connaissance ne propose aucune relation entre ces entités.

FIGURE 6.3 – Protocole de l'évaluation humaine des correspondances

Dans le but d'évaluer les valeurs des scores calculés dans l'alignement, nous avons choisi de construire un jeu de test à partir des correspondances de notre sortie d'alignement. Nous avons réalisé ce jeu de test sous la forme de trois listes de correspondances de type équivalence et association selon un seuil que nous fixons au préalable (dans ce cas, le seuil est égal à 0.5) :

- une liste « bonnes » correspondances : un ensemble de correspondances qui possèdent un score supérieur au seuil fixé ;
- une liste « moyennes » correspondances : un ensemble de correspondances qui possèdent un score juste au-dessous du seuil fixé (ex. si on a un seuil de 0.5, nous prenons les correspondances ayant un score entre [0.3 .. 0.49]) ;
- une liste « faibles » correspondances : un ensemble de correspondances qui possèdent un score très faible par rapport au seuil fixé.

Les figures 6.4 et 6.5 montrent l'accueil et l'interface que nous fournissons à l'ingénieur de la connaissance pour typer les relations entre les entités mises en correspondance. L'ingénieur de la connaissance a la possibilité de choisir les correspondances à valider des deux entités à rapprocher. Quand l'ingénieur de la connaissance a un doute pour nommer les relations entre une paire d'entités, il peut fouiller le texte annoté pour explorer les occurrences des entités.

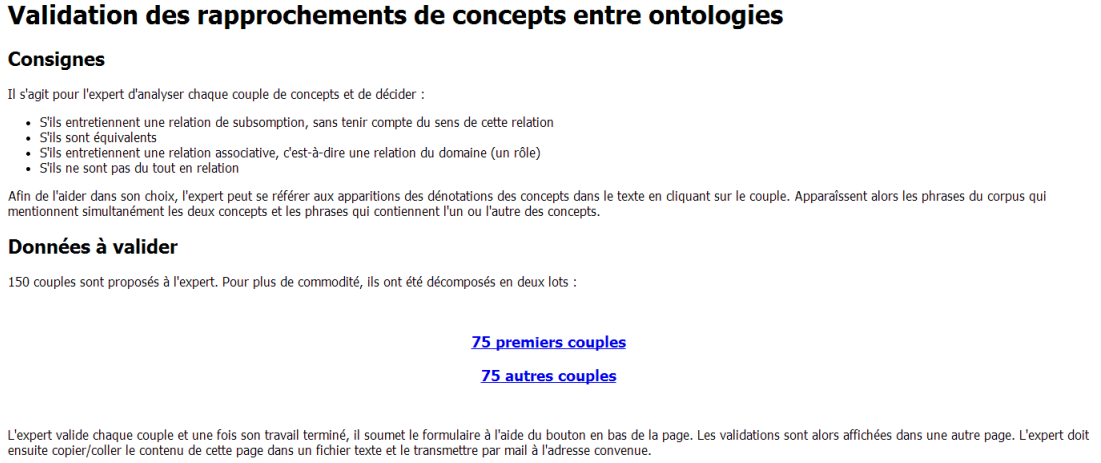


FIGURE 6.4 – Accueil de l'interface de l'évaluation humaine de l'alignement

L'interface de validation humaine comporte 4 parties :

- la partie 1 contient les entités mises en relation ;
- la partie 2 contient les possibilités que l'ingénieur de la connaissance peut associer à un couple d'entités ;
- la partie 3 indique le nombre d'occurrences isolées ou groupées des entités mises en correspondance dans le texte.

A partir des jugements de l'ingénieur de la connaissance qui représentent un alignement de référence, nous évaluons le typage de correspondances ainsi que le comportement des scores calculés de la manière suivante :

1. Pour chaque type de relations, nous comparons les correspondances de notre outil d'alignement par rapport aux correspondances de l'ingénieur de la connaissance.

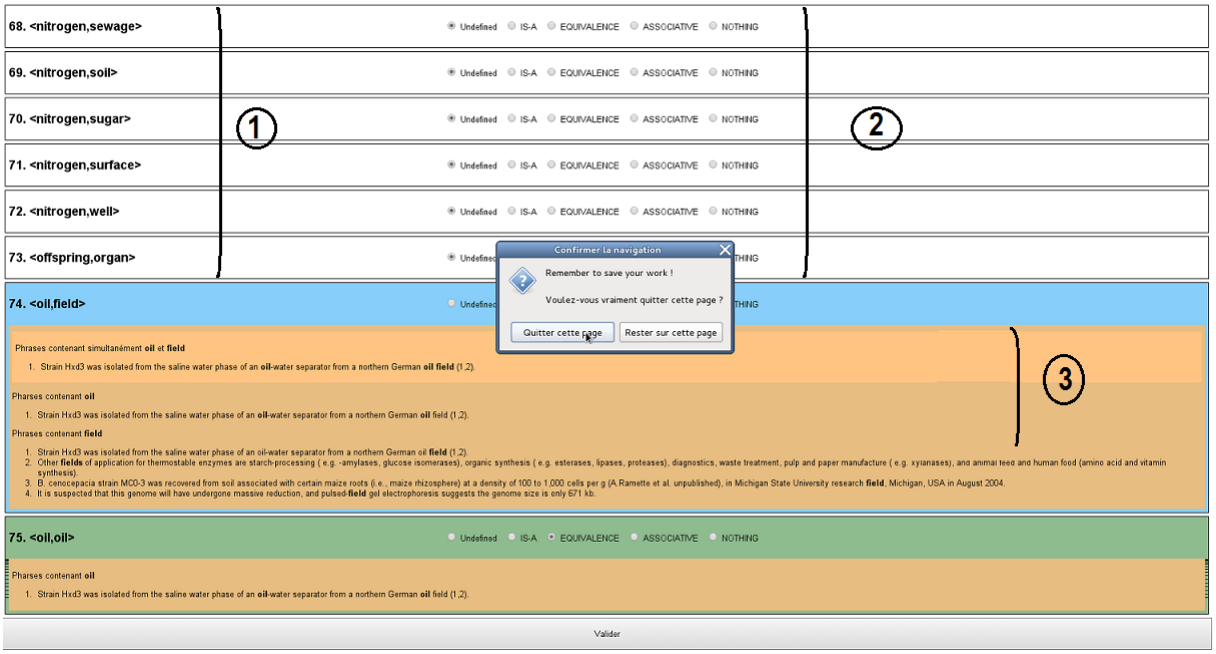


FIGURE 6.5 – Interface de l'évaluation humaine de l'alignement

Nous utilisons pour cela les mesures de précision et rappel. Ces mesures permettent d'évaluer la pertinence des correspondances ainsi que le typage de relations.

2. Pour tout type de relations, nous comparons les correspondances que l'ingénieur de la connaissance a typées par rapport à chaque liste de correspondances (bonne, moyenne, faible). Cette évaluation permet de mesurer la pertinence des valeurs de scores calculés.

Résultats de l'évaluation Nous avons mené une expérience utilisant le cas d'usage du domaine biologique en proposant à l'ingénieur de la connaissance 150 couples d'entités mises en correspondance. Il s'agit pour l'ingénieur de la connaissance d'analyser chaque couple d'entités et de décider le type de relations.

Afin de l'aider dans son choix (voir figure 6.5), l'ingénieur de la connaissance [3] peut se référer aux apparitions des dénotations des concepts dans le texte en cliquant sur le couple. Apparaissent alors les phrases du texte qui mentionnent simultanément les deux concepts et les phrases qui contiennent l'un ou l'autre des concepts.

Analyse de calcul de correspondances et de typage de relations Cette expérimentation consiste à vérifier si l'ingénieur de la connaissance a réussi à typer les relations que notre approche a fourni. Le tableau 6.3 montre que l'ingénieur de la connaissance a réussi à typer 23 relations d'équivalence sur 32, 48 relations d'association sur 118 et 79 relations non-typées. On constate que le rappel dans les deux types de relations est égal à 1. Cela dû au fait que les correspondances trouvées par l'ingénieur de la connaissance sont

3. Merci à Liliana IBANESCU, Maître de Conférences de AgroParisTech & I.N.R.A. Méta@risk

exactement typées de la même manière par notre méthode automatique d'alignement. Au total, nous avons obtenu une f-mesure de 71%.

	Précision	Rappel	F-mesure
# rel. equiv	0.71 (23/32)	1 (23/23)	0.83
# rel. assoc	0.41 (48/118)	1 (48/48)	0.59
# total	0.56	1	0.71

Tableau 6.3 – Pertinence de l'alignement fourni

Analyse du comportement des valeurs de scores calculés L'expérimentation consiste à vérifier si le comportement de scores calculés correspond bien à nos spécifications. Cette expérimentation montre la qualité de ces scores calculés par rapport à trois listes de correspondances que nous avons préalablement présentées.

Les résultats figurent dans le tableau 6.4. On constate sans surprise que 56% de « bonnes » correspondances ayant un score supérieur à 0.5 figurent en tête et de classement, 38% de « moyennes » correspondances en deuxième position et 35% de « faibles » correspondances sont en troisième position. Les valeurs de pourcentage obtenues sont essentiellement dues à la manière de calculer les scores ainsi que le choix de la mesure mais aussi à la façon de calculer le seuil.

	TOM	Expert	%
# « bonnes » correspondances (>= 0.5)	81	45	56
# « moyennes » correspondances ([0.30..0.49])	52	20	38
# « faibles » correspondances (< 0.30)	17	6	35

Tableau 6.4 – Pertinence de calcul de scores

6.5 Évaluation par rapport à la campagne d'évaluation OAEI

La troisième expérimentation dans ce chapitre a pour objectif de comparer notre outil d'alignement TOM aux systèmes d'alignement de l'état de l'art (participant à la campagne OAEI) et d'étudier l'impact du texte. Cette expérimentation est réalisée sur la base de test de OAEI. Cette base de test comporte 49 ontologies dont la première ontologie (ontologie #101) est considérée comme une ontologie de référence. Elle est composée de 37 classes, 115 relations *is-a*, 72 rôles et 57 instances. Les autres ontologies sont des variantes produites de façon systématique à partir de l'ontologie de référence (ex. par une suppression de quelques classes, de la hiérarchie). Les variantes de l'ontologie de référence permettent de tester le comportement des méthodes d'alignement quand on supprime des informations. Cela influe non seulement sur le résultat d'alignement mais aussi sur les mesures d'évaluation. Le contenu de ces ontologies est décrit dans l'annexe B.

Cette section est organisée comme suit : la section 6.5.1 décrit les ontologies ainsi que les tests effectués dans la campagne d'évaluation OAEI. La base de tests est organisée en

5 catégories selon les changements effectués sur l'ontologie de référence. La section 6.5.2 montre l'influence du choix du texte sur les résultats d'alignement. La dernière section présente une comparaison de notre outil d'alignement TOM par rapport aux systèmes d'alignement disponibles et testés dans OAEI.

6.5.1 Base de tests

Les ontologies de la base de test OAEI décrivent des références bibliographiques. Les tests réalisés sur ces ontologies permettent de comparer l'ontologie de référence aux variantes. Le but de ces tests est d'étudier l'impact de ces changements sur le résultat d'alignement.

La base de tests de OAEI comprend 50 tests et elle est divisée en plusieurs catégories selon les modifications établies sur l'ontologie de référence. Étant donnée que notre méthode d'alignement s'inscrit dans la catégorie des méthodes qui utilisent les techniques terminologiques, nous décrivons dans ce qui suit les catégories de la base de test où les ontologies possèdent le niveau lexical. Ces catégories sont au nombre de 4 :

- **cat #1**, tests 101-104 : cette catégorie comporte 3 ontologies contenant les mêmes noms de classes et les mêmes propriétés que l'ontologie de référence : (1) test 101-101 : l'ontologie de référence est comparée à elle même, (2) test 101-103 : l'ontologie de référence est comparée à une ontologie au format OWL-Lite dans laquelle il n'existe pas les propriétés de type owl:unionOf, owl:oneOf et owl:TransitiveProperty (on parle d'une ontologie de généralisation), et (3) test 101-104 : l'ontologie de référence est comparée à une ontologie au format OWL-Lite dans laquelle les contraintes sont supprimées (on parle d'une ontologie de restriction).
- **cat #2**, tests 201-210 : cette catégorie comporte 10 ontologies bien structurées :
  - le test 101-201 compare l'ontologie de référence avec une ontologie qui ne possède pas de noms de classes (les étiquettes de classes ainsi que les propriétés sont remplacés par des noms aléatoires) ;
  - le test 101-202 compare l'ontologie de référence avec une ontologie qui ne possède ni noms de classes ni commentaires ;
  - le test 101-203 compare l'ontologie de référence avec une ontologie qui ne possède pas de commentaires ;
  - les tests 101-204 et 101-208 comparent chacun l'ontologie de référence avec une ontologie dont les noms de classes sont obtenus en appliquant des règles de nommage (ex. majuscule, tiret, trait de soulignement) ;
  - les tests 101-205 et 101-209 comparent chacun l'ontologie de référence avec une ontologie dont les labels de classes sont remplacés par des synonymes ;
  - les tests 101-206, 101-207 et 101-210 comparent chacun l'ontologie de référence avec une ontologie dont les noms de classes sont traduits d'autres langues que l'anglais.
- **cat #3**, tests 221-247 : cette catégorie contient 18 ontologies lexicalisées ; les tests sont des variantes de l'ontologie de référence obtenues en enlevant la hiérarchie, en mettant la hiérarchie à plat (suppression de quelques liens hiérarchiques), en l'enrichissant (ajout de classes intermédiaires et de relations hiérarchiques ou supression de classes), ou encore en supprimant les instances et les propriétés.

– **cat #4**, tests 301-304 : cette catégorie contient 4 ontologies qui possèdent les deux niveaux lexical et structurel. Les tests comportent une comparaison de l'ontologie de référence avec les ontologies réelles de BibTeX/MIT, de BibTeX/UMBC, de Karlsruhe et de l'INRIA.

6.5.2 Analyse de choix du texte dans la méthode TOM

Cette évaluation expérimentale a été menée pour montrer l'importance du texte dans le cadre de la méthode proposée. L'objectif de TOM est de réaliser un alignement entre des ontologies, validé par le texte utilisé comme support d'information. L'alignement produit par TOM est donc dépendant du choix du texte de départ.

Nous avons construit manuellement deux textes (voir tableau 6.5) du domaine des références bibliographiques. Le choix de ce domaine a pour but de se comparer aux outils d'alignement existants de OAEI, en évaluant sur la base de test présentée plus haut.

Nous avons construit un premier texte à partir des pages Wikipédia décrivant une conférence, un article, un atelier, etc. Ce texte comporte 164 phrases. Un deuxième texte a été construit à partir des appels aux conférences et aux ateliers et il comporte 4944 phrases.

	#mots	#mots différents
Extrait de Wikipedia (texte 1)	35606	2083
Extraits d'appels aux conférences et aux ateliers (texte 2)	54122	3022

Tableau 6.5 – Nombre de mots dans les deux textes : extrait de Wikipedia et extraits d'appels aux conférences et aux ateliers

Nous avons aligné les ontologies de la base de test en prenant compte les deux textes tour à tour. Nous avons cherché à comparer nos résultats avec ceux obtenus dans OAEI pour chaque catégorie de la base de test. Le tableau 6.6 présente les résultats de cette comparaison. Notre méthode conduit à une augmentation significative de la précision, de rappel et de la f-mesure quand on passe d'un texte (texte 1) à l'autre (texte 2).

Les résultats de texte 2 comporte 48% de correspondances pour la catégorie cat#1 de la base de test alors qu'avec le texte 1, notre méthode ne donne que 13% de précision pour la même catégorie. Cela s'explique par le fait que plus on a des entités d'ontologies ancrées dans le texte plus les mesures de précision et rappel augmentent. Pour plus de détails sur ce surcroît de résultats, nous avons analysé les deux textes. Nous avons constaté que le texte de Wikipedia décrit d'une manière assez générale le domaine de références bibliographiques alors que le deuxième texte d'appels aux conférences et aux ateliers, comporte des mots techniques ou spécifiques à ce domaine, ce qui a permis d'obtenir plus d'entités d'ontologies ancrées dans ce texte.

Nous pouvons conclure que le texte est important dans notre méthode TOM car il nous permet de rapprocher lexicalement des entités ignorées par les outils existants. L'expérience ci-dessus montre qu'il faut bien choisir le texte sur lequel s'appuyer. Le choix de ce texte est guidé par l'application visée. Pour nous assurer de l'adéquation du texte, nous pouvons

	Texte 1			Texte 2		
	P	R	F-measure	P	R	F-measure
cat#1	13%	7%	9.1%	48%	79%	60%
cat#2	6%	5%	5.4%	48%	16%	24%
cat#3	10%	7%	8.2%	35%	45%	39%
cat#4	5%	10%	6.6%	32%	34%	33%

Tableau 6.6 – Expérience sur le choix du texte de référence

adapter des mesures de couverture existantes des ontologies au texte telles que celles de [Ninova *et al.*, 2005] ou celles de [Brewster *et al.*, 2004].

6.5.3 Comparaison aux outils d'alignement

Dans la campagne d'évaluation OAEI[4], 18 outils d'alignement ont été testés. Le tableau 6.7 récapitule les résultats d'alignement obtenus (précision, rappel et f-mesure) par les différents outils et TOM en utilisant les ontologies de la base de test. Le système edna est le système de base qui utilise l'algorithme de distance d'édition (distance de Levenshtein) pour rapprocher les entités des ontologies.

Presque tous les systèmes donnent des valeurs de précision moyenne élevées par rapport aux valeurs de rappel moyen. Les meilleurs résultats de valeurs globales de f-mesure sont obtenues par les systèmes AROMA (80%), YAM++ (86%) et MapSSS (91%) puisque ces systèmes prennent en compte les niveaux lexical et structurel des ontologies à aligner. Nous comparons le résultat obtenu de notre méthode TOM avec le deuxième texte (présenté dans la section 6.5.2) par rapport aux résultats d'évaluation des autres outils d'alignement. Nous proposons 3 volets de comparaison :

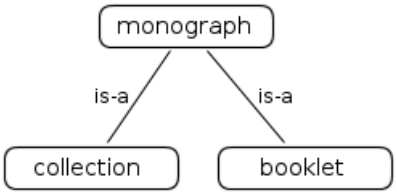
volet 1 : comparaison de TOM par rapport au système de base edna : les valeurs de f-mesure de chaque catégorie de test des deux systèmes sont proches. Le système de base edna présente une augmentation de la valeur de f-mesure dans la cat#3 car cette catégorie présente des variantes d'ontologies où on a ajouté ou supprimé des classes, ce qui peut agir sur le résultat global de cette catégorie dans notre méthode TOM. Le résultat de TOM dans cette catégorie (cat#3) s'explique par l'absence des concepts d'ontologies et leurs termes ajoutées dans le texte.

volet 2 : comparaison de TOM par rapport à TaxoMap : dans les catégories cat#1 et cat#4, les valeurs globales de f-mesure de TOM et TaxoMap sont proches. Cependant, dans la cat#2 où le niveau structurel est important, TaxoMap présente une f-mesure de 6% alors que TOM présente 24%. Cela s'explique par le fait que, dans cette catégorie, il existe des ontologies possédant des noms de classes (les labels de classes) ou des synonymes des labels de ces classes. De ce fait le résultat de TOM qui permet d'exploiter le niveau lexical de ces ontologies. Quant la catégorie cat#3, TaxoMap présente une valeur élevée de f-mesure (72%) par rapport à TOM. Cela est dû au fait que, dans cette catégorie, un ajout et une suppression de classes sont établis. Ce

4. Nous avons récupéré les résultats d'outils d'alignement évalués dans la campagne d'évaluation OAEI de l'année 2012 et le résultat de l'outil TaxoMap dans OAEI de l'année 2007

qui influe sur le résultat global de cette catégorie dans TOM car les noms des classes ajoutées ne sont pas présents dans le texte et aussi les noms des classes supprimées ne sont pas annotés dans le texte.

volet 3 : comparaison de TOM par rapport au système MapSSS qui a les meilleurs valeurs de précision, rappel et f-mesure. Ce n'est pas une surprise qu'un système a de meilleurs résultats que notre méthode TOM, car on se situe dans les méthodes complémentaires à ces outils. MapSSS est un outil qui prend en compte les niveaux structurel et lexical des ontologies. Ce qui permet d'avoir des meilleurs résultats. A titre d'exemple, les relations d'équivalence obtenues dans MapSSS se fondent essentiellement sur des mesures de similarité entre chaînes de caractères alors que dans TOM, on trouve des relations d'équivalence entre des concepts d'ontologies qui sont intéressantes à présenter à l'ingénieur de la connaissance dans le but de construire une ressource sémantique mais qui ne sont pas trouvées par une simple similarité. Par exemple « monograph » est équivalent à « collection » et « monograph » est équivalent à « booklet ». Ces deux relations d'équivalence sont intéressantes pour construire une ressource sémantique et elles peuvent être présentées comme des relations hiérarchiques.



En résumé, les résultats de notre méthode TOM où les valeurs sont moins bonnes se justifie par les caractéristiques de chaque catégorie de tests :

1. dans la cat#1, les classes des ontologies possèdent des étiquettes qui sont associées à des mots dans le texte ;
2. dans la cat#2, notre méthode TOM ne prend pas en compte le niveau structurel des ontologies lors du processus d'alignement ; la valeur de f-mesure 24% s'explique par l'absence des labels ou des synonymes dans les classes des ontologies ;
3. dans la cat#3, malgré le fait que notre méthode TOM prend en compte le niveau lexical dans les ontologies, nous avons obtenu 39% de f-mesure ; cela s'explique par la suppression et l'ajout de classes dans les variantes d'ontologies dans cette catégorie ;
4. dans la cat#4, on tient compte des niveaux lexical et structurel des ontologies ; notre méthode ne prend en compte que le niveau lexical. L'absence des classes d'ontologies associées aux mots du texte, a influé sur le résultat global dans cette catégorie.

6.6 Conclusion

L'évaluation des cartographies produites est beaucoup plus difficile parce qu'il faudrait pouvoir idéalement mesurer l'aide que les cartographies apportent aux ingénieurs de la connaissance et pour les différentes tâches de gestion d'ontologies qu'ils ont à mener. Ce point n'a pas été abordé.

	edna				AROMA				ASE				AUTOMSv2		
	P	R	FM		P	R	FM		P	R	FM		P	R	FM
cat #1	0,64	1	0,78	cat #1	1	1	1	cat #1	0,58	1	0,73	cat #1	0,96	0,92	0,94
cat #2	0,25	0,4	0,31	cat #2	1	0,81	0,87	cat #2	0,45	0,68	0,54	cat #2	0,97	0,68	0,73
cat #3	0,75	1	0,83	cat #3	1	1	1	cat #3	0,69	1	0,8	cat #3	0,97	1	0,99
cat #4	0,18	0,29	0,19	cat #4	0,93	0,12	0,36	cat #4	0,33	0,37	0,12	cat #4	0,96	0,29	0,3
Moyenne	0,45	0,67	0,52	Moyenne	0,98	0,73	0,80	Moyenne	0,51	0,76	0,47	Moyenne	0,96	0,72	0,74
	GOMMA				Hertuda				Hotmatch				LogMap		
	P	R	FM		P	R	FM		P	R	FM		P	R	FM
cat #1	0,84	1	0,91	cat #1	0,89	1	0,94	cat #1	0,91	0,76	0,83	cat #1	0,85	0,96	0,9
cat #2	0,82	0,76	0,77	cat #2	0,94	0,68	0,72	cat #2	0,98	0,63	0,71	cat #2	0,59	0,37	0,44
cat #3	0,74	1	0,85	cat #3	0,87	1	0,93	cat #3	0,94	0,87	0,9	cat #3	0,77	0,95	0,84
cat #4	0,48	1,36	0,62	cat #4	0,89	0,37	0,34	cat #4	1	0,23	0,5	cat #4	0,92	0,39	0,21
Moyenne	0,72	0,78	0,78	Moyenne	0,89	0,76	0,65	Moyenne	0,95	0,62	0,73	Moyenne	0,78	0,66	0,59
	LogMapLt				MaasMtch				MapSSS				MEDLEY		
	P	R	FM		P	R	FM		P	R	FM		P	R	FM
cat #1	0,83	0,99	0,91	cat #1	1	1	1	cat #1	1	1	1	cat #1	0,72	1	0,84
cat #2	0,59	0,39	0,46	cat #2	0,52	0,52	0,52	cat #2	0,99	0,9	0,94	cat #2	0,43	0,4	0,41
cat #3	0,76	1	0,86	cat #3	0,9	1	0,95	cat #3	0,99	0,99	0,99	cat #3	0,76	1	0,85
cat #4	0,87	0,27	0,3	cat #4	0,15	0,15	0,12	cat #4	0,98	0,36	0,72	cat #4	0,62	0,29	0,2
Moyenne	0,76	0,66	0,63	Moyenne	0,64	0,66	0,64	Moyenne	0,99	0,81	0,91	Moyenne	0,63	0,67	0,57
	Optima				ServOMap				ServOMapLt				WeSeE		
	P	R	FM		P	R	FM		P	R	FM		P	R	FM
cat #1	1	1	1	cat #1	1	0,98	0,99	cat #1	1	0,28	0,44	cat #1	0,97	0,94	0,95
cat #2	0,82	0,63	0,67	cat #2	0,99	0,39	0,5	cat #2	1	0,12	0,2	cat #2	0,99	0,65	0,72
cat #3	0,92	1	0,96	cat #3	0,91	0,91	0,91	cat #3	1	0,55	0,67	cat #3	0,99	0,98	0,98
cat #4	1	0,53	0,09	cat #4	0,59	0,48	0,06	cat #4	1	0,16	0,03	cat #4	1	0,31	0,31
Moyenne	0,93	0,79	0,67	Moyenne	0,87	0,69	0,61	Moyenne	1	0,27	0,33	Moyenne	0,98	0,72	0,74
	Wikimatch				YAM++				TOM				TaxoMap		
	P	R	FM		P	R	FM		P	R	FM		P	R	FM
cat #1	0,84	1	0,91	cat #1	1	1	1	cat #1	0,48	0,79	0,60	cat #1	1	0,34	0,50
cat #2	0,72	0,66	0,67	cat #2	0,97	0,88	0,91	cat #2	0,48	0,16	0,24	cat #2	0,26	0,04	0,06
cat #3	0,79	1	0,88	cat #3	1	1	1	cat #3	0,35	0,45	0,39	cat #3	0,92	0,60	0,72
cat #4	0,83	0,35	0,21	cat #4	0,98	0,13	0,55	cat #4	0,32	0,34	0,33	cat #4	1	0,26	0,41
Moyenne	0,79	0,75	0,66	Moyenne	0,98	0,75	0,86	Moyenne	0,40	0,43	0,41	Moyenne	0,79	0,31	0,42

Tableau 6.7 – Comparaison des systèmes d'alignement avec notre système TOM

Conclusion et perspectives

A La croisée de l'ingénierie de connaissances et du Web sémantique, cette thèse s'intéresse à la gestion des ontologies dans un contexte ouvert où l'on cherche avant tout à réutiliser des ressources existantes mais où l'hétérogénéité sémantique de ces ressources constitue un défi à l'interopérabilité et un frein à leur réutilisation.

Nous avons proposé une méthode et des outils pour cartographier un domaine sémantique et ainsi donner à un ingénieur de la connaissance une vue organisée des ressources sémantiques disponibles pour le domaine qui l'intéresse. L'originalité de notre approche vient de la place centrale qui est allouée aux textes. Comme l'ingénieur de la connaissance peut difficilement spécifier son domaine d'intérêt et la perspective qui est la sienne par une simple liste de mot-clés, nous avons jugé préférable de lui proposer de prendre un texte comme requête et de prendre appui sur ce texte pour sélectionner et aligner les ressources les plus pertinentes. Autour de ce texte utilisé comme pivot, nous proposons donc de construire automatiquement une cartographie de domaine. Sans modifier les ressources sémantiques sur lesquelles elle s'appuie, la cartographie se présente comme un ensemble de liens de correspondance entres les entités des ressources présélectionnées, un ensemble de liens d'annotation entre ces mêmes entités et leurs mentions dans le texte choisi comme référence et des outils d'exploration pour extraire et analyser des sous-ensembles de liens.

Nous avons d'abord passé en revue les travaux de l'état de l'art en mettant l'accent sur l'hétérogénéité des ressources sémantiques disponibles, la gestion de ces ressources sémantiques – notamment les méthodes d'alignement qui sont présentées comme une solution au problème de l'interopérabilité sémantique – et la notion de cartographie conceptuelle. En pratique, nous avons travaillé uniquement sur des ontologies lexicalisées mais nous pensons qu'une partie de nos propositions peut s'appliquer à des ressources moins fortement structurées.

Dans une seconde partie, nous avons présenté notre méthode de construction de cartographies de domaine et détaillant à la fois notre approche d'alignement guidé par le texte et la construction de la cartographie en tant que telle à partir des liens de correspondance établis dans la phase d'alignement.

La troisième et dernière partie présente les expériences que nous avons menées pour mettre au point et tester notre approche et celles que nous avons montées pour évaluer plus spécifiquement les résultats de l'alignement des ressources. Elles portent selon les cas sur l'un de nos cas d'usage dans les domaines biologique, géographique et alimentaire ou sur des données de campagnes d'évaluation.

Contributions

Outre la méthodologie d'ensemble de construction de cartographies de domaine à partir des ontologies hétérogènes disponibles et d'un texte pivot, cette thèse se décline en quatre points principaux.

Proposition d'une méthode d'alignement guidée par le texte L'avantage de la

méthode d'alignement proposée est qu'elle s'appuie sur un texte qui peut donner une description riche du domaine et de l'application visée. Cette méthode prend en entrée deux ontologies lexicalisées et un texte de référence. Elle exploite l'annotation du texte source au regard des ressources utilisées et établit des correspondances entre les entités des ressources à aligner en fonction de la cooccurrence et de la similarité distributionnelle de leurs labels dans le texte source. La sortie est un ensemble de correspondances filtrées qui établit un alignement de type n:m entre les ontologies d'entrée. Les correspondances calculées peuvent proposer des relations d'équivalence ou d'association entre les entités des ontologies alignées.

Proposition d'une méthode de construction de cartographies de domaine Nous proposons de compléter l'ensemble des liens d'annotation et de correspondance établis entre le texte et les ontologies d'entrée par des outils d'exploitation qui permettent d'identifier des configurations de liens particulières, éventuellement de réviser l'alignement obtenu et d'afficher des zones d'intérêt. Les configurations anormales mettent en évidence des écarts de conceptualisation entre les ressources alignées. Les configurations remarquables font ressortir des zones centrales qui peuvent être intéressantes à analyser pour l'ingénieur de la connaissance.

Implémentation des méthodes Nous avons implémenté 1) la méthode d'alignement *TOM*, 2) les méthodes de détection de configurations anormales et remarquables et 3) la méthode de résolution des anomalies. A partir de ces méthodes, nous proposons deux modules distincts pour l'alignement d'une part et pour la révision de l'alignement d'autre part.

Expériences et évaluation Nous avons conduit des expériences pour chacune des phases de notre méthodologie de construction de la cartographie de domaine, en utilisant des cas d'usage de trois domaines de spécialité différents, géographique, alimentaire et biologique. Nous avons également évalué la méthode d'alignement proposée, *TOM* 1) en la comparant à la méthode de l'état de l'art TaxoMap sur le cas d'usage biologique, 2) en préparant évaluation humaine (seul le protocole d'évaluation a été présenté à ce stade), et 3) en comparant notre méthode aux méthodes de l'état de l'art sur la base de test de la campagne d'évaluation de l'alignement d'ontologies (OAEI).

Perspectives

Ce travail ouvre différentes perspectives de travail.

Nous envisageons en premier lieu d'améliorer la méthodologie proposée. Certaines parties n'ont en effet pas fait l'objet d'une analyse très poussée et pourraient certainement être améliorées.

- L'annotation sémantique d'un texte au regard d'une ou plusieurs ontologies est faite de manière triviale, par une simple comparaison de chaînes de caractères entre les unités textuelles et les entités ontologique. A terme, il s'agira d'intégrer un outil d'annotation comme SemEx [Guissé *et al.*, 2011] ou un annotateur de l'état de l'art pour augmenter l'ancrage des ressources sémantiques dans le texte.
- Pour le calcul des cooccurrences et de la similarité distributionnelle, nous avons repris des mesures très classiques mais une étude et une évaluation plus précises

pourraient nous amener à en proposer de plus adaptées.

- Jusqu'à présent les ontologies et le texte source nous ont été fournis par des partenaires mais il faudrait prendre en compte l'étape préalable de sélection des ressources sémantiques. Nous pensons que cette sélection peut elle-même être guidée par le texte fourni par l'ingénieur de la connaissance pour représenter le domaine qui l'intéresse. Si on récupère des ontologies très générales, une étape de modularisation devra sans doute être intégrée pour circonscrire les parties intéressantes des ontologies par rapport au domaine de spécialité.
- Il est intéressant aussi d'adapter notre méthodologie à d'autres types de ressources sémantiques, à savoir les thésaurus où les connaissances sont de nature terminologique et les relations de nature lexicale.

Maintenant que nous sommes capables de construire des cartographies de domaine, il faut comprendre plus précisément comment elles peuvent être exploitées. A ce stade, nous avons pu mettre en oeuvre notre méthodologie sur différents cas d'usage et en évaluer les parties techniques mais nous n'avons pas pu valider avec des utilisateurs les deux grandes hypothèses qui la sous-tendent, à savoir 1) l'intérêt de la cartographie de domaine pour la gestion des ressources sémantiques et 2) la pertinence du texte comme point d'appui de l'ensemble du processus de cartographie. Notre travail a suscité l'intérêt de certains partenaires mais notre objectif est de monter des protocoles expérimentaux où la cartographie de domaine est utilisée dans un but précis, pour construire une nouvelle ontologie en fusionnant tout ou partie des ressources alignées ou au contraire pour les découper en modules.

Chapitre 7

Annexes

A Implémentation de la méthode d'alignement TOM

La méthode d'alignement *TOM* est implémenté en langage de programmation java sous l'environnement de développement intégré libre Eclipse IDE (Integrated Development Environment) dans l'application nommée TOM (Text-based Ontology Mapping). Notre choix du langage orienté objet java repose sur différents avantages à savoir sa puissance et sa richesse de librairies qui facilitent l'implémentation des applications sûres et stables. Parmi les librairies qui nous intéressent, on trouve les parsers XML et OWL (librairie Jena [1]). Java nous oriente à développer des applications bien structurées et modulables. L'autre avantage de java est qu'il permet l'exécution des applications indépendamment des systèmes d'exploitation et cela en utilisant une machine virtuelle (JVM - Java Virtual Machine). Java est donc un langage qui possède une portabilité excellente.

Eclipse est un environnement de développement que nous le considérons complet. En effet, il permet aussi bien de programmer que de concevoir, modéliser et tester des applications. C'est pour cette raison qu'il est un environnement de développement intégré. Eclipse permet aussi d'inclure différents langages de programmation ainsi que différents plugins prédéfinis (ex. Protégé-Owl API [2] qui permet de manipuler et interroger des fichiers OWL). La figure 7.1 montre l'interface d'Eclipse qui contient l'application TOM.

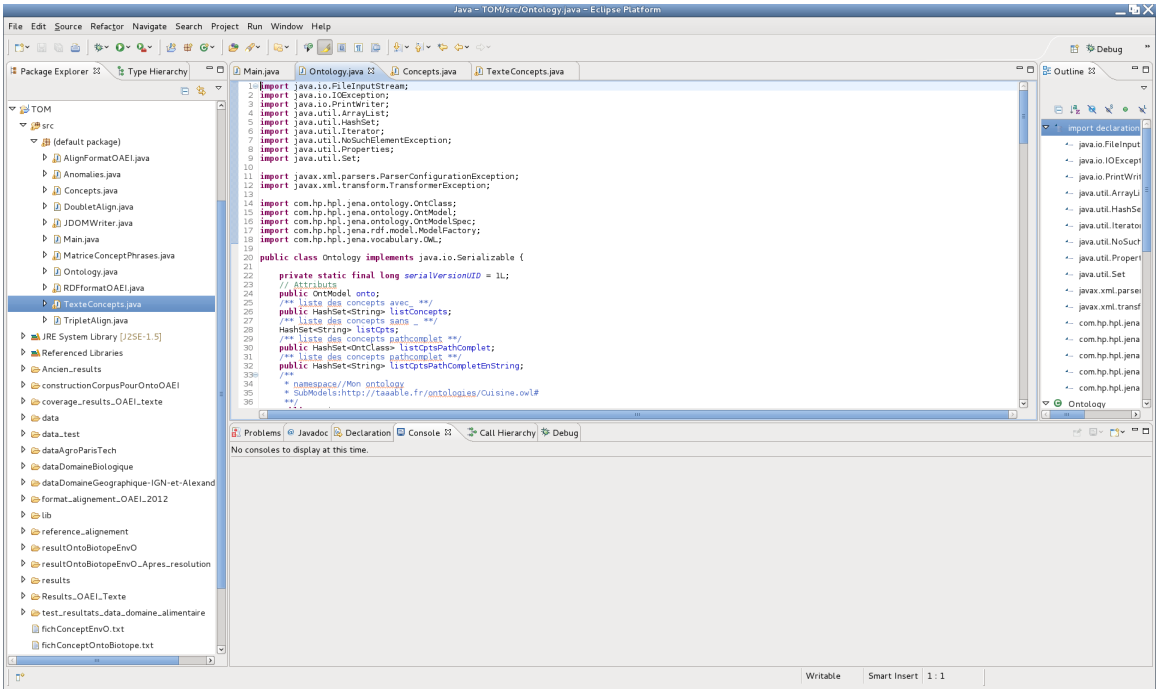


FIGURE 7.1 – Notre application *TOM* sous l'environnement Eclipse

1. http://jena.apache.org/
2. http://protege.stanford.edu/plugins/owl/api/

B Ontologies de la base de tests de la compagne OAEI

Ontologies	Description	Contenu
101	ontologie de référence	37 classes, 115 relations *is-a*, 72 rôles et 57 instances
103	ontologie généralisée	37 classes, 115 relations *is-a*, 72 rôles et 56 instances
104	ontologie de restriction	37 classes, 114 relations *is-a*, 72 rôles et 56 instances
201	ontologie avec aucun nom de classes	37 classes, 115 relations *is-a*, 72 rôles et 57 instances
202	ontologie avec aucun nom de classes et aucun commentaire	37 classes, 115 relations *is-a*, 72 rôles et 57 instances
203	ontologie avec aucun commentaire	37 classes, 115 relations *is-a*, 72 rôles et 57 instances
204	ontologie avec des noms de classes en appliquant des règles de nommage	50 classes, 115 relations *is-a*, 72 rôles et 57 instances
205	ontologie avec des labels de classes remplacés par des synonymes	50 classes, 115 relations *is-a*, 72 rôles et 57 instances
206	ontologie avec des noms de classes traduits	50 classes, 115 relations *is-a*, 71 rôles et 57 instances
207	ontologie avec des noms de classes et des commentaires traduits	50 classes, 115 relations *is-a*, 71 rôles et 57 instances
208	ontologie avec des noms de classes en appliquant des règles de nommage et aucun commentaire	50 classes, 115 relations *is-a*, 72 rôles et 57 instances
209	ontologie avec des labels de classes remplacés par des synonymes et aucun commentaire	50 classes, 115 relations *is-a*, 71 rôles et 57 instances
210	ontologie avec des noms de classes traduits et aucun commentaire	50 classes, 115 relations *is-a*, 71 rôles et 57 instances
221	ontologie avec aucune hiérarchie	37 classes, 88 relations *is-a*, 72 rôles et 57 instances
222	ontologie avec une hiérarchie aplatie	33 classes, 140 relations *is-a*, 72 rôles et 57 instances
223	ontologie avec une hiérarchie étendue	70 classes, 149 relations *is-a*, 73 rôles et 57 instances
224	ontologie avec aucune instance	38 classes, 115 relations *is-a*, 72 rôles et 0 instances
225	ontologie avec aucune restrictions sur les propriétés	37 classes, 27 relations *is-a*, 72 rôles et 56 instances
228	ontologie avec aucune propriété	36 classes, 27 relations *is-a*, 0 rôles et 56 instances
230	ontologie avec des classes aplaties	29 classes, 121 relations *is-a*, 60 rôles et 47 instances
231	ontologie avec des classes étendues	37 classes, 115 relations *is-a*, 72 rôles et 57 instances
232	ontologie avec aucune hiérarchie et aucune instance	38 classes, 0 relations *is-a*, 72 rôles et 0 instances
233	ontologie avec aucune hiérarchie et aucune propriété	36 classes, 0 relations *is-a*, 0 rôles et 56 instances
236	ontologie avec aucune instance et aucune propriété	36 classes, 27 relations *is-a*, 0 rôles et 0 instances
237	ontologie avec une hiérarchie aplatie et aucune instance	34 classes, 140 relations *is-a*, 72 rôles et 0 instances

Tableau 7.1 – Ontologies de la base de tests OAEI

Ontologies	Description	Contenu
238	ontologie avec une hiérarchie étendue et aucune instance	71 classes, 149 relations *is-a*, 73 rôles et 0 instances
239	ontologie avec une hiérarchie aplatie et aucune propriété	32 classes, 23 relations *is-a*, 0 rôles et 56 instances
240	ontologie avec une hiérarchie étendue et aucune propriété	69 classes, 60 relations *is-a*, 0 rôles et 56 instances
241	ontologie avec aucune hiérarchie, aucune instance et aucune propriété	36 classes, 0 relations *is-a*, 0 rôles et 0 instances
246	ontologie avec une hiérarchie aplatie, aucune instance et aucune propriété	32 classes, 23 relations *is-a*, 0 rôles et 0 instances
247	ontologie avec une hiérarchie étendue, aucune instance et aucune propriété	69 classes, 60 relations *is-a*, 0 rôles et 0 instances
248	ontologie avec aucun nom de classes, aucun commentaire et aucune hiérarchie	37 classes, 0 relations *is-a*, 72 rôles et 57 instances
249	ontologie avec aucun nom de classes, aucun commentaire et aucune instance	38 classes, 115 relations *is-a*, 72 rôles et 0 instances
250	ontologie avec aucun nom de classes, aucun commentaire et aucune propriété	36 classes, 27 relations *is-a*, 0 rôles et 56 instances
251	ontologie avec aucun nom de classes, aucun commentaire mais avec hiérarchie aplatie	33 classes, 140 relations *is-a*, 72 rôles et 57 instances
252	ontologie avec aucun nom de classes, aucun commentaire mais avec hiérarchie étendue	70 classes, 149 relations *is-a*, 73 rôles et 57 instances
253	ontologie avec aucun nom de classes, aucun commentaire, aucune hiérarchie, aucune instance	38 classes, 0 relations *is-a*, 72 rôles et 0 instances
254	ontologie avec aucun nom de classes, aucun commentaire, aucune hiérarchie, aucune propriété	36 classes, 0 relations *is-a*, 0 rôles et 56 instances

Tableau 7.2 – Ontologies de la base de tests OAEI (suite)

Ontologies	Description	Contenu
257	ontologie avec aucun nom de classes, aucun commentaire, aucune instance, aucune propriété	36 classes, 27 relations *is-a*, 0 rôles et 0 instances
258	ontologie avec aucun nom de classes, aucun commentaire, aucune instance mais avec hiérarchie aplatie	34 classes, 140 relations *is-a*, 72 rôles et 0 instances
259	ontologie avec aucun nom de classes, aucun commentaire, aucune instance mais avec hiérarchie étendue	71 classes, 149 relations *is-a*, 73 rôles et 0 instances
260	ontologie avec aucun nom de classes, aucun commentaire, aucune propriété mais avec hiérarchie aplatie	32 classes, 23 relations *is-a*, 0 rôles et 56 instances
261	ontologie avec aucun nom de classes, aucun commentaire, aucune propriété mais avec hiérarchie étendue	69 classes, 60 relations *is-a*, 0 rôles et 56 instances
262	ontologie avec aucun nom de classes, aucun commentaire, aucune propriété, aucune instance, aucune hiérarchie	36 classes, 0 relations *is-a*, 0 rôles et 0 instances
265	ontologie avec aucun nom de classes, aucun commentaire, aucune propriété, aucune instance mais avec hiérarchie aplatie	32 classes, 23 relations *is-a*, 0 rôles et 0 instances
266	ontologie avec aucun nom de classes, aucun commentaire, aucune propriété, aucune instance mais avec hiérarchie étendue	69 classes, 60 relations *is-a*, 0 rôles et 0 instances
301	ontologie réelle de BibTeX/MIT	15 classes, 55 relations *is-a*, 40 rôles et 0 instances
302	ontologie réelle de BibTeX/UMBC	16 classes, 22 relations *is-a*, 31 rôles et 0 instances
303	ontologie réelle de Karlsruhe	56 classes, 198 relations *is-a*, 72 rôles et 0 instances
304	ontologie réelle de l'INRIA	41 classes, 93 relations *is-a*, 51 rôles et 1 instances

Tableau 7.3 – Ontologies de la base de tests OAEI (suite)

Bibliographie

[Abadie, 2012] Abadie, N. (2012). *Formalisation, acquisition et mise en œuvre de connaissances pour l'intégration virtuelle de bases de données géographiques : les spécifications au cœur du processus d'intégration.* These, Université Paris-Est.

[Abadie et Mustière, 2010] Abadie, N. et Mustière, S. (2010). Constitution et exploitation d'une taxonomie géographique à partir des spécifications de bases de données. *Revue Internationale de Géomatique*, 20(2):145–174.

[Amardeilh et Francart, 2006] Amardeilh, F. et Francart, T. (2006). Enrichissement de bases de connaissances par l'annotation sémantique. plate-forme web sémantique couplée avec des outils linguistiques pour des activités de veille et d'édition. *Ingénierie des Systèmes d'Information*, 11(2):53–70.

[Aussenac-Gilles *et al.*, 2000] Aussenac-Gilles, N., Biébow, B. et Szulman, S. (2000). Corpus analysis for conceptual modelling. *In EKAW workshop Ontologies and texts*, pages 13–20, Toulouse, France. Université Paul Sabatier.

[Aussenac-Gilles *et al.*, 2013] Aussenac-Gilles, N., Buscaldi, D., Comparot, C. et Kamel, M. (2013). Enrichissement d'ontologies grâce à l'annotation sémantique de pages web. *In* Hermann, éditeur : *EGC - Atelier Extraction et Gestion Parallèles Distribuées des Connaissances*, volume E-24, pages 229–234. Revue des Nouvelles Technologies de l'Information (RNTI).

[Bach *et al.*, 2004] Bach, T. L., Dieng-Kuntz, R. et Gandon, F. (2004). On ontology matching problems - for building a corporate semantic web in a multi-communities organization. *In ICEIS (4)*, pages 236–243.

[Baneyx et Charlet, 2006] Baneyx, A. et Charlet, J. (2006). Evaluation, évolution et maintenance d'une ontologie en medecine : état des lieux et expérimentation. *In Revue I3 ; SI 2006 special issue on Ontological ressources.*

[Ben-Abacha et Zweigenbaum, 2010] Ben-Abacha, A. et Zweigenbaum, P. (2010). Annotation et interrogation sémantiques de textes médicaux. *In Atelier Web Sémantique Médical, Journées Francophones d'Ingénierie des Connaissances 21èmes 21es Journées Francophones d'Ingénierie des Connaissances*, pages 61–70, Nîmes France. Ecole des Mines d'Alès.

[Ben Abbès *et al.*, 2012] Ben Abbès, S., Scheuermann, A., Meilender, T. et D'Aquin, M. (2012). Characterizing modular ontologies. *In 6th International Workshop on Modular Ontologies - WoMO 2012*, pages 13–25, Graz, Autriche.

[Ben Abbès *et al.*, 2010] Ben Abbès, S., Zargayouna, H. et Nazarenko, A. (2010). Évaluation de classes sémantiques pour la construction d'ontologies. *In Acte des 21èmes 21es Journées Francophones d'Ingénierie des Connaissances 21èmes 21es Journées Francophones d'Ingénierie des Connaissances*, pages 297–308, Nîmes France. Ecole des Mines d'Alès.

[Berners-Lee et Hendler, 2001] Berners-Lee, T. et Hendler, J. (2001). Scientific publishing on the semantic web. *Nature*, 410:1023–1024.

[Borgo et Masolo, 2009] Borgo, S. et Masolo, C. (2009). Foundational choices in dolce. *In Handbook on Ontologies*. Springer.

[Bourigault *et al.*, 2004] Bourigault, D., Aussenac-Gilles, N. et Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18(1):87–110.

[Bourigault et Charlet, 2005] Bourigault, D. et Charlet, J. (2005). Construction d'un index thématique de l'ingénierie des connaissances. *In Conférence d'Ingénierie de Connaissances*, pages 29–47, De Régine Teulier.

[Brewster *et al.*, 2004] Brewster, C., Alani, H., Dasmahapatra, S. et Wilks, Y. (2004). Data-driven ontology evaluation. *In Proceedings of the Language Resources and Evaluation Conference (LREC 2004)*, pages 164–168, Lisbon, Portugal.

[Buche *et al.*, 2012] Buche, P., Dibie-Barthélémy, J., Ibanescu, L. et Soler, L. (2012). Fuzzy web data tables integration guided by a termino-ontological resource. *IEEE Transactions on Knowledge and Data Engineering*, in press:000–014.

[Cheng *et al.*, 2008] Cheng, C. P., Lau, G. T., Pan, J., Law, K. H. et Jones, A. (2008). Domain-specific ontology mapping by corpus-based semantic similarity.

[Choi *et al.*, 2006] Choi, N., Song, I.-Y. et Han, H. (2006). A survey on ontology mapping. *SIGMOD Rec.*, 35:34–41.

[Chrisment *et al.*, 2008] Chrisment, C., Haemmerlé, O., Hernandez, N. et Mothe, J. (2008). Méthodologie de transformation d'un thésaurus en une ontologie de domaine. *Revue d'Intelligence Artificielle*, 22:7–37.

[Cimiano, 2006] Cimiano, P. (2006). *Ontology Learning and Population from Text : Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Cimiano et Völker, 2005] Cimiano, P. et Völker, J. (2005). Text2onto - a framework for ontology learning and datadriven change discovery.

[Corby *et al.*, 2006] Corby, O., Dieng-Kuntz, R., Faron-Zucker, C. et Gandon, F. (2006). Searching the semantic web : Approximate query processing based on ontologies. *IEEE Intelligent Systems*, 21(1):20–27.

[Cruse, 1986] Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

[d'Aquin *et al.*, 2009] d'Aquin, M., Schlicht, A., Stuckenschmidt, H. et Sabou, M. (2009). Modular ontologies. chapitre Criteria and Evaluation for Ontology Modularization Techniques, pages 67–89. Springer-Verlag, Berlin, Heidelberg.

[de Bruijn *et al.*, 2006] de Bruijn, J., Ehrig, M., Feier, C., Martíns-Recuerda, F., Scharffe, F. et Weiten, M. (2006). *Ontology Mediation, Merging, and Aligning*, pages 95–113.

[Deladrière et Kilian, 2009] Deladrière, J. et Kilian, C. (2009). *Organisez vos idées avec le Mind Mapping*. Dunod.

[Desmontils et Jacquin, 2002] Desmontils, E. et Jacquin, C. (2002). Annotations sur le web : notes de lecture. *Journées scientifiques Web Sémantique, CNRS*, 171.

[Després et Szulman, 2008] Després, S. et Szulman, S. (2008). Réseau terminologique versus ontologie. *In Toht 2008*, pages 17–34, France.

[Dingli *et al.*, 2003] Dingli, A., Ciravegna, F. et Wilks, Y. (2003). Automatic semantic annotation using unsupervised information extraction and integration. *In Workshop on Knowledge Markup and Semantic Annotation*.

[Ehrig et Staab, 2004] Ehrig, M. et Staab, S. (2004). Qom - quick ontology mapping. *In International Semantic Web Conference*, pages 683–697.

[Ellouze *et al.*, 2012] Ellouze, N., Lammari, N. et Métais, E. (2012). Citom : An incremental construction of multilingual topic maps. *Data Knowl. Eng.*, 74:46–62.

[Euzenat, 2004] Euzenat, J. (2004). An api for ontology alignment. *In ISWC 2004*, pages 698–712. Springer.

[Euzenat *et al.*, 2004] Euzenat, J., Bach, T. L., Barrasa, J., Bouquet, P., Bo, J. D., Dieng-Kuntz, R., Ehrig, M., Hauswirth, M., Jarrar, M., Lara, R., Maynard, D., Napoli, A., Stamou, G., Stuckenschmidt, H., Shvaiko, P., Tessaris, S., Acker, S. V. et Zaihrayeu, I. (2004). State of the art on ontology alignment.

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

[Fürst et Trichet, 2006] Fürst, F. et Trichet, F. (2006). Raisonner sur des ontologies lourdes à l'aide de graphes conceptuels. *In INFORSID*, pages 879–894.

[Golick *et al.*, 2011] Golick, W., Warnier, P. et Nedellec, C. (2011). Corpus-based extension of termino-ontology by linguistic analysis - a use-case in biomedical event extraction. *In Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, pages 37–39, Paris, France. INALCO.

[Grefenstette, 1994] Grefenstette, G. (1994). Corpus-derived first, second and third order affinities. *In EURALEX*, Amsterdam.

[Gruber, 1995] Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Presented at the Padua workshop on Formal Ontology, March 1993, later published in International Journal of Human-Computer Studies*, 43:907–928.

[Guissé *et al.*, 2011] Guissé, A., Lévy, F. et Nazarenko, A. (2011). Un moteur sémantique pour explorer des textes réglementaires. *In 22èmes journées francophones d'ingénierie des connaissances*, page 8, Chambéry, France.

[Haase et Stojanovic, 2005] Haase, P. et Stojanovic, L. (2005). Consistent evolution of owl ontologies. *In ESWC*, pages 182–197.

[Handschuh et Staab, 2003] Handschuh, S. et Staab, S. (2003). Cream : Creating metadata for the semantic web. *Computer Network*, 42(5):579–598.

[Hanif *et al.*, 2006] Hanif, M. S., Seki, Y. et Aono, M. (2006). Automatic alignment of ontology eliminating the probable misalignments. *In ASWC*, pages 212–218.

[Hignette *et al.*, 2009] Hignette, G., Buche, P., Dibie-Barthélemy, J. et Haemmerlé, O. (2009). Fuzzy annotation of web data tables driven by a domain ontology. *In ESWC*, pages 638–653.

[Hoekstra *et al.*, 2007] Hoekstra, R., Breuker, J., Bello, M. D. et Boer, A. (2007). The LKIF core ontology of basic legal concepts. *In Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007)*.

[Huza *et al.*, 2006] Huza, M., Harzallah, M. et Trichet, F. (2006). Ontomas : vers un assistant d'alignement d'ontologies. *In 7ème Journées Francophones d'Ingénierie des Connaissances (IC)*, Nantes.

[Isaac *et al.*, 2007] Isaac, A., van der Meij, L., Schlobach, S. et Wang, S. (2007). An empirical study of instance-based ontology matching. *In ISWC/ASWC*, pages 253–266.

[Jaccard, 1901] Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272.

[Jacquemin et Ploux, 2006] Jacquemin, B. et Ploux, S. (2006). Corpus spécialisé et ressource de spécialité : l'information forme le sens. *In Actes des Journées Scientifiques du CRTT : Corpus et dictionnaires de langues de spécialité*, Lyon.

[Jaro, 1989] Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84:414–420.

[Jin *et al.*, 2009] Jin, L., Jian, W., Yin, J., Li, Y. et Deng, S. (2009). Ontology alignment based service interface adaptation. *In IEEE International Conference on Services Computing*, pages 494–497.

[Jouanot, 2000] Jouanot, F. (2000). Un modèle sémantique pour l'interopérabilité de systèmes d'information. pages 347–364, IAE - 15, Quai Claude Bernard, Lyon, 69007.

[Kahan et Koivunen, 2001] Kahan, J. et Koivunen, M.-R. (2001). Annotea : an open rdf infrastructure for shared web annotations. *In Proceedings of the 10th international conference on World Wide Web*, pages 623–632, New York, NY, USA. ACM.

[Kassaie *et al.*, 2012] Kassaie, B., Vazifedoost, A. et Rahgozar, M. (2012). Application of textual corpus in ontology matching. *International Journal of Information and Education Technology*.

[Kefi *et al.*, 2006] Kefi, H., Reynaud, C. et Safar, B. (2006). Techniques structurelles pour l'alignement de taxonomies sur le web. *In Atelier Fouille du Web - EGC 2006*.

[Kingkaew, 2012] Kingkaew, C. (2012). Using unstructured documents as background knowledge for ontology matching. *In International Conference on Machine Learning and Computer Science (IMLCS'2012)*, pages 147–151, Phuket (Thailand).

[Kiryakov *et al.*, 2004] Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D. et Goranov, K. M. (2004). Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, 2:49–79.

[Koo *et al.*, 2003] Koo, S. O., Lim, S. Y. et Lee, S. J. (2003). Building an ontology based on hub words for information retrieval. *In Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, pages 466–469, Washington, DC, USA. IEEE Computer Society.

[Kwak et Yong, 2010] Kwak, J. et Yong, H.-S. (2010). Ontology matching based on hypernym, hyponym, holonym, and meronym sets in wordnet. *International Journal of Web & Semantic Technology*, 1:1–14.

[Lame, 2002] Lame, G. (2002). *CONSTRUCTION D'ONTOLOGIE A PARTIR DE TEXTES. Une ontologie du droit dédiée à la recherche d'information sur le Web*. Thèse de doctorat, Centre de Recherches Informatiques de l'Ecole des Mines.

[Lassila et McGuinness, 2001] Lassila, O. et McGuinness, D. L. (2001). The role of frame-based representation on the semantic web. Rapport technique KSL-01-02, Stanford University, Stanford.

[Lefèvre, 2000] Lefèvre, P. (2000). *La recherche d'informations : du texte intégral au thésaurus*. Hermes Science.

[Levenshtein, 1966] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.

[Lin, 1998] Lin, D. (1998). An information-theoretic definition of similarity. *In In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

[Ma *et al.*, 2013] Ma, Y., Lévy, F. et Nazarenko, A. (2013). Annotation sémantique pour des domaines spécialisés et des ontologies riches. *In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 464–478, Les Sables d'Olonne, France.

[Maedche *et al.*, 2002] Maedche, A., Pekar, V. et Staab, S. (2002). *Ontology Learning Part One - On Discovering Taxonomic Relations from the Web*, pages 301–322. Springer Verlag.

[Martin *et al.*, 2004] Martin, A. F., Garofolo, J. S., Fiscus, J. G., Le, A. N., Pallett, D. S., Przybocki, M. A. et Sanders, G. A. (2004). Nist language technology evaluation cookbook. *In LREC*.

[Microsystems, 2001] Microsystems, S. (2001). The Upper Cyc Ontology in XTM. Rapport technique, Sun Microsystems.

[Niles et Pease, 2001] Niles, I. et Pease, A. (2001). Towards a standard upper ontology. *In Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA.

[Ninova *et al.*, 2005] Ninova, G., Nazarenko, A., Thierry, H. et Szulman, S. (2005). Comment mesurer la couverture d'une ressource terminologique pour un corpus ? *In Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2005*, pages 293–302. ATALA.

[Novak et Cañas, 2006] Novak, J. D. et Cañas, A. J. (2006). The theory underlying concept maps and how to construct them. Rapport technique, Institute for Human and Machine Cognition.

[Quaero, 2008] Quaero (2008). Quaero. `http://www.quaero.org/`, consulté le 01/01/12.

[Quix *et al.*, 2011] Quix, C., Roy, P. et Kensche, D. (2011). Automatic selection of background knowledge for ontology matching. *In Proceedings of the International Workshop on Semantic Web Information Management*, pages 51–57, New York, NY, USA. ACM.

[Rada *et al.*, 1989] Rada, R., Mili, H., Bicknell, E. et Blettner, M. (1989). Development and application of a metric on semantic nets. *In IEEE Transactions on Systems, Man and Cybernetics*, pages 17–30.

[Rahm et Bernstein, 2001] Rahm, E. et Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350.

[Sabou et Motta, 2006] Sabou, M. et Motta, E. (2006). Modularization : a key for the dynamic selection of relevant knowledge components. *In In Proceedings of the ISWC 2006 Workshop on Modular Ontologies.*

[Safar et Reynaud, 2009] Safar, B. et Reynaud, C. (2009). Alignement d'ontologies basé sur des ressources complémentaires : Illustration sur le système taxomap. *Revue Technique et Science Informatiques (TSI)*, pages 1211–1232.

[Sampson, 2005] Sampson, J. (2005). Ontology alignment in agent systems : Current and future challenges. *In Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, volume 1, pages 168–173, Washington, DC, USA. IEEE Computer Society.

[Sellami *et al.*, 2012] Sellami, Z., Aussenac-Gilles, N., Gleizes, M. P. et Camps, V. (2012). Dynamo, un outil de construction et d'évolution d'ontologies à partir de textes. *Technique et Science Informatiques*, 31(1):97–124.

[Stuckenschmidt *et al.*, 2005] Stuckenschmidt, H., Serafini, L. et Wache, H. (2005). Reasoning about ontology mappings. Rapport technique, ITC-IRST, Trento, Italy.

[Tricot et Roche, 2004] Tricot, C. et Roche, C. (2004). Cartographie sémantique : des connaissances à la carte. *In EGC*, page 171.

[Valtchev, 1999] Valtchev, P. (1999). *Construction automatique de taxonomies pour l'aide à la représentation de connaissances par objets.* Thèse de doctorat, Univ. Joseph Fourier - Grenoble 1, Rocquencourt.

[van Heijst *et al.*, 1997] van Heijst, G., Schreiber, A. T. et Wielinga, B. J. (1997). Using explicit ontologies in kbs development. *Int. J. Hum.-Comput. Stud.*, 46:183–292.

[Vandenbussche *et al.*, 2011] Vandenbussche, P.-Y., vatant, B. et Rozat, L. (2011). Qualité et robustesse dans le web de données : Linked open vocabularies. *In Atelier Qualité et Robustesse pour le Web de Données IC 2011.*

[Wang et Xu, 2008] Wang, P. et Xu, B. (2008). Debugging ontology mappings : A static approach. *Computing and Informatics*, 27(1):21–36.

[Winkler, 1999] Winkler, W. E. (1999). The state of record linkage and current research problems.

[Wu et Palmer, 1994] Wu, Z. et Palmer, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL'94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Zablith *et al.*, 2008] Zablith, F., d'Aquin, M., Sabou, M. et Motta, E. (2008). Using background knowledge for ontology evolution. *In Proceedings of the ISWC International Workshop on Ontology Dynamics (IWOD.*

[Zargayouna *et al.*, 2007] Zargayouna, H., Safar, B. et Reynaud, C. (2007). TaxoMap in the OAEI 2007 alignment contest. *In Proceedings of The Second International Workshop on Ontology Matching (OM'07)*, pages 268–275, Busan, Corée, République De.

[Zhdanova et Shvaiko, 2006] Zhdanova, A. V. et Shvaiko, P. (2006). Community-driven ontology matching. *In Proceedings of the 3rd European conference on The Semantic Web : research and applications*, pages 34–49, Berlin, Heidelberg. Springer-Verlag.