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.

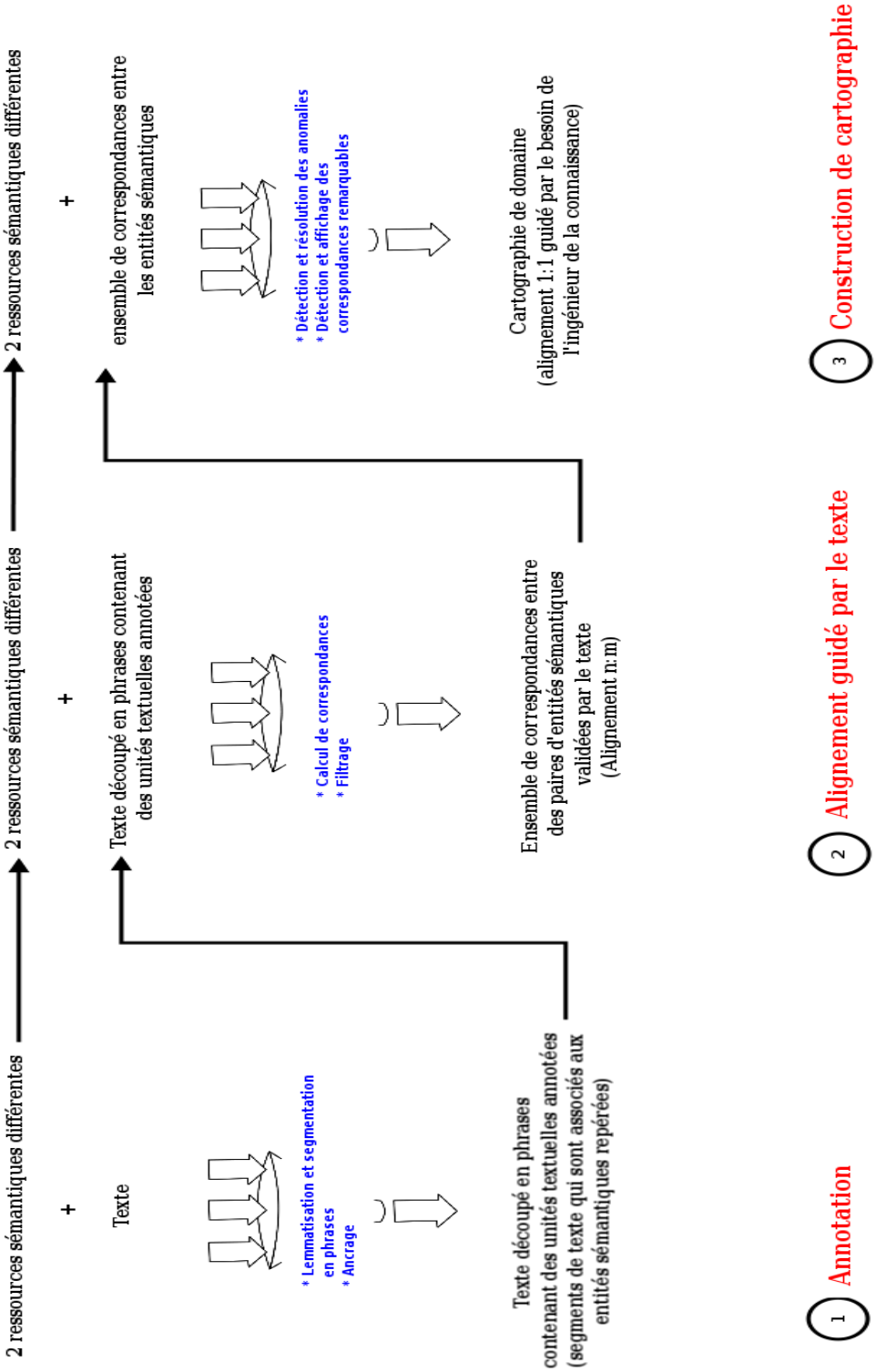


FIGURE 2.1 – Méthodologie de construction d'une cartographie à partir de ressources sémantiques

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

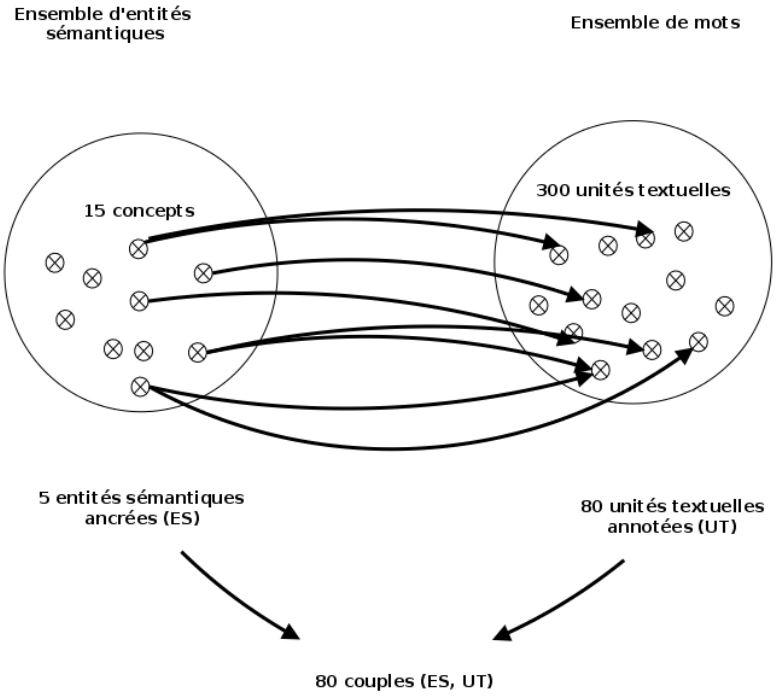


FIGURE 2.2 – Principe d'annotation dans notre méthodologie de construction de la cartographie de domaine

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.

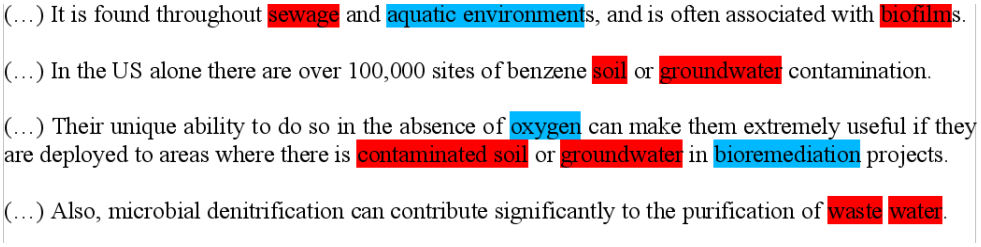


FIGURE 2.4 – Exemple de texte annoté par les ontologies de la figure 2.5

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.

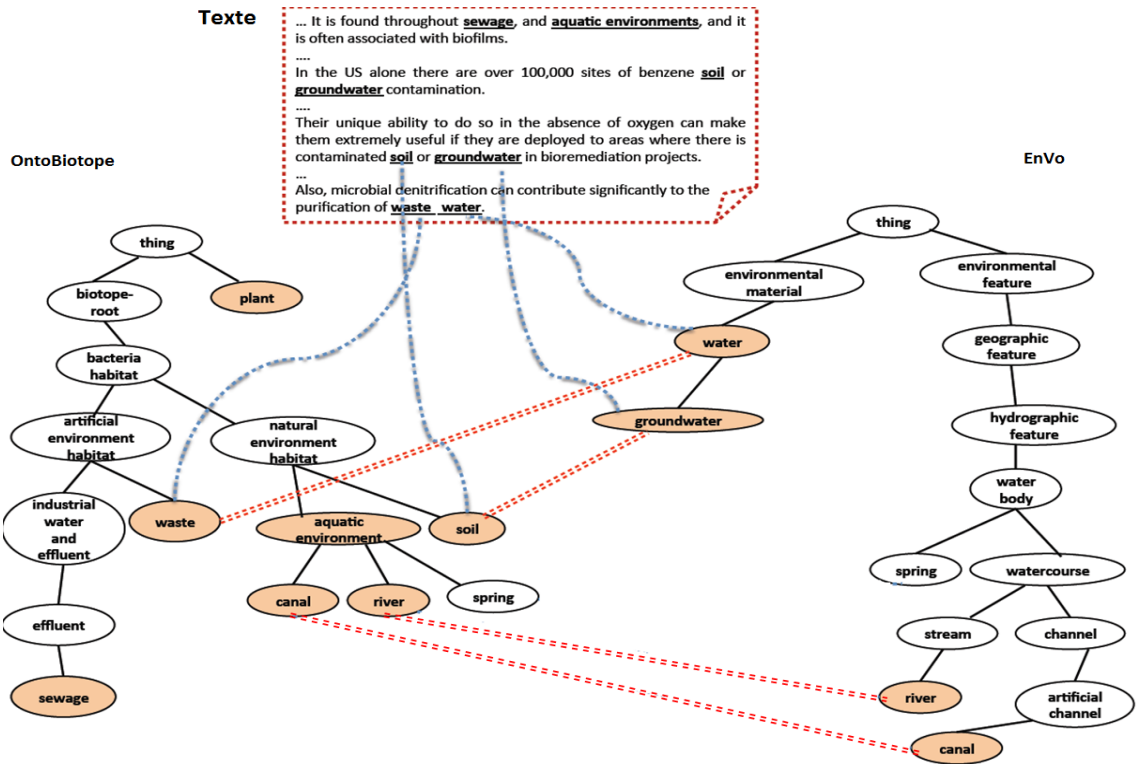


FIGURE 2.5 – Exemple d'alignement entre deux ontologies

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/

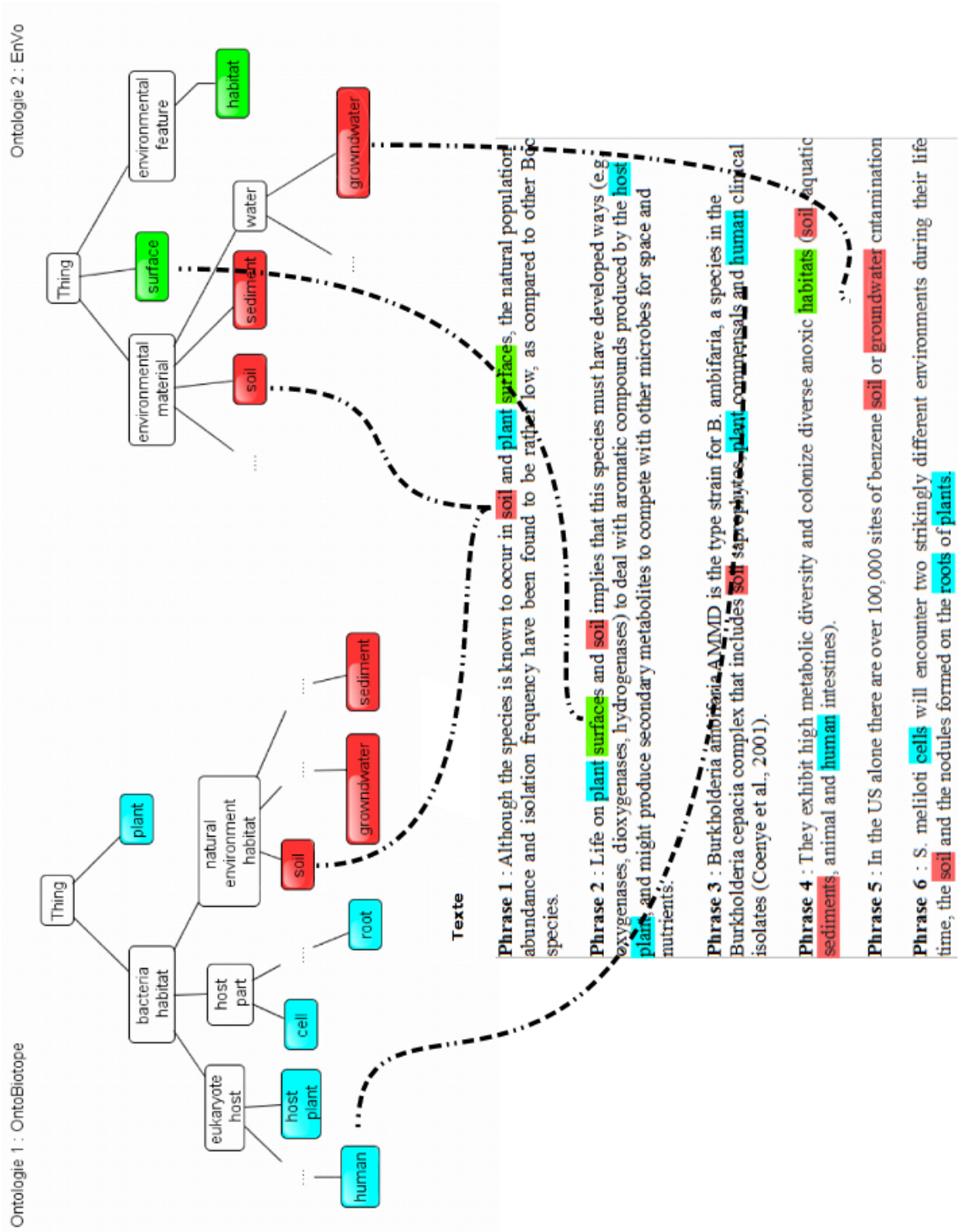


FIGURE 2.7 – Annotation d'un texte par deux ontologies (OntoBiotope et EnVo)

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.

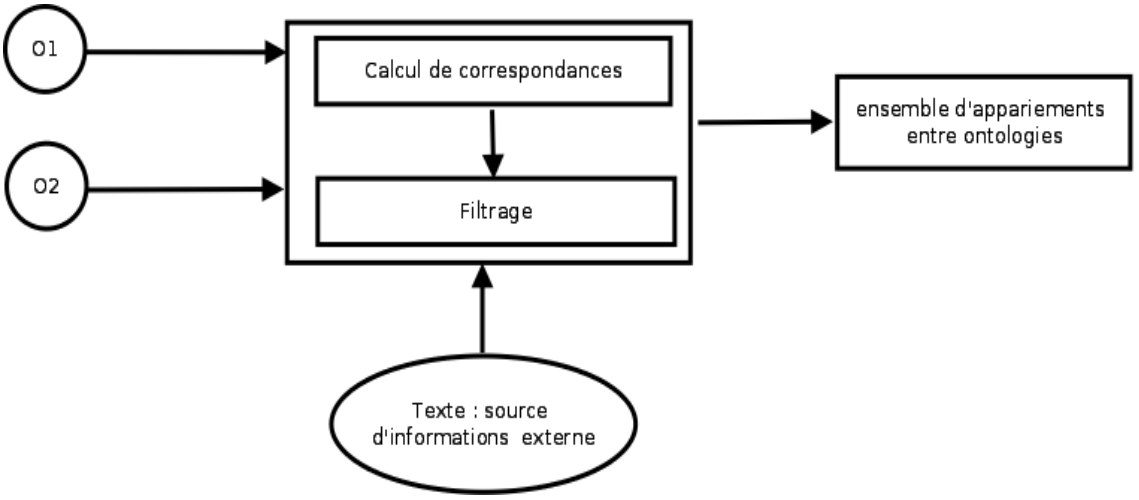


FIGURE 3.1 – Processus d'alignement guidé par le texte

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

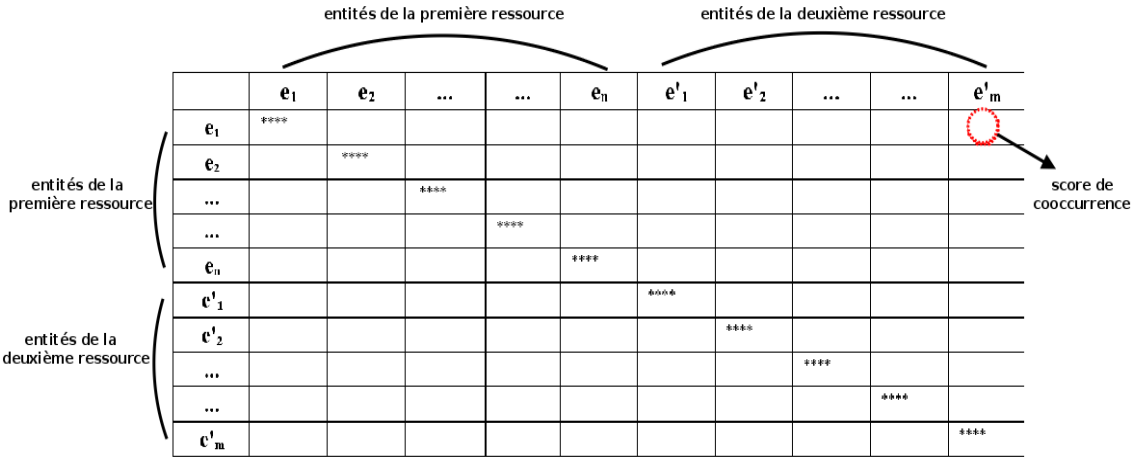


FIGURE 3.2 – Matrice de cooccurrences entre entités sémantiques

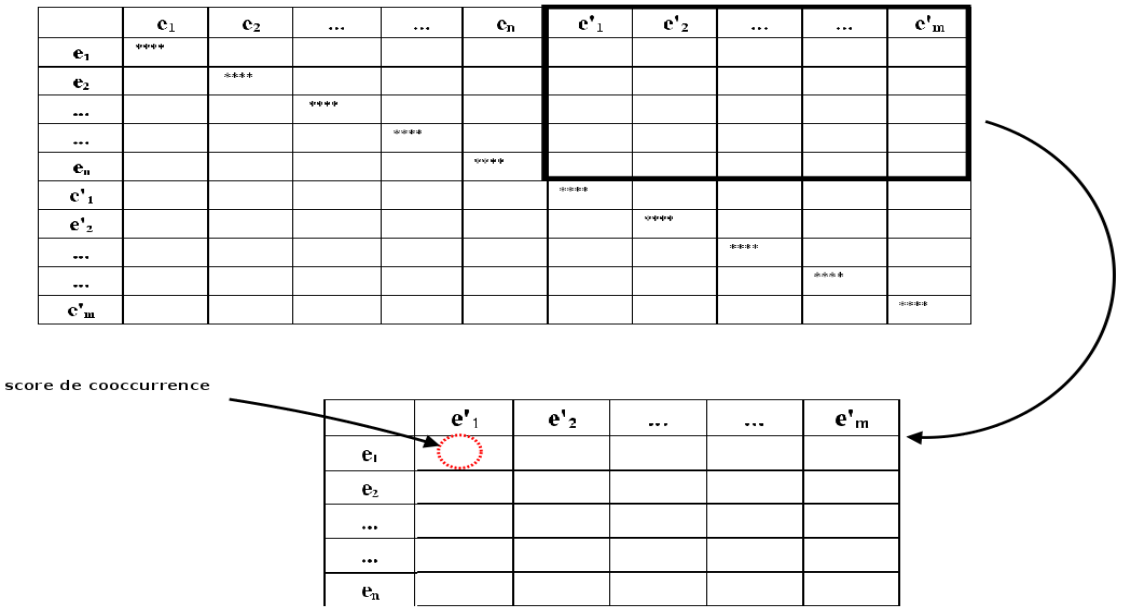


FIGURE 3.3 – Partie de la matrice de cooccurrences pour extraire les relations d'association

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 :

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

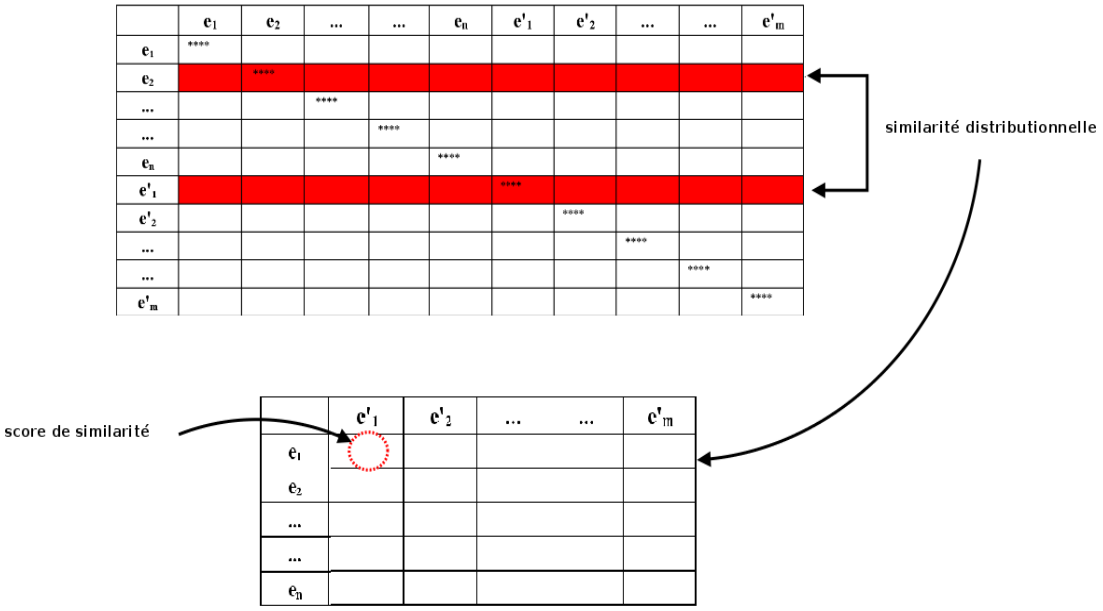


FIGURE 3.4 – Matrice de calcul de similarité pour dériver les relations d'équivalence

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).

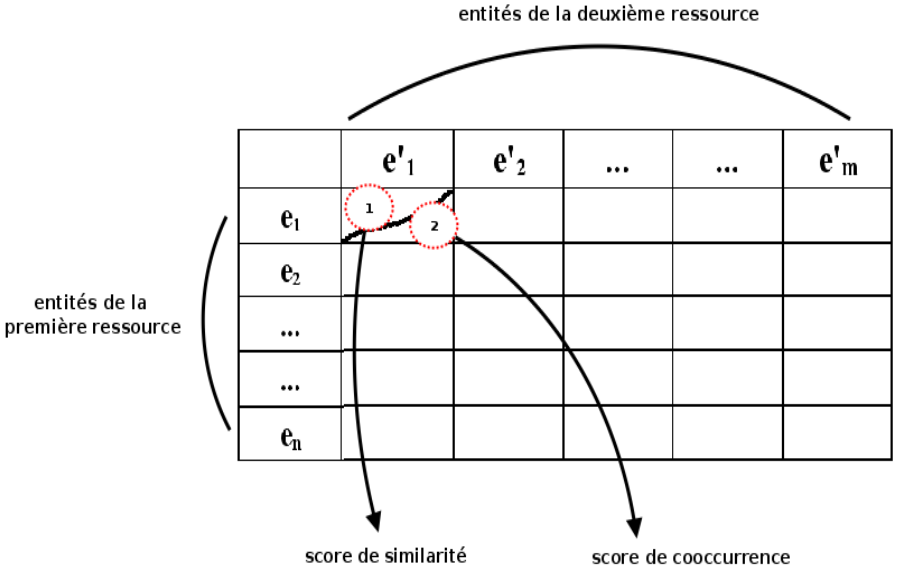
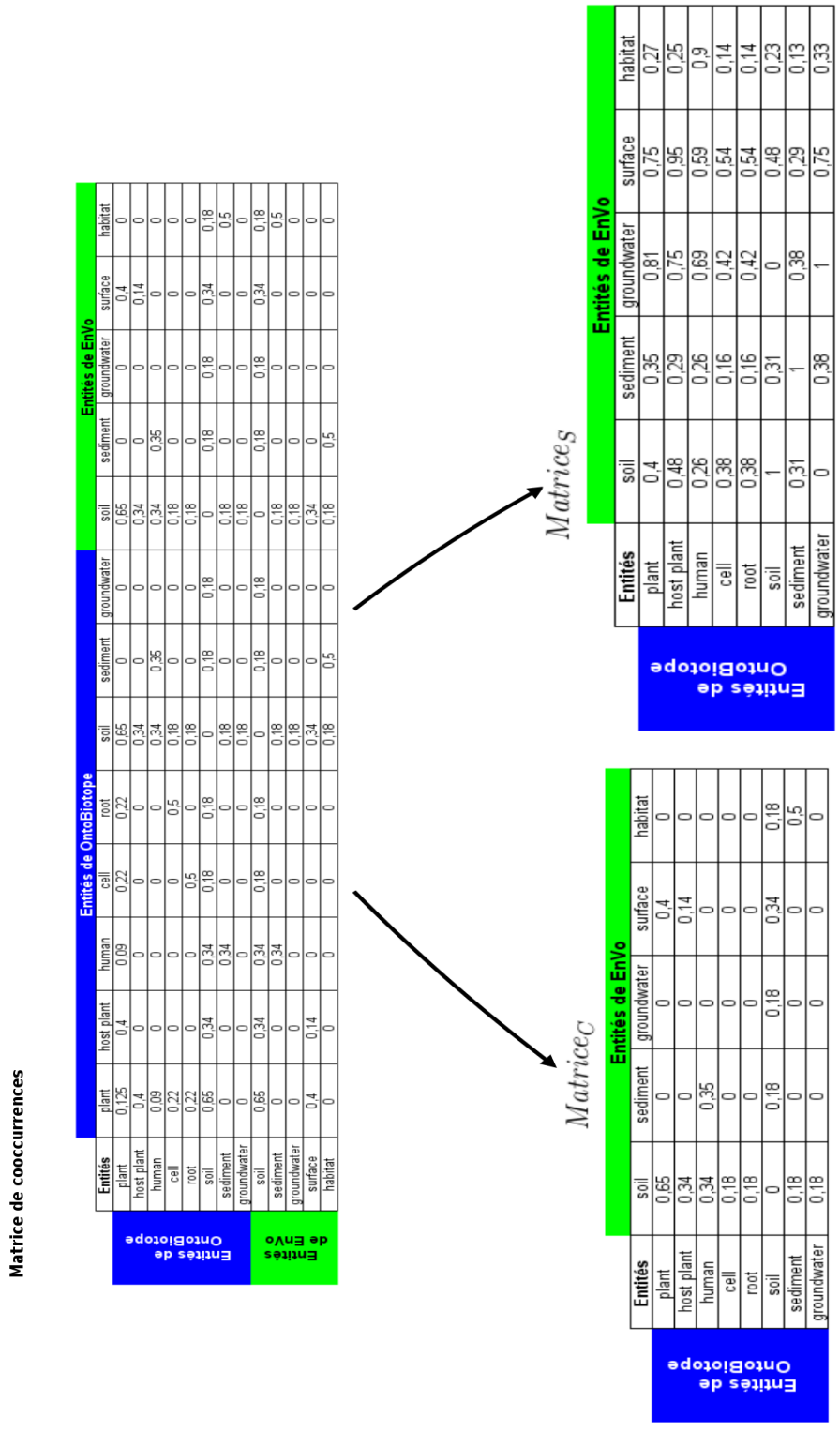


FIGURE 3.5 – Matrice globale contenant les deux matrices : $Matrice_C$ et $Matrice_S$

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.

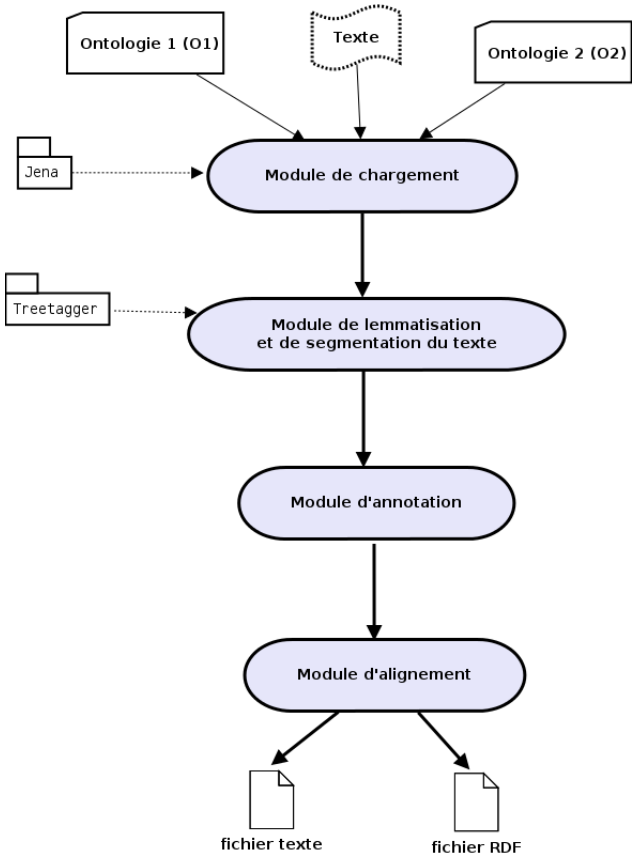


FIGURE 3.7 – Architecture de l'application TOM

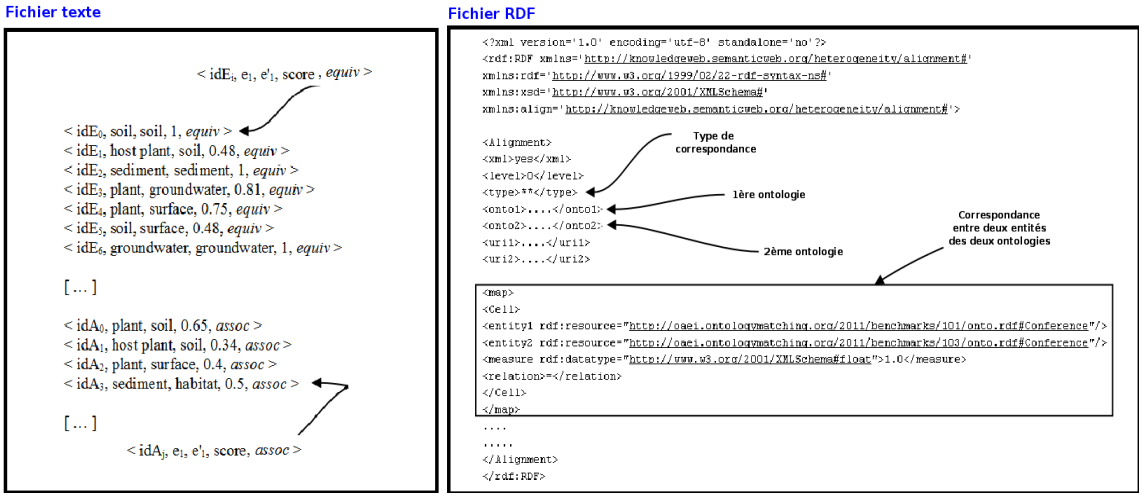


FIGURE 3.8 – Exemple de résultats obtenus par l'alignement des deux ontologies OntoBiotope et EnvO

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.

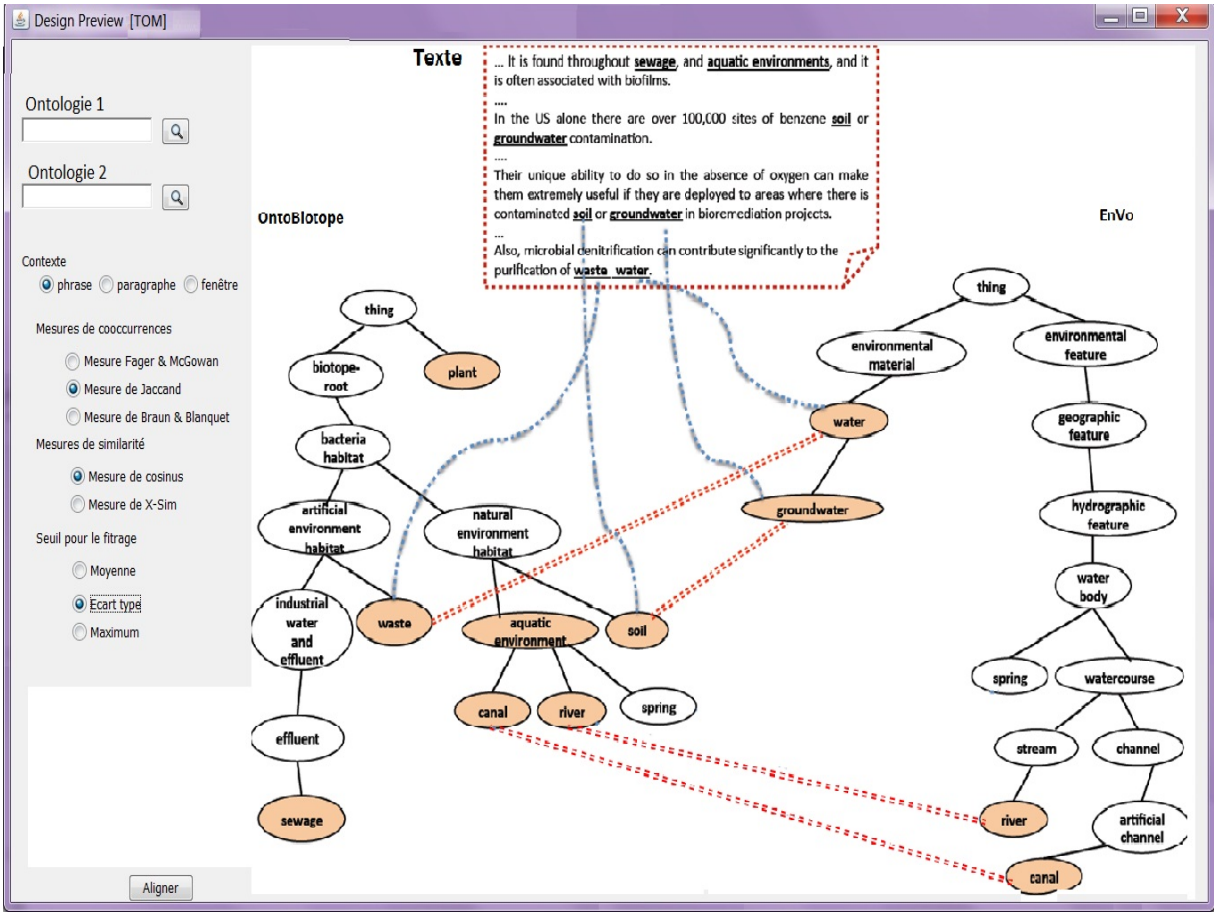


FIGURE 3.9 – Maquette de l'interface de notre méthode d'alignement TOM

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.