Alignement d'ontologies : exploitation des ontologies liées sur le web de données

Thomas Hecht[*,**], Patrice Buche[**,***], Juliette Dibie-Barthélemy[*,****],
Liliana Ibănescu[*,****], Cássia Trojahn dos Santos[♯]

[*]INRA - Mét@risk, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France
[**]INRA - UMR IATE, 2 place Pierre Viala, F-34060 Montpellier Cedex 2, France
Patrice.Buche@supagro.inra.fr
[***]LIRMM, Montpellier, France
[****]AgroParisTech, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France
{Juliette.Dibie, Liliana.Ibanescu}@agroparistech.fr
[♯]IRIT & UTM2, 5 allées Antonio Machado, F-31058 Toulouse Cedex 9, France
Cassia.Trojahn@irit.fr

Résumé. Nous proposons dans cet article une méthode d'alignement d'une ontologie source avec des ontologies cibles déjà publiées et liées sur le web de données. Nous présentons ensuite un retour d'expérience sur l'alignement d'une ontologie dans le domaine des sciences du vivant et de l'environnement avec AGROVOC et NALT.

1 Introduction

Les ontologies sont nées d'un besoin de standardisation des vocabulaires ressenti dans de nombreux domaines et connaissent, depuis quelques années, un succès qui a en partie été porté par l'explosion du web de données (cf. Heath et Bizer (2011)), dont la promesse est le partage à grande échelle de données ouvertes et liées. Le web d'aujourd'hui est donc non seulement une source inépuisable d'informations, mais il est devenu, à travers le web de données, un vecteur de développement et de diffusion incontournable pour les ingénieurs de la connaissance et les fournisseurs de données.

Participer au web de données suppose d'être capable de s'interconnecter avec les données et ontologies déjà présentes et disponibles. Publier une nouvelle ontologie nécessite donc au préalable de la relier avec les "bonnes" ontologies publiées sur le web de données comportant des concepts similaires dans des domaines similaires. Relier deux ontologies consiste en fait à les aligner, i.e. trouver des correspondances entre les entités (concepts, propriétés ou instances) des deux ontologies. L'alignement d'ontologies est un domaine en plein essor qui a donné lieu à de nombreux travaux de recherche. Nous pouvons notamment citer Shvaiko et Euzenat (2013), Bernstein et al. (2011), Rahm (2011) ou encore Euzenat et Shvaiko (2007). Ces travaux ont permis de définir des formalismes et des outils, qui sont régulièrement évalués dans le cadre de la campagne d'évaluation de l'OAEI [1] (*Ontology Alignment Evaluation Initiative*) sur des jeux de données de différents domaines.

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

Nous proposons dans cet article une méthode d'alignement d'une ontologie source avec des ontologies cibles déjà publiées sur le web de données et liées entre elles. Ces ontologies peuvent en fait être des thésaurus, des ontologies ou encore des ressources termino-ontologique qui peuvent être exprimées dans différents langages de représentation. Notre méthode repose sur un principe de raffinage d'alignements qui exploite des ontologies liées sur le web de données. Cette méthode prend en entrée un ensemble d'alignements qui peuvent être générés en utilisant différentes approches d'alignement.

Nous proposons d'illustrer la méthode proposée par un retour d'expérience sur l'alignement d'une ontologie dans le domaine des sciences du vivant et de l'environnement. Dans ce domaine, plusieurs thésaurus ont été créés au niveau international et publiés sur le web de données pour faire face au besoin de standardisation des vocabulaires. Les deux plus importants sont actuellement AGROVOC [2] et NALT [3]. AGROVOC a été créé dans les années 1980 par la FAO (Food and Agriculture Organization of the United Nations) comme un thésaurus structuré multilingue pour les domaines de l'agriculture, de la sylviculture, de la pêche, de l'alimentation et de domaines apparentés (comme l'environnement). Il est actuellement disponible en 19 langues, avec une moyenne d'environ 40 000 termes dans chaque langue (cf. Caracciolo et al. (2012)). NALT est un thésaurus bilingue comparable à AGROVOC en terme de domaine couvert et maintenu par la USDA (United States Department of Agriculture). Il est actuellement composé d'environ 91 000 termes en anglais et espagnol. A titre d'exemple, le vocabulaire d'AGROVOC est actuellement relié à celui de 11 ressources internationales comme GeoNames [4], DBpedia [5] et GEMET [6]. De plus, 13 390 termes d'AGROVOC sont actuellement reliés aux termes de NALT (cf. Caracciolo et al. (2012)). Dans ce papier, nous nous intéressons à l'alignement d'une ressource termino-ontologique naRyQ (n-ary Relations between Quantitative experimental data) non encore publiée (cf. (Buche et al., 2013)) avec AGROVOC et NALT. naRyQ contient environ 1 100 concepts structurés en plusieurs sous-domaines, tels que les produits alimentaires, les microorganismes et les emballages.

Nous présentons, dans la section 2, notre méthode d'alignement d'une ontologie avec des ontologies liées et, dans la section 3, les résultats de notre expérimentation dans le domaine des sciences du vivant et de l'environnement. Enfin, nous concluons dans la section 4.

2 Méthode d'alignement d'une ontologie avec des ontologies liées

Nous présentons dans cette section une méthode permettant d'aligner une ontologie source, O_s, avec deux ontologies liées O_c^1 et O_c^2. Cette méthode propose de prendre en compte les spécificités des ontologies à lier qui peuvent être des ontologies, des thésaurus, des ressources termino-ontologique et qui peuvent être exprimées dans différents langages de représentation avec différents niveaux d'expressivité, ici OWL DL et SKOS. Une Ressource Termino-Ontologique (RTO) (cf. Reymonet et al. (2007); Roche et al. (2009); McCrae et al. (2011))

2. http ://aims.fao.org/standards/agrovoc/about
3. http ://agclass.nal.usda.gov/agt.shtml
4. http ://www.geonames.org/
5. http ://dbpedia.org/
6. http ://www.eionet.europa.eu/gemet

est un modèle hybride combinant une composante conceptuelle et une composante terminologique : les termes, associés aux concepts qu'ils dénotent, sont des manifestations lexicales différentes d'une même entité (synonymes, abréviations, etc.). Nous parlerons dans la suite de méthode d'alignement d'ontologies par souci de simplicité.

Cette méthode propose de s'appuyer sur des méthodes d'alignements existantes en utilisant des variantes des ontologies à aligner dans différents langages de représentation et d'exploiter les liens définis entre les deux ontologies cibles. Elle se décompose en deux étapes : une première étape qui consiste à aligner l'ontologie source O_s avec chacune des deux ontologies cibles O_c^1 et O_c^2 et une deuxième étape qui consiste à affiner les correspondances trouvées dans la première étape à l'aide des liens entre les ontologies cibles. Notre méthode s'appuie sur les définitions de Euzenat et Shvaiko (2007) dont nous rappellerons certaines par soucis de clarté.

2.1 Première étape : alignement

La première étape de notre méthode consiste à aligner l'ontologie source O_s avec chacune des deux ontologies cibles O_c^1 et O_c^2. Elle exploite d'une part les spécificités des ontologies à aligner qui peuvent être de différentes natures et exprimées dans différents langages, et, d'autre part, les méthodes d'alignements existantes.

2.1.1 Ensemble de variantes d'ontologie

Les ontologies source et cible à aligner pouvant être des thésaurus, des ontologies ou encore des RTO exprimés dans différents langages de représentation, nous proposons d'associer à chaque ontologie, un ensemble de variantes, qui pourront être exploitées différemment par les différentes méthodes d'alignement existantes, défini comme suit :

Définition 2.1 (Ensemble de variantes d'une ontologie) *L'*ensemble V_O des variantes d'une ontologie O *est constitué de ses transformations en différents langages de représentation* $L_1, L_2, \ldots$. *Il contient nécessairement la version « originale », O^{orig}, de l'ontologie.*
$V_O = \{O^{orig}, O^{L_1}, O_1^{L_2}, O_2^{L_2}, \ldots\}$, *où $O_j^{L_i}$ est la $j^{ème}$ transformation de l'ontologie O en utilisant le langage de représentation L_i.*

2.1.2 Alignement des variantes des ontologies

Le processus d'alignement d'ontologies prend en entrée deux ontologies et produit en sortie un ensemble de correspondances entre les entités de ces ontologies. Selon Euzenat et Shvaiko (2007), ce processus est défini comme suit :

Définition 2.2 (Processus d'alignement (cf. Euzenat et Shvaiko (2007))) *Le* processus d'alignement d'ontologies *est une fonction f qui, appliquée à deux ontologies O_s et O_c et à un alignement initial A^{orig}, produit un alignement orienté $A^f_{O_s,O_c}$ entre les deux ontologies ($O_s \rightarrow O_c$). Le processus dépend d'un ensemble de paramètres p et peut faire appel à un ensemble de ressources externes r : $A^f_{O_s,O_c} = f(O_s, O_c, A^{orig}, p, r)$*

Définition 2.3 (Correspondance (cf. Euzenat et Shvaiko (2007))) *Soient deux ontologies O_s et O_c, une* correspondance *c^f issue des résultats d'un processus d'alignement f est une relation r établie entre deux entités e_s et e_c, notée $c^f = \langle id, e_s, e_c, r, n\rangle$, où : $c^f \in A^f_{O_s,O_c}$; $e_s \in O_s$ et $e_c \in O_c$; $r \in \{\equiv, \sqsubseteq, \sqsupseteq\}$; n est un degré de confiance (en général, $n \in [0, 1]$).*

Notre méthode consiste à lancer plusieurs processus d'alignement sur les variantes des ontologies à aligner. Soient O_s l'ontologie source et O_c une des deux ontologies cibles. Soit $F = \{f_1, f_2, \ldots\}$ l'ensemble de processus d'alignement qui sont lancés pour aligner les ontologies O_s et O_c. Chaque processus d'alignement f_i est lancé sur un couple de variantes (O_s^j, O_c^k) où $O_s^j \in V_{O_s}$ et $O_c^k \in V_{O_c}$ et permet d'obtenir l'ensemble de correspondances suivant :

$$A^{f_i}_{O_s^j, O_c^k} = f_i(O_s^j, O_c^k, \emptyset, p, r) \tag{1}$$

Le résultat de l'alignement de l'ontologie source O_s avec l'une des deux ontologies cibles O_c est l'ensemble, noté $\overset{agr}{C}_{O_s \to O_c}$, des ensembles d'alignement obtenus pour chaque processus d'alignement et chaque couple de variantes d'ontologies retenus. Ce résultat est noté :

$$\overset{agr}{C}_{O_s \to O_c} = \bigoplus_{i,j,k} A^{f_i}_{O_s^j, O_c^k} \tag{2}$$

On remarquera que le nombre total de processus d'alignement lancés pour obtenir l'alignement de l'ontologie source O_s avec chacune des deux ontologies cibles O_c^1 et O_c^2 est :

$$|V_{O_s}| \times |V_{O_c^1}| \times |F| + |V_{O_s}| \times |V_{O_c^2}| \times |F| \tag{3}$$

2.2 Deuxième étape : raffinage des correspondances trouvées

La deuxième étape consiste à raffiner les ensembles d'ensembles de correspondances $\overset{agr}{C}_{O_s \to O_c^1}$ et $\overset{agr}{C}_{O_s \to O_c^2}$. Ces deux ensembles d'ensembles qui résultent de la concaténation des résultats de plusieurs processus d'alignement lancés sur plusieurs variantes d'ontologies permettent d'obtenir de nombreuses correspondances (qui laisse présager une bonne couverture), mais aussi beaucoup de bruit, i.e. des mauvaises correspondances, qu'il convient de réduire.

Nous proposons deux méthodes de raffinage pour améliorer la qualité des correspondances trouvées lors de la première étape : la première méthode permet de supprimer des correspondances considérées comme ambiguës et donc potentiellement éronnées (section 2.2.1) ; la deuxième méthode permet d'identifier les correspondances considérées comme potentiellement correctes (section 2.2.2).

2.2.1 Suppression des correspondances ambiguës

Nous distinguons trois types d'ambiguïté entre correspondances. Le premier type d'ambiguïté concerne les correspondances obtenues à partir d'une même méthode d'alignement lancée sur différentes variantes des ontologies source et cible, correspondances qui ont la même entité source, la même entité cible et la même relation. Nous proposons de lever les ambiguïtés de type 1 en ne conservant que la correspondance ayant le degré de confiance le plus élevé.

Définition 2.4 (Correspondances ambiguës de type 1) *Soient deux processus d'alignement f_1 et f_2 générés par une même méthode entre deux ontologies O_s et O_c, avec O_s^j et O_c^k leurs variantes respectives. Deux correspondances c^{f_1} et c^{f_2} des ensembles $A^{f_1}_{O_s^{j_1}, O_c^{k_1}}$ et $A^{f_2}_{O_s^{j_2}, O_c^{k_2}}$ sont* ambiguës selon le type 1 *si :*

$$c^{f_1} = \langle id_1, e_s^1, e_c^1, r_1, n_1 \rangle \wedge c^{f_2} = \langle id_2, e_s^2, e_c^2, r_2, n_2 \rangle \wedge e_s^1 = e_s^2 \wedge e_c^1 = e_c^2 \wedge r_1 = r_2.$$

L'ensemble des ensembles de correspondances non ambigüs selon le type 1 est :

$$\overset{agr}{C^*}_{O_s \to O_c} = \bigoplus_{i,j,k} (A^{f_i}_{O^j_s, O^k_c} \backslash \{c^{f_k}\}) \text{ où } c^{f_k} = \begin{cases} c^{f_1} & \text{si } n_1 \leq n_2 \\ c^{f_2} & \text{sinon} \end{cases}$$

Le deuxième type d'ambiguïté identifié entre des correspondances correspond au cas où une entité d'une ontologie source O_s est alignée, par la relation d'équivalence, avec deux entités distinctes d'une ontologie cible O_c. Nous proposons dans ce cas de ne conserver que la correspondance la plus pertinente, i.e. celle qui a *a priori* le degré de confiance le plus élevé. Cependant lorsque ces correspondances n'ont pas été générées par la même méthode d'alignement, leur degré de confiance ne sont pas comparables. Nous proposons de calculer une mesure de similarité *sim*, indépendante des méthodes d'alignement utilisées, pour les deux correspondances à comparer, qui peut, par exemple, s'appuyer sur les mesures de similarité syntaxiques usuelles implémentées dans l'Alignment API (David et al., 2011).

Définition 2.5 (Correspondances ambiguës de type 2) *Soit un ensemble de processus d'alignement $F = \{f_1, f_2, \ldots\}$ entre deux ontologies O_s et O_c. Deux correspondances c^{f_i} et c^{f_j} sont* ambiguës selon le type 2 *si :*

$$c^{f_i} = \langle id_1, e^1_s, e^1_c, \equiv, n_1 \rangle \wedge c^{f_j} = \langle id_2, e^2_s, e^2_c, \equiv, n_2 \rangle \wedge e^1_s = e^2_s \wedge e^1_c \neq e^2_c.$$

L'ensemble des ensembles de correspondances non ambiguës de type 2 est :

$$\overset{agr}{C^{2*}}_{O_s \to O_c} = \bigoplus_{i,j,k} (A^{f_i}_{O^j_s, O^k_c} \backslash \{c^{f_k}\}) \text{ où } c^{f_k} = \begin{cases} c^{f_i} & \text{si } sim(c^{f_i}) \leq sim(c^{f_j}) \\ c^{f_j} & \text{sinon} \end{cases}$$

Le troisième type d'ambiguïté identifié entre des correspondances correspond au cas où deux entités distinctes d'une ontologie source O_s sont alignées, par la même relation, avec une même entité d'une ontologie cible O_c. Nous proposons dans ce cas de ne conserver que la correspondance la plus pertinente, i.e. celle avec la mesure de similarité *sim* la plus élevée.

Définition 2.6 (Correspondances ambiguës de type 3) *Soit un ensemble de processus d'alignement $F = \{f_1, f_2, \ldots\}$ entre deux ontologies O_s et O_c. Deux correspondances c^{f_i} et c^{f_j} sont* ambiguës selon le type 3 *si :*

$$c^{f_i} = \langle id_1, e^1_s, e^1_c, r_1, n_1 \rangle \wedge c^{f_j} = \langle id_2, e^2_s, e^2_c, r_2, n_2 \rangle \wedge e^1_s \neq e^2_s \wedge e^1_c = e^2_c \wedge r_1 = r_2.$$

L'ensemble des ensembles de correspondances non ambiguës de type 3 est :

$$\overset{agr}{C^{3*}}_{O_s \to O_c} = \bigoplus_{i,j,k} (A^{f_i}_{O^j_s, O^k_c} \backslash \{c^{f_k}\}) \text{ où } c^{f_k} = \begin{cases} c^{f_i} & \text{si } sim(c^{f_i}) \leq sim(c^{f_j}) \\ c^{f_j} & \text{sinon} \end{cases}$$

2.2.2 Identification des correspondances potentiellement correctes

Lorsque des redondances apparaissent entre des correspondances qui ont été générées à partir d'au moins deux méthodes d'alignement distinctes, nous faisons l'hypothèse que ces correspondances peuvent être considérées comme ayant plus de "chance" d'être bonnes. Nous les conserverons dans un ensemble distinct, noté $\overset{recT}{C}_{O_s \to O_c}$ pour ensemble de recouvrement, afin de les présenter à l'utilisateur comme des correspondances potentiellement correctes.

Définition 2.7 (Ensemble de recouvrement) *Soient deux processus d'alignement f_1 et f_2 générés par deux méthodes distinctes entre deux ontologies O_s et O_c.* L'ensemble de recouvrement $\overset{recT}{C}_{O_s \to O_c}$ *est défini comme suit :*

Si $c^{f_1} = \langle id_1, e_s^1, e_c^1, r_1, n_1 \rangle \wedge c^{f_2} = \langle id_2, e_s^2, e_c^2, r_2, n_2 \rangle \wedge e_s^1 = e_s^2 \wedge e_c^1 = e_c^2 \wedge r_1 = r_2$
alors $c^{f_k} \in \overset{recT}{C}_{O_s \to O_c}$, où $c^{f_k} = \begin{cases} c^{f_1} & \text{si } n_1 \geq n_2 \\ c^{f_2} & \text{sinon} \end{cases}$

Supposons maintenant qu'il existe un alignement $A^{LOD}_{O_c^1 \to O_c^2}$ défini sur le web de données [7] entre les ontologies cibles O_c^1 et O_c^2. Nous proposons de nous appuyer sur la même hypothèse que précédemment à savoir que des correspondances comparables qui apparaissent dans au moins deux processus d'alignement peuvent être considérées comme ayant plus de "chance" d'être bonnes, les correspondances étant ici considérées comme "comparables" si elles permettent d'aligner une même entité à deux entités distinctes mais liées sur le LOD. Ces correspondances seront conservées dans deux ensembles distincts, notés $\overset{LOD}{C}_{O_s \to O_c^1}$ et $\overset{LOD}{C}_{O_s \to O_c^2}$ pour ensembles de recouvrement du LOD, afin de les présenter à l'utilisateur comme des correspondances potentiellement correctes.

Définition 2.8 (Ensemble de recouvrement du LOD) *Soit $A^{LOD}_{O_c^1 \to O_c^2}$ le résultat d'un processus d'alignement sur le web de données entre deux ontologies O_c^1 et O_c^2. Soient un ensemble de processus d'alignement $F^1 = \{f_1^1, f_2^1, \ldots\}$ entre deux ontologies O_s et O_c^1, et, un ensemble de processus d'alignement $F^2 = \{f_1^2, f_2^2, \ldots\}$ entre deux ontologies O_s et O_c^2. Les* ensembles de recouvrement du LOD $\overset{LOD}{C}_{O_s \to O_c^i}$, $i \in [1, 2]$, *sont définis comme suit :*

Si $\exists c \in A^{LOD}_{O_c^1 \to O_c^2} \wedge c = \langle id, e_{c_1}, e_{c_2}, \equiv, n \rangle \wedge c^{f_i^1} \in A^{f_i^1}_{O_s, O_c^1} \wedge c^{f_i^1} = \langle id_1, e_s, e_{c_1}, \equiv, n_1 \rangle \wedge c^{f_j^2} \in A^{f_j^2}_{O_s, O_c^2} \wedge c^{f_j^2} = \langle id_2, e_s, e_{c_2} \equiv, n_2 \rangle$, alors $c^{f_i^1} \in \overset{LOD}{C}_{O_s \to O_c^1}$ et $c^{f_j^2} \in \overset{LOD}{C}_{O_s \to O_c^2}$.

Nous considérons dans la section 3 que l'ensemble des *correspondances potentiellement correctes* entre une ontologie source O_s et une ontologie cible O_c est l'ensemble obtenu par l'union des deux ensembles de recouvrement définis ci-dessus, noté $U^{*,2*,3*}_{O_s \to O_c} = \overset{recT}{C}_{O_s \to O_c} \cup \overset{LOD}{C}_{O_s \to O_c}$, dans lequel ont été supprimées les ambiguïtés de type 1, 2 et 3.

3 Expérimentation

Nous illustrons dans cette section la méthode présentée dans la section 2 pour aligner une ontologie source naRyQ , présentée dans la section 3.1, avec chacune des deux ontologies cibles, AGROVOC et NALT. L'alignement de l'ontologie naRyQ avec AGROVOC est noté naRyQ $\rightarrow$ AGROVOC et l'alignement de l'ontologie naRyQ avec NALT est noté naRyQ $\rightarrow$ NALT.

7. LOD pour Linked Open Data en anglais

3.1 L'ontologie source naRyQ

L'ontologie naRyQ (n-ary Relations between Quantitative experimental data) a été définie pour représenter des relations n-aires entre des données quantitatives expérimentales (Buche et al., 2013). Les spécificités de cette ontologie sont les suivantes : (i) c'est une ressource termino-ontologique qui est un modèle hybride ; (ii) les labels sont disponibles en français et en anglais ; (iii) elle est représentée en OWL DL et SKOS ; (iv) la composante conceptuelle contient environ 1 100 concepts structurés en plusieurs sous-domaines, les plus importants en effectif étant les produits alimentaires ($\approx$ 460 concepts), les microorganismes ($\approx$ 180 concepts) et les emballages ($\approx$ 150 concepts).

3.2 Production des alignements de références

Pour évaluer la qualité des alignements produits et pour comparer entre eux les résultats des processus d'alignement, nous utilisons les mesures de précision et de rappel adaptées à l'alignement d'ontologies (Euzenat et Shvaiko, 2007). Ces mesures s'appuient sur une comparaison avec un alignement de référence, R. La production d'un alignement complet n'étant pas envisageable car elle demanderait de nombreuses personnes, du temps et une expertise pointue, nous avons construit deux alignements, partiels, notés $\overline{R}_{\text{AGROVOC}}$ pour l'alignement naRyQ $\rightarrow$ AGROVOC, et $\overline{R}_{\text{NALT}}$ pour l'alignement naRyQ $\rightarrow$ NALT. Dans la suite, nous n'étudierons que la relation d'équivalence $\equiv$.

Pour chaque ontologie et pour tout concept, nous avons extrait ses labels (e.g. skos :prefLabel, skos :altLabel, rdfs :label, rdfs :comment) en anglais ou en français ainsi que des éléments de structure (e.g. skos :broader, rdfs :subClassOf). Un premier alignement a été produit en utilisant SMOA (*A String Metric for Ontology Alignment*) (Stoilos et al., 2005), une similarité syntaxique destinée à l'alignement d'ontologie. Cet alignement a permis de générer 1 453 correspondances, qui ont ensuite été validées par deux experts en double aveugle et ont été réconciliées, i.e. les experts se sont *a posteriori* mis d'accord. Cette validation a été réalisée en quatre heures, en utilisant un outil de visualisation spécifiquement développé pour cette tâche.

Afin d'améliorer les premiers alignements produits $\overline{R}_{\text{AGROVOC}}$ et $\overline{R}_{\text{NALT}}$, nous les avons enrichis en exploitant les résultats d'alignement obtenus avec la méthode proposée dans ce papier. Les alignements ainsi enrichis, notés $\overline{R}^{+}_{\text{AGROVOC}}$ et $\overline{R}^{+}_{\text{NALT}}$, quoique partiels, seront, dans la suite, utilisés comme alignements de référence. Plus précisèment, les correspondances obtenues avec notre méthode, considérés comme faux positifs par rapport à $\overline{R}^{+}_{\text{AGROVOC}}$ et $\overline{R}^{+}_{\text{NALT}}$, ont été validées par deux experts en double aveugle. Après validation, nous avons obtenu :

- pour l'alignement naRyQ $\rightarrow$ AGROVOC : 368 correspondances validées, 361 concepts de naRyQ étant alignés avec des concepts d'AGROVOC.
- pour l'alignement naRyQ $\rightarrow$ NALT : 428 correspondances validées, 424 concepts de naRyQ étant alignés avec des concepts de NALT.
- 303 concepts de naRyQ sont alignés à la fois avec des concepts d'AGROVOC et des concepts de NALT.

3.3 Protocole expérimental

Deux outils d'alignement, suffisamment génériques, couvrant des problèmes différents et ayant obtenu de bons scores lors des éditions 2011 et 2012 des compétitions OAEI (Aguirre

et al., 2012), ont été retenus : LogMap [8] (Jiménez-Ruiz et Grau, 2011) et Aroma [9] (David, 2007). Quelque soit l'outil utilisé, nous avons, dans la suite, fait l'hypothèse qu'une correspondance est jugée *acceptable* si elle est obtenue avec un degré de confiance supérieur ou égal à *0.5*, seuil qui a été obtenu empiriquement suite à de nombreux tests.

Nous avons retenus les variantes d'ontologies suivantes, en ne retenant pour AGROVOC que les labels en français et en anglais et pour NALT que les labels en anglais :

$$\begin{aligned} V_{\mathrm{naRyQ}} &= \{\mathrm{naRyQ}^{OWL-SKOS}, \mathrm{naRyQ}^{OWL}, \mathrm{naRyQ}^{SKOS}\} \\ V_{\textsc{Agrovoc}} &= \{\textsc{Agrovoc}^{SKOS}, \textsc{Agrovoc}_1^{OWL}, \textsc{Agrovoc}_2^{OWL}, \textsc{Agrovoc}_3^{OWL}\} \\ V_{\textsc{Nalt}} &= \{\textsc{Nalt}^{SKOS}, \textsc{Nalt}^{OWL}\} \end{aligned}$$

Remarque 3.1 *La variante* naRyQ^{OWL} *a été obtenue en gardant la structure de la composante conceptuelle et en transformant les skos :prefLabel et skos :altLabel en rdfs :label. La variante* naRyQ^{SKOS} *a été obtenue en gardant les labels de la composante terminologique et en transformant la hiérarchie conceptuelle en éléments de hiérarchie SKOS.*

Sur les 24 processus d'alignements lancés entre naRyQ et AGROVOC, seuls 9 processus ont produit des résultats d'alignement. Sur les 12 processus d'alignement lancés entre naRyQ et NALT, seuls 4 processus ont produit des résultats d'alignement. Les premiers ensembles générés suite aux lancement des deux outils (cf. équation 2) sont : $\overset{agr}{C}_{naRyQ \rightarrow \textsc{Agrovoc}}$, noté $\overset{agr}{C}_{\textsc{Agrovoc}}$, qui contient 3 196 correspondances, et $\overset{agr}{C}_{naRyQ \rightarrow \textsc{Nalt}}$, noté $\overset{agr}{C}_{\textsc{Nalt}}$, qui contient 1 676 correspondances.

3.4 Résultats expérimentaux

3.4.1 Meilleurs résultats obtenus par les outils d'alignement

Afin de pouvoir évaluer nos résultats, nous présentons dans le tableau 1 les meilleurs résultats obtenus par les deux outils d'alignements sélectionnés sur les différents couples de variantes de l'ontologie source et des ontologies cibles, en nous appuyant sur nos deux alignements de référence partiels $\overline{R}^{+}_{\textsc{Agrovoc}}$ et $\overline{R}^{+}_{\textsc{Nalt}}$. Il est important de noter que ces résultats ne sont donc qu'une approximation, puisque calculé en fonction d'alignements de référence partiels, ce qui peut influencer la précision des résultats. Les valeurs sur une même ligne représentent chacune le meilleur score obtenu par un outil d'alignement pour chaque indicateur (nombre de bonnes correspondances, précision, rappel ou F-mesure). « $\#^*$ » signifie « plus grand nombre de bonnes correspondances » ; « P^* » correspond à « meilleure précision » ; « R^* » correspond à « meilleur rappel » ; « $F\text{-}m^*$ » correspond à « meilleure F-mesure ».

Alignement	$\#^*$	P^*	R^*	$F\text{-}m^*$
naRyQ $\rightarrow$ AGROVOC	300	0.90	0.82	0.85
naRyQ $\rightarrow$ NALT	359	0.80	0.84	0.81

TAB. 1 – *Meilleurs scores d'alignements obtenus avec les deux outils utilisés.*

8. http ://www.cs.ox.ac.uk/isg/projects/LogMap/
9. http ://aroma.gforge.inria.fr/

3.4.2 Évaluation de naRyQ → AGROVOC

Le tableau 2 présente l'évaluation des différents ensembles de correspondances produits durant la phase de raffinage par rapport à l'ensemble d'alignement de référence partiel, $\overline{R}^{+}_{\text{AGROVOC}}$. Sur la dernière ligne, le symbole $\star$ indique, pour l'indicateur de la colonne, un résultat meilleur que celui du meilleur outil pour cet indicateur, en termes de $\overline{R}^{+}_{\text{AGROVOC}}$ (cf. le tableau 1).

Ensemble	# total	# bons	P	R	F-m
$\overset{agr}{C^{*}}_{\text{AGROVOC}}$	1583	366	0.23	0.99	0.37
$\overset{recT}{C}_{\text{AGROVOC}}$	582	354	0.61	0.96	0.74
$\overset{LOD}{C}_{\text{AGROVOC}}$	336	254	0.76	0.69	0.72
U_{AGROVOC}	620	363	0.58	0.99	0.73
$U^{*,2*,3*}_{\text{AGROVOC}}$	447	344*	0.77	0.93*	0.84$^{\simeq}$

TAB. 2 – *Évaluation de naRyQ* → AGROVOC *par rapport à* $\overline{R}^{+}_{\text{AGROVOC}}$.

3.4.3 Évaluation de naRyQ → NALT

Le tableau 3 présente l'évaluation des différents ensembles de correspondances produits durant la phase de raffinage par rapport à l'ensemble d'alignement de référence partiel, $\overline{R}^{+}_{\text{NALT}}$. Sur la dernière ligne, le symbole $\star$ indique, pour l'indicateur de la colonne, un résultat meilleur que celui du meilleur outil pour cet indicateur, en termes de $\overline{R^{+}}_{\text{NALT}}$ (cf. le tableau 1).

Ensemble	# total	# bons	P	R	F-m
$\overset{agr}{C^{*}}_{\text{NALT}}$	850	415	0.49	0.97	0.65
$\overset{recT}{C}_{\text{NALT}}$	480	368	0.77	0.86	0.81
$\overset{LOD}{C}_{\text{NALT}}$	337	255	0.76	0.59	0.67
U_{NALT}	551	404	0.733	0.94	0.82
$U^{*,2*,3*}_{\text{NALT}}$	400	348	0.87*	0.81	0.84*

TAB. 3 – *Évaluation de naRyQ* → NALT *par rapport à* $\overline{R}^{+}_{\text{NALT}}$.

3.4.4 Discussion

Comme nous pouvons l'observer dans les tableaux 2 et 3 et comme nous pouvions nous y attendre, (1) l'augmentation de l'ensemble d'alignements permet d'améliorer les valeurs de rappel (au détriment de la précision) pour la plupart des ensembles produits (meilleur rappel avec $\overset{agr}{C^{*}}$), et, (2) en combinant les différentes méthodes de raffinage, nous obtenons de meilleurs résultats en termes de précision (ensemble $U^{*,2*,3*}$). En comparant ces résultats avec les meilleurs scores obtenus par les outils d'alignement (cf. tableau 1), dans le cas de

AGROVOC, notre approche obtient des performances similaires en termes de F-mesure, tandis que dans le cas de NALT une augmentation de la F-mesure est observée. De plus, notre approche surpasse les meilleurs résultats individuels en rappel pour AGROVOC et en précision pour NALT. Ce qui représente, dans l'ensemble, des résultats très encourageants.

Notre approche d'exploitation des alignements existants sur le web de données pour raffiner des résultats d'alignement est une piste prometteuse. D'autres travaux exploitent également le web de données pour aider à la tâche d'alignement. Pernelle et Sais (2011) proposent une approche qui combine à la fois la découverte de liens entre les données du web de données et l'alignement entre les concepts de deux ontologies. Parundekar et al. (2012) exploitent les liens entre différents sources de données du web de donnés pour aligner deux ontologies.

La plupart des outils d'alignement utilisent des stratégies pour combiner différentes méthodes d'alignement de base (i.e., lexicale, structurale, etc.) au sein d'un processus d'alignement et pour filtrer leurs résultats (seuil, agrégation pondérée, règles, etc.) (cf. Euzenat et Shvaiko (2007)). Nous nous intéressons ici à la fois au raffinage des ensembles d'alignements produits par différents outils d'alignements et à la discrimination de l'ensemble de correspondances trouvées. Dans le premier cas, nous avons identifié trois types d'ambiguïté à résoudre pour raffiner l'ensemble de correspondances en supprimant un certain nombre. Dans le deuxième cas, nous proposons deux méthodes originales pour discriminer l'ensemble des correspondances. La première méthode consiste à considérer la redondance entre des correspondances trouvées dans au moins deux processus d'alignement issus de méthodes distinctes comme gage de validité (i.e. une correspondance apparaissant dans le résultat d'au moins deux processus d'alignement est susceptible d'être correcte). La deuxième méthode consiste à exploiter les alignements définis sur le web de données pour appuyer la validité de certaines correspondances (i.e. les correspondances permettant d'aligner une même entité à deux entités distinctes mais liées sur le web de données sont susceptibles d'être correctes). D'autres travaux (Mochol et Jentzsch, 2008; Steyskal et Polleres, 2013) ont, comme nous, eu l'idée d'utiliser des outils existants et de combiner leurs résultats pour aligner des ontologies. Tandis que Mochol et Jentzsch (2008) utilisent un ensemble de règles pour sélectionner le meilleur outil, Steyskal et Polleres (2013) proposent une méthode itérative basée sur le vote où, à chaque tour, les correspondances acceptées par la majorité des outils sont considérées comme valides. Cependant, ces travaux n'exploitent pas des alignements sur le web de données.

Un autre point original de notre approche réside dans l'exploitation de différentes variantes qui permet de prendre en compte les spécificités des ontologies à lier qui peuvent être des ontologies, des thésaurus, des RTO et qui peuvent être exprimées dans différents langages de représentation avec différents niveaux d'expressivité. Cela nous donne la possibilité de couvrir un panel large et hétérogène de ressources.

4 Conclusion et persectives

Nous avons proposé une méthode originale d'alignement d'ontologies qui permet de relever l'un des défis des alignements d'ontologies cité dans Shvaiko et Euzenat (2013) : « Matching with background knowledge ». Notre méthode permet, d'une part, d'obtenir de nombreuses correspondances en utilisant et combinant des méthodes existantes pour aligner des ontologies, des thésaurus et des RTO exprimés dans différents langages. Elle permet, d'autre part, de discriminer les correspondances obtenues en en supprimant certaines ambiguës, et,

en exploitant la redondance et les alignements existants sur le web de données pour identifier un sous-ensemble de correspondances potentiellement correctes qui pourra être soumis à l'utilisateur pour validation.

Ce travail est un travail préliminaire pour publier des ontologies sur le web de données. L'alignement d'ontologies permet non seulement d'enrichir l'ontologie source avec de nouveaux concepts et/ou termes, mais également de la lier à des ontologies existantes et reconnues sur le web de données afin d'assoir sa place dans le domaine étudié et ciblé. De plus, la liaison des ontologies dans l'esprit du web de données permet de lier les données indexées par ces ontologies et de les rendre disponibles et réutilisables par la communauté du web de données.

Afin d'améliorer notre processus de raffinage des correspondances trouvées, nous envisageons, à court terme, (i) de prendre en compte l'expressivité des variantes des ontologies pour lever les ambiguïtés de type 1 ; (ii) d'étudier d'autres relations dans les ambiguïtés de type 2 où nous n'avons étudié que la relation d'équivalence. Ainsi, au lieu de supprimer les correspondances ambigües de type 2, nous envisageons de proposer une méthodologie permettant de choisir la meilleur correspondance via une approche de raisonnement, par exemple, en supprimant les correspondances introduisant une incohérence logique ; (iii) de lever des ambiguïtés entre des correspondances trouvées à l'aide de relations différentes. Dans ce cas, nous pourrions utiliser une algèbre pour définir une relation combinant celles trouvées. De plus, nous envisageons d'étudier comment utiliser la relation de subsomption pour aider à l'identification des correspondances potentiellement correctes. À long terme, nous envisageons d'exploiter les alignements indirects entre différentes sources d'alignements sur le web de données pour discriminer l'ensemble de correspondances. Nous voudrions également étendre notre approche pour pouvoir prendre en compte des entités plus complexes telles que les unités de mesure et les relations n-aires.

Références

Aguirre, J., K. Eckert, J. Euzenat, A. Ferrara, W. R. van Hage, L. Hollink, C. Meilicke, A. Nikolov, D. Ritze, F. Scharffe, O. S.-Z. Pavel Shvaiko, C. Trojahn, E. Jiménez-Ruiz, B. C. Grau, et B. Zapilko (2012). Results of the ontology alignment evaluation initiative 2012. In *Proc. 7th ISWC workshop on ontology matching (OM)*, pp. 73–115.

Bernstein, P. A., J. Madhavan, et E. Rahm (2011). Generic schema matching, ten years later. *PVLDB 4*(11), 695–701.

Buche, P., S. Dervaux, J. Dibie-Barthelemy, L. Ibanescu, L. Soler, et R. Touhami (2013). Intégration de données hétérogènes et imprecise guide par une resource termino-ontologique. application au domaine des sciences du vivant. *RSTI série Revue d'Intelligence Artificielle 27*(4-5), 539–568.

Caracciolo, C., A. Stellato, S. Rajbhandari, A. Morshed, G. Johannsen, J. Keizer, et Y. Jaques (2012). Thesaurus maintenance, alignment and publication as linked data : the AGROVOC use case. *IJMSO 7*(1), 65–75.

David, J. (2007). *AROMA : une méthode pour la découverte d'alignements orientés entre ontologies à partir de règles d'association*. Ph. D. thesis, Université de Nantes.

David, J., J. Euzenat, F. Scharffe, et C. Trojahn (2011). The alignment api 4.0. *Semantic web 2*(1), 3–10.

Euzenat, J. et P. Shvaiko (2007). *Ontology matching*, Volume 18. Springer Heidelberg.

Heath, T. et C. Bizer (2011). *Linked Data : Evolving the Web into a Global Data Space*, Volume 1. Morgan & Claypool.

Jiménez-Ruiz, E. et B. C. Grau (2011). Logmap : Logic-based and scalable ontology matching. In *The Semantic Web–ISWC 2011*, pp. 273–288. Springer.

McCrae, J., D. Spohr, et P. Cimiano (2011). Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In G. Antoniou, M. Grobelnik, E. P. B. Simperl, B. Parsia, D. Plexousakis, P. D. Leenheer, et J. Z. Pan (Eds.), *ESWC (1)*, Volume 6643 of *Lecture Notes in Computer Science*, pp. 245–259. Springer.

Mochol, M. et A. Jentzsch (2008). Towards a rule-based matcher selection. In A. Gangemi et J. Euzenat (Eds.), *Knowledge Engineering : Practice and Patterns*, Volume 5268 of *Lecture Notes in Computer Science*, pp. 109–119. Springer Berlin Heidelberg.

Parundekar, R., C. A. Knoblock, et J. L. Ambite (2012). Discovering concept coverings in ontologies of linked data sources. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, et E. Blomqvist (Eds.), *International Semantic Web Conference (1)*, Volume 7649 of *Lecture Notes in Computer Science*, pp. 427–443. Springer.

Pernelle, N. et F. Sais (2011). LDM : Link Discovery Method for new Resource Integration. In M.-E. V. Zoé Lacroix, Edna Ruckhaus (Ed.), *Fourth International Workshop on Resource Discovery*, Volume 737, Heraklion, Grèce, pp. 94–108.

Rahm, E. (2011). Towards large-scale schema and ontology matching. In Z. Bellahsene, A. Bonifati, et E. Rahm (Eds.), *Schema Matching and Mapping*, pp. 3–27. Springer.

Reymonet, A., J. Thomas, et N. Aussenac-Gilles (2007). Modelling ontological and terminological resources in OWL DL. In *OntoLex 2007 - Workshop at ISWC07*, Busan, South-Korea.

Roche, C., M. Calberg-Challot, L. Damas, et P. Rouard (2009). Ontoterminology - a new paradigm for terminology. In J. L. G. Dietz (Ed.), *KEOD*, pp. 321–326. INSTICC Press.

Shvaiko, P. et J. Euzenat (2013). Ontology matching : state of the art and future challenges. *IEEE Trans. Knowl. Data Eng. 25*(1), 158–176.

Steyskal, S. et A. Polleres (2013). Mix'n'match : An alternative approach for combining ontology matchers. In R. Meersman, H. Panetto, T. S. Dillon, J. Eder, Z. Bellahsene, N. Ritter, P. D. Leenheer, et D. Dou (Eds.), *OTM Conferences*, Volume 8185 of *Lecture Notes in Computer Science*, pp. 555–563. Springer.

Stoilos, G., G. Stamou, et S. Kollias (2005). A string metric for ontology alignment. In *The Semantic Web–ISWC 2005*, pp. 624–637. Springer.

Summary

In this paper, we propose an alignment method of a source ontology with two target ontologies published and linked on the Web of Data. Then we present some results about the alignment of a source ontology from life sciences with AGROVOC and NALT.