



HAL
open science

Réconciliation d'alignements multilingues dans BioPortal

Amina Annane, Vincent Emonet, Faïçal Azouaou, Clement Jonquet

► **To cite this version:**

Amina Annane, Vincent Emonet, Faïçal Azouaou, Clement Jonquet. Réconciliation d'alignements multilingues dans BioPortal. IC 2016 - 27es Journées francophones d'Ingénierie des Connaissances, Jun 2016, Montpellier, France. lirmm-01395900

HAL Id: lirmm-01395900

<https://hal-lirmm.ccsd.cnrs.fr/lirmm-01395900>

Submitted on 12 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Réconciliation d'alignements multilingues dans BioPortal

Amina Annane¹, Vincent Emonet², Faïçal Azouaou¹ et Clement Jonquet^{2,3}

¹ Ecole nationale Supérieure d'Informatique, Algérie
{a_annane, f_azouaou}@esi.dz

² Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier, France
Université de Montpellier & CNRS
{jonquet, vincent.emonet}@lirmm.fr

³ Center for Biomedical Informatics Research, Stanford University, USA

Résumé : De nos jours, les ontologies sont souvent développées de manière multilingue. Cependant, pour des raisons historiques, dans le domaine biomédical, de nombreuses ontologies ou terminologies ont été traduites d'une langue à une autre ou sont maintenues explicitement dans chaque langue. Cela génère deux ontologies potentiellement alignées mais avec leurs propres spécificités (format, développeurs, versions, etc.). Souvent, il n'existe pas de représentation formelle des liens de traduction reliant les ontologies traduites aux originales et ils ne sont pas accessibles sous forme de *linked data*. Cependant, ces liens sont très importants pour l'interopérabilité et l'intégration de données biomédicales multilingues. Dans cet article, nous présentons les résultats d'une étude de réconciliation des liens de traduction entre ontologies sous forme d'alignements multilingues. Nous avons réconcilié et représenté à l'aide de vocabulaire du web sémantique, plus de 228K mappings entre dix ontologies anglaises hébergées sur le NCBO BioPortal et leurs traductions françaises. Ensuite, nous avons stocké à la fois les ontologies et les mappings sur une version française de la plate-forme, appelée SIFR BioPortal, pour rendre le tout disponible en RDF (données liées). La réconciliation des alignements s'est avérée plus complexe que ce qu'on pourrait penser car les traductions ne sont que rarement l'exacte copie des originales comme nous le discutons.

Mots-clés : Mappings ou alignements multilingues, réconciliations de mappings, web sémantique, données liées, alignement d'ontologies, entrepôt d'ontologies, ontologies biomédicales, BioPortal, interopérabilité et intégration de données, diffusion de données et de connaissances.

1. Introduction

Le domaine biomédical produit de nombreuses ontologies¹. Cependant la majorité de ces ontologies sont en anglais (Névél, Grosjean et al. 2014) et même quand des ontologies sont disponibles dans d'autres langues, il y a un manque d'outils pour les utiliser. L'entrepôt de référence d'ontologies biomédicales NCBO BioPortal (<http://biportal.bioontology.org>) regroupe plus de 433 ontologies, dont uniquement 6 ne sont pas en anglais, 5 en français et une en espagnol (Jonquet, Emonet et al. 2015). Aussi, le méta-thésaurus d'UMLS (Unified Medical Language System) même s'il couvre 21 langues, 75.1% de ses termes sont en anglais et uniquement 1.82% sont en français (Bollegala, Kontonatsios et al. 2015). Cela représente un frein pour les communautés non anglophones qui produisent et manipulent des données biomédicales dans d'autres langues que l'anglais. Cela les empêche d'exploiter ces ontologies pour la recherche, l'annotation et l'indexation sémantique de leurs données, ainsi que pour l'intégration et l'extraction de connaissance à partir de ces données. En effet, lorsque les ressources biomédicales contiennent des éléments textuels, il est important que la langue de ces ressources soit la même que celles des ontologies qui vont permettre de les exploiter sémantiquement, d'où l'intérêt d'avoir des ontologies multilingues ou de traduire les ontologies d'une langue à une autre (ex : traduction du MeSH par l'INSERM) (Deléger, Merkel et al. 2009; Meilicke, García-Castro et al. 2012). Afin d'assurer l'interopérabilité sémantique, ce qui est un des rôles clés des ontologies, il ne suffit pas de les traduire, mais il

¹ Dans notre contexte le mot ontologie englobe les ontologies et les ressources terminologiques. De la même manière nous parleront de mapping ou d'alignement de façon interchangeable.

faut aussi conserver explicitement les liens entre les objets de l'ontologie traduite et ceux de l'ontologie originale (Buitelaar, Cimiano et al. 2009), d'autant plus si les ontologies continuent à évoluer après leur traduction. (Re)Établir ces liens est l'objet de ce travail que nous appelons réconciliation² de mappings multilingues. Nous verrons que même si ce travail ne demande pas du tout la même expertise que la traduction (expertise médicale), ou l'extraction d'alignements (Euzenat and Shvaiko 2013), la tâche devient vite assez difficile si on considère : (i) le nombre d'ontologie et de termes concernés, (ii) leur hétérogénéité, disponibilité, et les façons dont elles ont été traduites et (iii) la variété des champs (e.g., code, identifiant, propriété) qui peuvent être utilisés pour réconcilier les mappings. Ces mappings multilingues, une fois établis et représentés d'une manière formelle, auront de multiples applications (Fu, Brennan et al. 2010). Par exemple, ils peuvent être utilisés pour effectuer une indexation sémantique multilingue de ressources. Ou encore pour intégrer des bases de données biomédicales de langues différentes afin d'étudier des problématiques de recherche telle que la résistance aux maladies, étant donné que les jeux de données spécifiques à une population sont généralement décrits dans la langue de cette population.

Notre travail s'inscrit dans le cadre du projet SIFR (Indexation Sémantique de Ressources biomédicales Francophones – www.lirmm.fr/sifr). Le projet s'intéresse à exploiter les ontologies dans la construction de services d'indexation, de fouille, et de recherche de données pour les ressources biomédicales françaises. Dans ce projet, nous développons un workflow d'indexation basé sur les ontologies (i.e., un annotateur) similaire à celui qui existe pour les ressources anglaises (Jonquet, Shah et al. 2009), mais destiné au français. Pour améliorer le fonctionnement du workflow et relier les ontologies francophones utilisées (hébergées dans une instance locale de BioPortal) à leurs équivalentes anglaises (hébergées sur le NCBO BioPortal ou n'importe où ailleurs), le projet s'intéresse à la réconciliation de mappings multilingues. Cette étude concerne ainsi une dizaine d'ontologies françaises (des traductions) du SIFR BioPortal (<http://bioportal.lirmm.fr/ontologies>) que nous souhaitons aligner formellement avec les ontologies anglaises originales dans le NCBO BioPortal. Nous nous intéressons à cela seulement pour des ontologies monolingues pour lesquelles il n'existe pas de version multilingues ; en parallèle nous étudions comment à l'avenir, gérer le multilinguisme dans BioPortal (Jonquet, Emonet et al. 2015).

Dans la suite de l'article, nous allons discuter brièvement l'état de l'art du domaine d'alignement des ontologies et la représentation de ressources terminologiques multilingues sur le web (section 2). Ensuite, nous allons étudier la gestion des mappings dans BioPortal et les ontologies traitées dans ce travail (section 3 et 4) ; Expliquer la méthodologie suivie (section 5), analyser les résultats obtenus (section 6), et enfin conclure le travail (section 7).

2. Etat de l'art

Le multilinguisme est l'un des problèmes majeurs d'interopérabilité et d'intégration de données et de connaissances. Ce problème est ressenti de plus en plus à cause de l'explosion des données sur le web sémantique (Buitelaar and Cimiano 2014). Afin de le traiter, différents défis ont été identifiés (Gracia 2012), notamment les alignements multilingues d'ontologies et la représentation des informations lexicales dans les ontologies. Dans la littérature, plusieurs approches ont été proposées pour extraire des alignements multilingues. La première approche adoptée est l'approche manuelle. Il s'agit d'extraire les mappings par les experts humains (Liang and Sini 2006). Cette approche donne des mappings de qualité mais elle est fastidieuse pour les ontologies volumineuses. Par conséquent, les chercheurs se sont retournés vers des approches automatiques. Différentes techniques ont été utilisées : machine learning (Spohr, Hollink et al. 2011), machine translation (Fu, Brennan et al. 2012), extraction de mappings à l'aide de ressources multilingues (Background Knowledge) (Tigrine,

² Nous parlons de réconciliation d'alignements car l'information pour formaliser ces alignements existent plus ou moins (comme discuté ci-après) dans les ontologies traduites comparativement aux approches d'extraction de d'alignements qui doivent explicitement détecter des alignements avec des méthodes plus ou moins complexes (structure, syntaxe, etc.).

Bellahsene et al. 2015), etc. Une classification détaillée de ces approches peut être trouvée dans (Trojahn, Fu et al. 2014). Malgré la richesse des travaux dans le domaine de l'alignement d'ontologies, le problème de réconciliation de mappings entre une ontologie et leurs traductions (Euzenat and Shvaiko 2013; Trojahn, Fu et al. 2014) a été ignoré, étant considéré comme une tâche plus facile. Cependant, cette étude montre qu'il n'est pas si facile car les traductions ne suivent pas la même évolution que les ontologies originales ce qui les rend souvent différentes les unes des autres. D'un autre côté et afin de formaliser les liens multilingues entre les ressources terminologiques sur le web (thésaurus, ontologie, etc.), des travaux de recherche ont essayé de définir un modèle capable de représenter les descriptions linguistiques de ces ressources. Le modèle GOLD (General Ontology for Linguistic Description) (Farrar and Langendoen 2003) permet de représenter formellement des concepts linguistiques en utilisant une ontologie owl. Le modèle Lemon (LEXicon Model for ONtologies) (McCrae, Spohr et al. 2011) est le modèle le plus complet actuellement pour la publication de ressources lexicales riches sous forme de données liées. En effet, Lemon est le résultat de l'évolution de plusieurs modèles: LMF (Lexical markup framework), LexInfo (Cimiano, Buitelaar et al. 2011), et Linguistic Information Repository (Montiel-Ponsoda, Aguado de Cea et al. 2008). Lemon permet également de représenter les informations lexicales relatives à une ontologie publiée sur le web sémantique. Il a été étendu récemment pour inclure de nouveaux composants comme celui qui traite les traductions (Gracia, Montiel-Ponsoda et al. 2014). Ce qui a donné le modèle OntoLex/Lemon (Bosque-Gil, Gracia et al. 2015). Il est certainement intéressant de disposer de tels modèles pour représenter tous les détails linguistiques d'une ressource, mais il faut réfléchir à leur mise en œuvre aussi. Les modèles riches tels que Lemon sont difficiles à implémenter notamment pour des ontologies très riches et complexes. Par conséquent, les outils qui implémentent ces modèles et facilitent leurs utilisations sont nécessaires pour qu'ils soient adoptés par la communauté (Gracia, Montiel-Ponsoda et al. 2012).

A l'heure actuelle, le domaine biomédical comporte beaucoup d'ontologies monolingues pour lesquelles une traduction a été produite. Parfois par un autre groupe/projet que celui qui a développé l'originale (e.g., MeSH, MedlinePlus, ICD, MEDDRA, et ICPC). La plupart de ces ontologies sont hébergées dans plusieurs plateformes qui veillent à rassembler les ontologies biomédicales et les intégrer dans le but d'offrir des services à la communauté (chacune selon sa vision et son orientation). La gestion du multilinguisme diffère d'une plateforme à l'autre. Le NCBO BioPortal n'est pas multilingue même s'il héberge les ontologies de différentes langues (Jonquet and Musen 2014). Le Métha-thésaurus UMLS, un ensemble de terminologies intégrées manuellement et distribuées d'une manière publique (sauf quelques exceptions) par la bibliothèque NLM (United States National Library of Medicine) (Bodenreider 2004), contient des terminologies (ainsi que leurs alignements) de langues différentes autres que l'anglais (mais ils représentent une minorité, pour le français 1.82% uniquement). Le portail HeTOP (Grosjean, Merabti et al. 2012) offre aussi des termes biomédicaux traduits dans plusieurs langues, notamment le français, il permet également une recherche multilingue mais la plupart de son contenu n'est pas public ou accessible via des services web. Dans ces deux dernières plateformes, l'approche implémentée consiste à intégrer toutes les ontologies en suivant un méta-modèle commun. Cela signifie qu'il existe une abstraction unique pour les concepts de différentes sources (par exemple, les identifiants uniques de concepts d'UMLS (CUI³)). Ceci est différent de l'approche BioPortal que nous suivons également. Ce dernier ne construit pas un thésaurus global mais garde chaque ontologie séparée et utilise les mappings pour les interconnecter (Ghazvinian, Noy et al. 2009). Une autre différence avec BioPortal, est que ni UMLS ni HeTOP sont construits de manière native avec les technologies du Web sémantique et donc ne proposent pas de représentation sémantique pour rendre les ontologies ou les mappings multilingues disponibles sous format de données liées. En outre, aucune des plateformes citées précédemment n'offre la possibilité de récupérer une ontologie traduite alignée formellement

³ Concept Unique Identifier : un identifiant unique d'un concept au sein du MetaThesaurus UMLS

avec son origine. Cela représente un frein qui empêche l'exploitation de ces ontologies par des applications multilingues.

L'état de l'art fait ainsi ressortir le besoin (au moins pour le français) de réconcilier les mappings multilingues entre les ontologies traduites et leurs origines ainsi que le besoin de rendre ces mappings disponibles et public sous forme de données liées.

3. Représentation de mappings multilingues dans le BioPortal

Choix des propriétés sémantiques : BioPortal utilise une propriété appartenant aux standards du web sémantique pour étiqueter un mapping entre deux concepts⁴. Par exemple la propriété *exactMatch* du vocabulaire SKOS pour indiquer que deux concepts sont identiques. Les mappings sont sauvegardés dans un format spécifique à la plateforme (e.g., URI1 *skos:exactMatch* URI2), mais cet étiquetage permet de les interroger facilement. En effet, tous les mappings (comme n'importe quel contenu dans le portail) sont disponibles soit à travers un SPARQL endpoint ou via une API REST qui renvoie du JSON-LD. Nous proposons de représenter les mappings multilingues dans BioPortal comme n'importe quel mapping, mais avec des propriétés sémantiques supplémentaires. Des propriétés qui marquent l'aspect linguistique et formalisent le lien de traduction entre deux concepts (l'un est la traduction de l'autre). Par exemple, le concept *Mélanome* de la version française de MeSH sur SIFR BioPortal doit être aligné avec le concept *Melanoma* de la version anglaise de MeSH sur le NCBO BioPortal en précisant qu'en plus du fait que *Mélanome* et *Melanoma* ont la même signification, *Mélanome* est la traduction française de *Melanoma*. Nous pouvons continuer à utiliser les propriétés de SKOS pour représenter la correspondance sémantique entre les concepts, mais pour l'aspect linguistique nous avons besoin d'autres propriétés pour décrire la relation de traduction. Nous proposons d'utiliser les propriétés du modèle Lemon ou GOLD (Jonquet, Emonet et al. 2015). Pour ce travail, nous avons choisi d'utiliser les propriétés de GOLD : la propriété *gold:freeTranslation* pour représenter une traduction exacte (quand les deux concepts ont exactement le même sens, quel que soit leurs libellés), et *gold:translation* pour représenter une traduction moins précise.⁵ Ces propriétés répondent à notre besoin (représenter le lien de traduction), sont simples à utiliser, et les ontologies traduites ne contiennent pas les informations liées à leurs traductions qui nous permettrait d'adopter des vocabulaires plus riches. Par exemple, le concept traduit n'a pas une propriété qui permet de classer sa traduction tel que le vocabulaire Lemon le propose avec *directEquivalent*, *culturalEquivalent* ou *lexicalEquivalent*.

Changements dans l'architecture de BioPortal : Pour pouvoir sauvegarder les mappings multilingues, nous avons dû modifier leur représentation dans BioPortal, notamment pour : (1) permettre d'étiqueter le même mapping avec plusieurs propriétés du web sémantique (cette façon de faire évite de dupliquer les mappings au lieu d'un mapping de traduction et un mapping sémantique, l'étiquetage avec deux propriétés va regrouper les deux dans un seul) ; (2) permettre à BioPortal de stocker des alignements vers des ontologies qui sont soit dans une autre instance de BioPortal (inter-portal), soit vers des ontologies externes qui ne sont dans aucune instance de BioPortal (mappings externes).

4. Les ontologies à aligner

Nous avons traité un ensemble de 20 ontologies (voir Table 1); dix en français et dix en anglais. Ce sont des ontologies très utilisées dans le domaine biomédical pour les deux communautés francophone et anglophone. Par exemple, la Classification Internationale des Maladies (CIM-10) est utilisée dans les hôpitaux pour coder les actes médicaux, le Medical Subject Heading (MeSH) est utilisé pour l'indexation de documents (travaux de la NLM et de

⁴http://www.bioontology.org/wiki/index.php/BioPortal_Mappings

⁵ A noter que nous n'utilisons pas la 3^{ème} propriété disponible (*gold:literalTranslation*) qui identifie des traductions mots à mots et qui en conséquence ne permet pas de représenter une relation entre concepts qui peuvent avoir de multiples labels (nom préféré et synonymes). Voir les définitions de GOLD (<http://linguistics-ontology.org/gold>) pour plus de détails.

CISMeF). Chaque ontologie française est couplée avec une ontologie anglaise dont elle est issue. Les ontologies anglaises proviennent toutes du Metathesaurus UMLS (version 2015AA) et ont été importées dans le NCBO BioPortal à l'aide de l'outil umls2rdf développé par le NCBO (<https://github.com/ncbo/umls2rdf>). Les ontologies françaises proviennent soit de l'UMLS (qui contient quelques ontologies directement en français), soit elles ont été produites par le CISMeF du CHU de Rouen qui nous a fourni un fichier OWL pour les inclure dans le SIFR BioPortal. Dans ce deuxième cas, les traductions ont été en général produites ou synthétisées par CISMeF.

TABLE 1 – Les ontologies traitées dans cette étude (les acronymes sont les identifiants respectifs dans le NCBO ou SIFR BioPortal).

N°	Ontologie	Acronyme	Version	Format	Source
01	Systematized Nomenclature of MEDicine	SNMI	2015AA	RDF/TTL	UMLS
	Systematized Nomenclature of MEDicine, version française	SNMIFRE	3.5	OWL	CISMeF
02	International Classification of Functioning, Disability and Health	ICF	1.0.2	OWL	UMLS
	Classification Internationale du Fonctionnement, du handicap et de la santé	CIF	2001	OWL	CISMeF
03	MedlinePlus Health Topics	MEDLINEPL US (EN)	2015AA	RDF/TTL	UMLS
	MEDLINEPLUS FR	MEDLINEPL US (FR)	-	OWL	CISMeF
04	Minimal Standard Terminology of Digestive Endoscopy	MSTDE	2015AA	RDF/TTL	UMLS
	Terminologie minimale standardisée en endoscopie digestive	MTHMSTFRE	2011ab	RDF/TTL	UMLS
05	Semantic Types Ontology	STY (EN)	2015AA	RDF/TTL	UMLS
	Réseau sémantique UMLS	STY (FR)	2014AB	RDF/TTL	CISMeF
06	Medical Subject Headings	MESH	2015AA	RDF/TTL	UMLS
	Medical Subject Headings, version française	MSHFRE	2015AA	RDF/TTL	UMLS
07	Medical Dictionary for Regulatory Activities	MEDDRA	2015AA	RDF/TTL	UMLS
	Dictionnaire médical pour les activités réglementaires en matière de médicaments	MDRFRE	2015AA	RDF/TTL	UMLS
08	World Health Organization (WHO) Adverse Reaction Terminology	WHO-ART	2015AA	RDF/TTL	UMLS
	World Health Organization (WHO) Adverse Reaction Terminology, version française	WHO-ARTFRE	1997	OWL	CISMeF
09	International Classification of Diseases, Version 10	ICD10	2015AA	RDF/TTL	UMLS
	Classification Internationale des Maladies, version 10	CIM-10	10	OWL	CISMeF
10	International Classification of Primary Care - 2 PLUS	ICPC2P	2015AA	RDF/TTL	UMLS
	Classification Internationale de Soins Primaires	CISP-2	1998	OWL	CISMeF

5. Méthodologie

5.1. Récupération des triplets à partir des fichiers des ontologies

Les fichiers qui contiennent les ontologies sont au format OWL ou bien RDF/TTL tels que fournis par les BioPortals. Nous avons utilisé l'API Jena pour extraire les triplets RDF à partir des ontologies (sujet-prédicat-objet). Afin de récupérer uniquement les triplets dont nous avons besoin, nous avons filtré selon une propriété précise. En général, cette propriété a pour valeur le code ou l'identifiant que nous souhaitons utiliser pour réconcilier l'alignement. Pour déterminer la propriété adéquate nous avons dû étudier les fichiers contenant les ontologies un par un. La propriété la plus fréquente est *skos:notation*, utilisée pour 12 des 20 ontologies.

D'autres propriétés ont été utilisées telles que *skos:altLabel* (MEDLINEPLUS (FR)) ou *icd:icdCode* (ICF). Dans les cinq cas où la propriété qui contient le code n'existait pas, nous avons extrait le code à partir des URI des concepts. Pour ces derniers cas et afin de ne récupérer que des classes correspondantes à un concept dans l'ontologie, nous avons filtré et conservé uniquement les classes qui ont une propriété *skos:prefLabel*. La Table 2 résume pour chaque ontologie la propriété utilisée pour avoir le code de mapping avec des exemples. Techniquement, nous avons développé une fonction pour chaque ontologie afin d'extraire le code utilisé pour l'alignement. D'éventuels traitements étaient nécessaires tels que l'élimination du type attaché à la valeur, ou l'utilisation d'expression régulière pour isoler la chaîne de caractère exacte du code.

TABLE 2– RÉCUPÉRATION DES CODES À UTILISER POUR RÉCONCILIER LES MAPPINGS

Ontology	Code source	Example
SNMI MTHMSTFRE MSTDE STY (FR) STY (EN) MDRFRE MEDDRA WHO-ARTICD10 MESH MSHFRE MEDLINEPLUS (EN)	Le code interne est affecté à la propriété <i>skos:notation</i>	Le concept suivant de l'ontologie MSTDE a pour code interne "MT200025". <http://purl.bioontology.org/ontology/MTHMST/MT200025> a owl:Class ; skos:prefLabel Gastric angioectasia (diagnosis)@eng ; skos:notation MT200025 ^^xsd:string ;
ICF	Le code interne est affecté à la propriété <i>icd:icdCode</i>	<owl:Class rdf:ID=s1208> <icd:icdCode rdf:datatype=http://www.w3.org/2001/XMLSchema#string >s1208</icd:icdCode>
SNMIFRE CIF WHO-ARTFRE CIM-10 CISP-2	Le code interne est extrait à partir des URI de concepts. Le filtre de triplets se basait sur la propriété <i>skos:prefLabel</i>	Le code "M-40030" est extrait à partir de l'URI http://churouen.fr/cismef/SNOMED_int/#M-40030 d'un concept de SNMIFRE.
MEDLINEPLUS (FR)	Les concepts n'ont pas de codes internes, nous avons utilisé le CUI pour réconcilier les mappings. Nous avons filter selon la propriété <i>skos:altLabel</i> , ensuite nous avons utilisé une expression régulière pour ne récupérer que les CUI et non pas les labels.	<rdf:Description rdf:about=http://churouen.fr/cismef/MedlinePlus#T351> <skos:prefLabel xml:lang=fr>douleur</skos:prefLabel> <skos:altLabel xml:lang=fr>C0008031</skos:altLabel> <skos:altLabel xml:lang=fr>C0030193</skos:altLabel> <skos:altLabel xml:lang=fr>C0030231</skos:altLabel> </rdf:Description>
ICPC2P	Le code interne est affecté à la propriété <i>icpc2p:icpccode</i>	<http://purl.bioontology.org/ontology/ICPC2P/ICPCCODE> A01 ^^xsd:string ;

5.2. Réconciliation de mappings

A cette étape nous avons stocké les données extraites à partir des fichiers contenant les ontologies dans une base de données relationnelle (une table par ontologie). La table comporte trois champs : (1) ID, un numéro séquentiel qui identifie chaque enregistrement dans la table ; (2) Code, une chaîne de caractère, qui contient le code extrait précédemment et qui peut être un code interne des concepts au niveau de l'ontologie, le CUI ou tout autre critère pertinent de mapping ; (3) URI, une chaîne de caractère pour enregistrer l'URI des concepts de l'ontologie étudiée. Ces URI seront nécessaires pour identifier les concepts dans les BioPortals. Chaque triplet récupéré lors de l'étape précédente entraîne l'insertion d'un nouvel enregistrement dans la table appropriée. Par exemple le triplet (<http://purl.bioontology.org/ontology/MSHFRE/D001542> ; *skos:notation* ; D001542) extrait de la version française de MeSH génère l'enregistrement illustré ci-dessous. Dans la table, il

n'y a pas de contrainte d'unicité pour les deux champs Code et URI. Cela est justifié par le fait que pour un URI donnée il est possible d'avoir plusieurs codes. Nous avons eu ce cas avec la version française de l'ontologie MEDLINEPLUS qui contient 442 concepts ayant plus d'un CUI (identifiant de concept d'UMLS). Par exemple, le concept *minéraux* possède neuf CUI distincts. Nous avons aussi rencontré des cas où le code cible plusieurs URI au sein de la même ontologie (ICPC2P, MEDLINEPLUS FR, CIM-10). Nous allons aborder ces cas plus en détails dans la section 6. Une fois les deux ontologies chargées dans la base de données, nous faisons une jointure sur le champ Code entre les deux tables correspondantes. Comme le code utilisé lors de la jointure n'est pas forcément unique au sein de la même ontologie, le nombre de couples (URI fr, URI en) générés peut être supérieur au nombre de concepts de l'une des deux ontologies (ou les deux). C'est le cas par exemple de l'ontologie CISP2 qui a généré 5063 couples de mappings alors qu'elle n'a que 745 concepts.

TABLE 3—EXEMPLE DE CONTENU STOCKÉ DANS UNE TABLE RELATIONNELLE.

Id	Code	URI
1	D001542	http://purl.bioontology.org/ontology/MSHFRE/D001542

5.3. Chargement des mappings dans le SIFR BioPortal

C'est la dernière étape, elle consiste à représenter les alignements produits lors de l'étape précédente d'une manière formelle et permanente dans SIFR BioPortal. Nos mappings ont été systématiquement qualifiés avec une propriété de traduction de GOLD et une propriété d'alignement de SKOS. Au final nous avons utilisé les quatre combinaisons suivantes : (1) *skos:exactMatch/gold:freeTranslation* : utilisée lorsque le concept traduit est exactement le même dans sa version d'origine. C'est généralement le cas lorsque le mapping se base sur une égalité totale entre le code interne du concept français et le code interne du concept anglais. C'est le cas le plus fréquent. (2) *skos:broadMatch/gold:Translation* : pour décrire le lien qui va d'un concept source plus précis que le concept cible. Par exemple le concept *agression par d'autres moyens précisés/établissement collectif* de CIM-10 ayant pour code **Y08.1** n'a pas de concept anglais dans ICD10 avec le même code. Cependant, nous pouvons l'aligner avec le concept *Assault by other specified means* qui a le code **Y08**. (3) *skos:narrowMatch/gold:Translation* : Utilisée pour le cas contraire, c'est lorsque le concept cible est plus précis que le concept source, par exemple le mapping entre CISP2/ICPC2P via le code interne (le code ICPC). Un concept de la version française est aligné avec plusieurs concepts de la version anglaise qui ont le même code ICPC mais différenciés à l'aide d'un suffixe (à cause d'une spécialisation de l'ontologie anglaise après la traduction). Par exemple, le concept tumeur bénigne ayant le code B75 est mappé avec huit concepts anglais plus précis, e.g., (*benign neoplasm of the blood*, **B75001**), (*benign neoplasm of the lymphatics*, **B75002**), etc. (4) *skos:closeMatch/gold:Translation* : En absence d'un identifiant interne de concepts, nous étions obligés d'utiliser des identifiants moins précis tels que le CUI pour les ontologies qui viennent d'UMLS. Les CUI sont des identifiants au niveau du méta-thésaurus, et non pas au niveau des ontologies sources. Donc, il ne s'agit pas d'une traduction exacte du concept d'une langue à une autre mais plutôt des concepts qui signifient la même chose étant donné qu'on leur a affecté le même CUI.

Tous les alignements produits ont été enregistrés sur le SIFR BioPortal à l'aide d'un script qui utilise l'API web service du SIFR BioPortal (<http://data.biportal.lirmm.fr/documentation>). Désormais, pour toutes les ontologies traitées dans ce travail, nous pouvons consulter pour un concept donné, ses mappings multilingues. Le lien ainsi généré permet à l'utilisateur de passer directement du SIFR BioPortal au concept cible dans le NCBO BioPortal afin d'effectuer certains traitements tel que la recherche de ressources anglaises indexées avec ce concept. En plus de l'interface graphique, ces mappings multilingues sont également disponibles via une API web service et un SPARQL endpoint (<http://sparql.biportal.lirmm.fr/test/>) ce qui en fait des éléments à part entière du Web de

données et qui les rend facilement réutilisables et exploitables par des applications tierces (via l'utilisation ou non des technologies du web sémantique).

6. Résultats

Notre objectif étant de fournir des alignements pour les versions françaises d'ontologies. Nous exprimerons nos résultats sous forme de pourcentage de concepts de l'ontologie française pour lesquels nous avons pu fournir au moins un mapping de traduction.

Les trois couples d'ontologies (STY FR ; STY EN), (MDRFRE ; MEDDRA), (CIF ; ICF) ont été alignés parfaitement (un pourcentage de 100%) grâce au code interne des concepts. Dans les paragraphes suivants, nous revenons sur l'ensemble des autres couples traités.

MSHFRE/MeSH : Le nombre de concepts de la version anglaise (252242 concept) est dix fois plus grand que le nombre de concepts dans la version française (26142 concept) car cette dernière ne contient que les descripteurs MeSH et pas les concepts supplémentaires⁶. Notre alignement couvre presque la totalité des concepts français avec un pourcentage de 99.79%. Seuls 55 concepts de la version française n'ont pas été alignés car leurs codes n'existent pas dans la version anglaise. De plus, même en essayant de les mapper à l'aide de CUI, nous n'avons pas retrouvé leurs CUI dans le MeSH anglais. Dans le cas d'utilisation de CUI, nous estimons qu'il s'agit sans doute d'erreurs commises par les traducteurs ou des problèmes apparus lors de l'intégration de la nouvelle traduction dans UMLS. En effet, les versions de MeSH devraient être parfaitement alignées car elles viennent les deux d'UMLS. Nous envisageons de contacter l'INSERM et la NLM pour leur indiquer ces résultats.

MTHMSTFRE/MSTDE : Parmi les 1700 concepts de la version française, uniquement deux concepts n'ont pas été alignés car leurs codes n'existent pas dans la version anglaise. Nous avons trouvé un seul concept non mappé dans la version anglaise, il a pour code la valeur *NOCODE*. Cependant ce concept possède deux CUI, qui sont ceux affectés aux concepts français non mappés. Ainsi, pour ces deux concepts français le mapping était par CUI, ce qui a permis d'obtenir 100% d'alignement. Nous estimons que ce cas représente une erreur commise lors de l'intégration de MSTDE dans UMLS, car en principe chaque classe devrait avoir un code. C'est d'ailleurs le cas de la version française MTHMSTFRE.

WHO-ARTFRE/WHO : Le code interne des concepts n'est pas renseigné à travers une propriété dans le fichier contenant l'ontologie française, nous avons dû l'extraire à partir des URI de concepts. Dans la version anglaise WHO-ART, ce code est bien renseigné à l'aide de la propriété *skos:notation*. Nous avons constaté que la version française a subi une modification. En effet, un code de la version anglaise peut référencer plusieurs sous concepts français qui ont le même code de base avec des suffixes différents. Par exemple, le concept anglais qui a le code **1723** référence quatre concepts français ayant pour code : **1723-IT0**, **1723-IT1**, **1723-IT2** et **1723-PT**. Pour cela, le nombre de concepts français est supérieur au nombre de concepts anglais (3320 vs 1724). La version française est plus détaillée, ses concepts sont plus précis que ceux de la version anglaise, par conséquent nous avons utilisé les relations *skos:broadMatch/gold:translation* pour décrire ces mappings.

MEDLINEPLUS FR/MEDLINEPLUS EN : En raison de l'absence d'un code interne qui distingue les concepts, nous avons utilisé la propriété CUI pour l'alignement de ces deux ontologies. La version française de MedlinePlus contient 795 concepts. Chaque concept a pour propriété un ou plusieurs CUI (442 concepts en ont plus d'un), ce qui a donné 1686 couples (concept, CUI) distincts. La version anglaise contient 1986 concepts distincts, et chaque concept n'a pour propriété qu'un seul CUI. Il est surprenant de noter que des concepts de la version française ont pour propriété des CUI qui n'appartiennent pas à la version anglaise. 123 concepts parmi ces 147 ont pour propriété d'autre CUI appartenant à la version anglaise mais les 24 qui restent n'ont aucun CUI appartenant à la version anglaise (e.g., C0021311, C2362506, etc.). Ces concepts n'existent donc pas ou plus dans la version

⁶ Voir <http://mesh.inserm.fr/mesh> pour des éléments sur la traduction de MeSH.

anglaise. Par conséquent, 24 concepts français n'ont pas été alignés et nous avons obtenu un pourcentage de 97% de concepts français alignés. En essayant de raffiner l'étude, nous avons pris huit de ces concepts qui n'apparaissent pas dans la version anglaise, et nous leur avons appliqué le traitement suivant : (1) Chercher le terme préféré de concept dans l'ontologie française ; (2) Traduire le terme français en anglais manuellement, en utilisant un portail terminologique TermSciences (www.termosciences.fr) ou une autre ressource lexicale (e.g., BabelNet ou même Google translation); (3) Chercher dans la version anglaise, le concept anglais qui a le terme anglais trouvé précédemment comme label et si le concept anglais existe, noter son CUI. Dans 7 cas sur 8, nous avons trouvé le concept anglais qui correspond au concept français (l'origine du concept français) mais avec un CUI différent, comme illustré dans le tableau ci-dessous. Ces résultats nous font penser que ces 24 concepts non mappés sont dû à des erreurs dans le choix des identifiants (CUI) lors du processus de traduction. Nous envisageons de communiquer ces concepts aux traducteurs pour qu'ils détectent les erreurs possibles et éventuellement mettent à jour leur traduction.

TABLE 4 – RÉSULTAT DU TRAITEMENT EFFECTUÉ SUR 8 CONCEPTS

Cui	PrefLab	Cui	PrefLabel
C0156543	Avortement	C0392535	Abortion
C2362506	Fitness et exercice	C1456706	Fitness and Exercise
C0021311	Infections	C3714514	Infections
C1456593	santé mentale et comportement	C1832070	mental health and behavior
C1456620	vivre avec le SIDA	C2963182	Living with HIV/AIDS
C1456571	nutrition des nourrissons et des bébés	/	"nutrition of infants and babies" non trouvé
C2362562	sécurité du patient	C1113679	patient safety
C0002808	Anatomie	C0700276	Anatomy

CISP2/ICPC2P : La version française contient 745 concepts alors que le nombre mapping est de 5141. Cela s'explique par le fait que la version anglaise a subi une modification. Un concept avec un code ICPCCODE a été spécialisé pour générer plusieurs concepts fils (spécialisation des concepts). Par exemple, pour le code **A01**, dans la version française on trouve un seul concept *douleur générale/de sites multiples* alors que dans la version anglaise ce code référence quatre concepts plus précis (**A01001**; generalised aches), (**A01004**; body pain), etc. Par conséquent, un seul concept de CISP2 génère autant de mappings qu'il y'en a des concepts anglais ayant le même code. Pour cette raison, nous avons étiqueté ces mappings avec les propriétés *skos:narrowMatch/gold:translation*. 59 concepts de l'ontologie française n'ont pas été alignés grâce au code ICPC, en étudiant ces concepts de près, nous avons constaté qu'ils n'ont pas un code ICPC comme le reste des concepts. De plus, ces concepts n'ont aucun CUI comme propriété également. Il semblerait qu'ils ont été ajoutés lors de la traduction, ou supprimés de la version anglaise.

CIM-10/ICD10 : La CIM-10 contient 19853 concepts tandis que sa version anglaise, ICD10, contient 12318 concepts. Pour ce cas aussi nous avons remarqué que la version française a été modifiée. Elle a été enrichie avec de nouveaux concepts résultants d'une spécialisation des concepts originaux. Une jointure selon le code interne des concepts entre les deux ontologies a généré un pourcentage de mapping de 62% (12308 concepts ont été mappés). Nous avons également constaté qu'il existe 6 chapitres dans la version française n'ont pas les mêmes identifiants que leurs correspondants anglais. Par exemple, dans CIM-10 le code d'un chapitre est (**B99**) alors que dans ICD10 le code est **B99-B99.8**. Ces chapitres ont la particularité de ne contenir qu'une seule entrée. Nous avons dû les traiter manuellement vu que la jointure selon le code ne les prenne pas. Tous les mappings discutés précédemment ont été étiquetés avec les propriétés *skos:exactMatch/gold:freeTranslation*. Quant aux concepts générés par une spécialisation (leurs codes n'existent pas dans la version anglaises), nous avons extrait le code de leurs unique direct concept père (les trois premiers digits de leurs codes) et nous les avons mappés avec leurs concepts père anglais. Ces derniers mappings ont été étiquetés avec les propriétés *skos:broadMatch/gold:translation*. Par

exemple, le concept français (*Agression par d'autres moyens précisés /domicile* ; **Y08.0**) a été mappés avec le concept anglais (*Assault by other specified means* ; **Y08**). En suivant ce processus nous avons réduit le nombre de concepts non mappés de 7545 à 40.

SNMIFRE/SNMI : La version française SNMIFRE comporte 106266 concept français, tandis que la version anglaise, contient 109150 concepts, ce qui présente un écart de 2884 concepts en plus dans la version anglaise. Sur la base du code interne, 102093 concepts français ont été mappés, soit 96% de l'ontologie française. Cependant il restait 4173 concepts de la version française sans mapping. En outre, nous n'avons pas pu utiliser non plus la propriété CUI pour ces concepts non mappés, car ils font partie d'un ensemble de 9510 concepts de SNMIFRE qui ne l'ont pas, tandis que les concepts de la version anglaise ont tous cette propriété. Nous n'avons pas trouvé d'autres solutions pour aligner ces 4173 concepts restants.

TABLE 5– RÉCAPITULATIF DES RÉSULTATS OBTENUS

ontologie française	nombre de concepts	ontologie anglaise	nombre de concepts	Nombre de concepts mappés	% de concepts mappés	Nombre de Mappings générés	Propriétés (skos ; gold)
MSHFRE	26142	MeSH	252242	26220	99.79%	26220	exactMatch ; freeTranslation
MTHMSTFRE	1700	MSTDE	1699	1700	100%	1700	exactMatch ; freeTranslation
STY	133	STY	133	133	100%	133	exactMatch ; freeTranslation
MDRFRE	66378	MEDDRA	66378	66378	100%	66378	exactMatch ; freeTranslation
MEDLINEPLUS	795	MEDLINEPLUS	2113	771	97%	1520	closeMatch ; translation
CIF	1495	ICF	1495	1495	100%	1495	exactMatch ; freeTranslation
WHO-ARTFRE	3482	WHO	1724	3482	100%	3482	broadMatch ; translation
CISP2	745	ICPC2P	7537	665	70%	5063	narrowMatch ; translation
CIM-10	19853	ICD10	12318	19813	99%	19813	exactMatch ; freeTranslation 62% broadMatch ; translation 37%
SNMIFRE	106266	SNMI	109150	102093	96%	102093	exactMatch ; freeTranslation

7. Conclusion

Dans ce travail nous avons proposé une approche pour la représentation formelle des liens sémantiques reliant une ontologie traduite à son origine. Notre approche consiste à représenter ces liens sous forme de mappings multilingues à l'aide des propriétés sémantiques. Cependant, ce travail ne doit pas être confondu avec l'extraction de mappings multilingues. En effet, nous nous sommes basés dans la majorité des cas sur les codes internes pour réconcilier les liens entre les concepts français et les concepts anglais. Malgré les difficultés que nous avons rencontrées dans certains cas, le lien sémantique entre le concept traduit et son origine existait à travers le code interne des concepts. Notre mission était de le rétablir et de le représenter d'une manière formelle, afin de le rendre disponible pour la communauté. Alors que l'extraction de mappings multilingues consiste à aligner deux ontologies différentes, qui n'ont aucun lien entre elles et qui ne sont pas de la même langue. Cependant, notre approche pour représenter et sauvegarder les mappings peut être utilisée pour représenter les deux types de mappings : réconciliés ou extraits à condition de choisir les propriétés sémantiques adéquates. Dans notre cas, nous avons choisi les propriétés de SKOS et GOLD. Ces propriétés sont complémentaires; en effet, la propriété *gold:translation* ne représente pas la différence entre une relation de traduction *narrow* (vers un concept plus précis que l'autre), une traduction *broad* (vers un concept plus général que l'autre) ou une

traduction *close* (forte relation de similarité) (ces types de traduction sont également identifiés par l'étude de Chen et Chen (S. jiu and H. hua 2012.)), mais en la combinant avec les propriétés de SKOS nous obtenons la description exacte de la relation qui lie les deux concepts. Par exemple, le couple de propriété *skos:narrowMatch/gold:translation* décrit une traduction de type *narrow*. Un aspect important dans le domaine d'alignement d'ontologies est l'évaluation des mappings résultants (Euzenat and Shvaiko 2013; Trojahn, Fu et al. 2014). Cependant, dans notre cas il s'agit d'une réconciliation de mappings entre concepts basée sur les codes internes de concepts (les codes trouvés à la fois dans l'ontologie originale et l'ontologie traduite) et non pas sur des mesures de similarité (Shvaiko and Euzenat 2013). Par conséquent les mappings obtenus sont systématiquement corrects. Sauf les mappings de MedlinePlus (1% des mappings produits) déduits grâce à la propriété CUI. Vu l'affectation multiples de CUI aux concepts français (section 6), ces mappings par contre doivent être vérifiés. Ainsi que l'ensemble d'anomalies constatées dans certains couples d'ontologies. Nous prévoyons communiquer les résultats aux traducteurs afin de les revoir et les rectifier.

Il est aussi, important de noter que quel que soit la richesse d'une ontologie en termes de langues (2,3 ou même 10), elle ne couvrira jamais toutes les langues. La traduction d'ontologie restera une solution inévitable pour pouvoir exploiter une ontologie dans d'autres langues, autres que celles supportées en natif. Nous espérons que cette étude va sensibiliser les traducteurs concernant l'utilisation des mêmes identifiants au moment de la traduction. Nous pensons que l'idéal est d'utiliser les principes du web sémantique, notamment la réutilisation du même URI comme identifiant unique d'un concept au lieu de créer un nouveau ou l'utilisation d'un alignement *owl:sameAs* qui garantit que les deux concepts sont les mêmes et qu'ils portent en plus la même logique. De plus, dans le processus de création des ontologies multilingues, le défi est d'adopter un standard lexical tel que Lemon au lieu de se contenter de la simple utilisation de la propriété *xmllang* pour spécifier la langue des labels.

Les mappings multilingues produits dans cette étude peuvent avoir de multiples applications, notamment l'intégration de données biomédicales de langues différentes, la recherche et l'indexation sémantiques multilingues. De plus, et dans la continuité du projet SIFR, ces mappings vont être intégrés dans la version française de l'annotateur sémantique (Jonquet, Shah et al. 2009) qui va étendre les annotations directes extraites à partir d'ontologies françaises avec (i) leurs concepts anglais correspondants, (ii) autres ontologies anglaises alignées l'une à l'autre dans le NCBO BioPortal. Nos mappings peuvent aussi être utilisés comme un corpus pour développer des outils de traduction automatique (ou semi-automatique) d'ontologies biomédicales, ce qui peut s'avérer très utile pour les traducteurs d'ontologies. Ils peuvent également servir de *background knowledge* à des approches d'extraction d'alignements entre ontologies différentes. C'est d'ailleurs une perspective que nous regardons actuellement.

Bien entendu, les ontologies évoluent et change à travers le temps (y inclus leur traduction) d'où la nécessité de mettre en œuvre une politique de mise à jour de nos réconciliations (Hartung, Kirsten et al. 2008). Actuellement, nous exécutons le script à nouveau quand une nouvelle version des ontologies traitées est chargée dans le portail, nous supprimons tous les anciens mappings multilingues pour sauvegarder les nouveaux. Ce traitement pourrait être automatisé si le script de réconciliation de mappings multilingues était intégré dans le SIFR BioPortal.

Références

- BODENREIDER, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32(suppl 1): D267-D270.
- BOLLEGALA, D., G. KONTONATSIOS, ET AL. (2015). A Cross-lingual Similarity Measure for Detecting Biomedical Term Translations. *PLoS ONE* 10(6).
- BOSQUE-GIL, J., J. GRACIA, ET AL. (2015). Applying the OntoLex Model to a Multilingual Terminological Resource. *The Semantic Web: ESWC 2015 Satellite Events*. F. Gandon, C. Guéret, S. Villata et al, Springer International Publishing. 9341: 283-294.

- BUITELAAR, P. AND P. CIMIANO (2014). *Towards the Multilingual Semantic Web*, Springer.
- BUITELAAR, P., P. CIMIANO, ET AL. (2009). Towards linguistically grounded ontologies. *The semantic web: research and applications*, Springer: 111-125.
- CIMIANO, P., P. BUITELAAR, ET AL. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web* 9(1): 29-51.
- DELEGER, L., M. MERKEL, ET AL. (2009). Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics* 42(4): 692-701.
- EUZENAT, J. AND P. SHVAIKO (2013). *Ontology Matching*, Springer Berlin Heidelberg.
- FARRAR, S. AND D. T. LANGENDOEN (2003). A linguistic ontology for the semantic web. *Glott International* 7(3): 97-100.
- FU, B., R. BRENNAN, ET AL. (2010). *Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web*. Multilingual Semantic Web, Raleigh, North Carolina, USA.
- FU, B., R. BRENNAN, ET AL. (2012). A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. *Web Semantics: Science, Services and Agents on the World Wide Web* 15: 15-36.
- GHAZVINIAN, A., N. NOY, ET AL. (2009). What Four Million Mappings Can Tell You about Two Hundred Ontologies. 8th International Semantic Web Conference, ISWC'09, Washington DC, USA, Springer.
- GRACIA, J. (2012). Cross-lingual ontology matching as a challenge for the Multilingual Semantic Web. *Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik*.
- GRACIA, J., E. MONTIEL-PONSODA, ET AL. (2012). Challenges for the multilingual Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web* 11: 63-71.
- GRACIA, J., E. MONTIEL-PONSODA, et al. (2014). Enabling language resources to expose translations as linked data on the web. *Proc. of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik (Iceland), Reykjavik, Iceland, European Language Resources Association.
- GROSJEAN, J., T. MERABTI, ET AL. (2012). Multi-terminology cross-lingual model to create the Health Terminology/Ontology Portal. *American Medical Informatics Association Annual Symposium*.
- HARTUNG, M., T. KIRSTEN, ET AL. (2008). *Analyzing the evolution of life science ontologies and mappings*. *Data Integration in the Life Sciences*, Springer.
- JONQUET, C., V. EMONET, ET AL. (2015). Roadmap for a multilingual BioPortal. *MSW4'15: 4th Workshop on the Multilingual Semantic Web*, Portoroz, Slovenia.
- JONQUET, C. AND M. MUSEN, A. (2014). Gestion du multilinguisme dans un portail d'ontologies: étude de cas pour le NCBO BioPortal. *TOTh'14: Terminology and Ontology : Theories and applications Workshop*, Bruxelles, Belgium.
- JONQUET, C., N. H. SHAH, ET AL. (2009). The Open Biomedical Annotator *American Medical Informatics Association Symposium on Translational Bioinformatics*, AMIA-TBI'09 56-60.
- LIANG, A. C. AND M. SINI (2006). Mapping AGROVOC and the Chinese Agricultural Thesaurus: definitions, tools, procedures. *New Review of Hypermedia and Multimedia* 12(1): 51-62.
- MCCRAE, J., D. SPOHR, ET AL. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *The semantic web: research and applications*, Springer: 245-259.
- MEILICKE, C., R. GARCIA-CASTRO, ET AL. (2012). MultiFarm: A benchmark for multilingual ontology matching. *Web Semantics: Science, Services and Agents on the World Wide Web* 15: 62-68.
- MONTIEL-PONSODA, E., G. AGUADO DE CEA, ET AL. (2008). Modelling multilinguality in ontologies.
- NEVEOL, A., J. GROSJEAN, ET AL. (2014). Language resources for French in the biomedical domain. 9th International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland.
- S. JIUN, C. AND C. H. HUA (2012.). Mapping multilingual lexical semantics for knowledge organization systems. *The Electronic Library* 30(2): 278-294.
- SHVAIKO, P. AND J. EUZENAT (2013). *Ontology Matching: State of the Art and Future Challenges*. *Knowledge and Data Engineering*, IEEE Transactions on 25(1): 158-176.
- SPOHR, D., L. HOLLINK, ET AL. (2011). A machine learning approach to multilingual and cross-lingual ontology matching. *The Semantic Web—ISWC 2011*, Springer: 665-680.
- TIGRINE, A., Z. BELLAHSENE, ET AL. (2015). Light-Weight Cross-Lingual Ontology Matching with LYAM++. *On the Move to Meaningful Internet Systems: OTM 2015 Conferences*. C. Debruyne, H. Panetto, R. Meersman et al, Springer International Publishing. 9415: 527-544.
- TROJAHN, C., B. FU, ET AL. (2014). state-of-the-art in multilingual and cross-lingual ontology matching. towards the multilingual semantic web. p. buitelaar and p. cimiano, springer berlin heidelberg: 119-135.