Interopérabilité sémantique entre vocabulaires contrôlés

Évaluation de la qualité des alignements sur des données de standards du diagnostic *in vitro*

Mélissa Mary[1,2], **Lina F. Soualmia**[2,3], **Xavier Gansel**[1]

1. bioMérieux SA, Département développement et intégration
3 route de Port Michaud, 38390 La Balme Les Grottes, France
{melissa.mary, xavier.gansel}@biomerieux.com

2. LITIS EA 4108 et NormaSTIC CNRS 3638,
Normandie Université, Université de Rouen de Normandie, 76000 Rouen, France
Lina.Soualmia@chu-rouen.fr

3. LIMICS INSERM UMR_1142
Sorbonne Universités, 75000 Paris, France

RÉSUMÉ. L'informatisation des données de santé doit relever le défi de l'interopérabilité syntaxique, mais surtout sémantique, entre les systèmes d'information ainsi que et les systèmes d'organisation des connaissances (SOC) sur lesquels ils reposent. L'intégration des connaissances entre SOC est une problématique largement étudiée notamment dans des domaines de la biologie ou la santé. Le diagnostic in vitro *(DIV) à l'interface de la médecine et de la biologie, doit répondre aux mêmes problématiques d'interopérabilité avec des outils adaptés à la nature multi et transdisciplinaire de ses données. Dans cet article, nous proposons une revue de littérature des algorithmes existants pour le liage des données. À partir de cette revue nous proposons une évaluation d'alignement de concepts issus du DIV présents dans les SOC de référence disponibles en ligne. Les méthodes que nous proposons reposent sur trois mesures de similarité syntaxique et un algorithme heuristique. Les résultats que nous obtenons dans cette étude montrent que les métriques de similarité syntaxique ne se révèlent pas suffisamment probantes pour se voir appliquer de manière systématique au domaine des tests de laboratoire. En revanche, la qualité des alignements obtenus* via *l'algorithme heuristique, filtré* a posteriori *en fonction d'une dimension sémantique, permet de conforter les critères de performance que nous avions établis. Cet algorithme est notre piste privilégiée pour obtenir des alignements de qualité dans le domaine du DIV. Une seconde version de cet algorithme, en cours de développement, intègre l'ensemble des métriques syntaxiques et sémantiques étudiées dans l'article.*

ABSTRACT. Medical data numerization raises syntactic but also semantic interoperability challenges between information systems and knowledge organisation systems. Knowledge integration was largely studied into general purposes as in specific domain such as clinical and biology. As in vitro diagnostics is transdisciplinary domain it should answer to the same

knowledge integration issues, which are encountered in clinical and biological field, using tools adapted to its multidisciplinary knowledge. In this article we propose a literature review about knowledge integration and linked data state of art with a specific focused on IVD data. We present an evaluation of concepts alignment extracted from two standards used in DIV and available on line. Methods we propose are based on three lexical semantic similarity measures and one heuristic algorithm. Results we obtained illustrates that lexical measures are not enough efficient to be used into laboratory domain. However, alignments obtained with the heuristic approach and filtered with a semantic dimension comply with our performance criteria. This strategy is under improvement process by the integration of semantic similarity and the refinement of lexical parameter into the heuristic approach.

MOTS-CLÉS : intégration de données, alignement d'ontologies, terminologies biomédicales.

KEYWORDS: data integration, ontology alignment, biomedical terminology.

DOI:10.3166/ISI.21.5-6.55-83 © 2016 Lavoisier

1. Introduction

La centralisation des données du patient dans un répertoire électronique est gérée par les instituts de santé publique afin d'améliorer sa prise en charge et de maîtriser les coûts médicaux (Fieschi, 2009 ; Macary, 2007 ; Stroetmann, 2009). L'objectif de ces répertoires uniques est de rendre ces données accessibles et éditables par l'ensemble des acteurs de la chaîne de soins. De ce fait, l'interopérabilité entre les différents systèmes d'information utilisés par les professionnels de santé est un enjeu majeur dans la mise en place de ces dossiers électroniques. L'utilisation de vocabulaires standards est une des solutions pour résoudre la problématique d'interopérabilité au sein de ces dossiers. Ces vocabulaires standard sont choisis pour représenter un domaine particulier de connaissance et peuvent prendre la forme de terminologie, classification ou ontologie que nous désignons sous le terme général de système d'organisation des connaissances (SOC) (Hodge, 2000). Les dossiers médicaux électroniques utilisent un véritable écosystème de SOC, représentant des informations complémentaires qu'il faut rendre interopérables pour fluidifier l'accès et la réutilisation des données patients. C'est notamment le cas dans le domaine du diagnostic *in vitro* (DIV) dont l'information principale est portée par le couple (*test de laboratoire réalisé, et résultats obtenus*). Les instances de standardisation nationales et internationales (Blumenthal, 2010 ; Stroetmann, 2009) recommandent l'utilisation de deux SOC : la terminologie LOINC®[1] (*Logical Observation Identifiers Names and Codes*) pour décrire les tests de laboratoire, et l'ontologie SNOMED CT®[2] (*Systematized Nomenclature of MEDicine – Clinical Terms*) pour exprimer le résultat. Les informations exprimées par la terminologie LOINC® et l'ontologie SNOMED CT® dans les dossiers patients sont interdépendantes, ce qui conduit à l'utilisation conjointe de ces SOC pour interroger

1. http://loinc.org/

2. http://www.ihtsdo.org/snomed-ct

et agréger les données d'un compte rendu d'analyse. Une collaboration a récemment été mise en place (IHTSDO et Regenstrief Institute, 2013) afin d'aligner LOINC® et SNOMED CT® et ainsi améliorer l'interopérabilité des données de comptes rendus de laboratoire (Vreeman, 2015).

De plus en plus d'initiatives visent à réaliser des alignements entre SOC dans le domaine clinique, mais également entre des SOC disponibles sur le web. L'obtention des alignements est un défi majeur, tant du fait du volume des données à aligner, de l'hétérogénéité de la représentation des concepts, que de la qualité attendue des alignements. De nombreuses méthodes, stratégies et métriques ont été développées permettant de réaliser des alignements notamment entre ontologies (Bellahsene *et al.*, 2011 ; Brahma et Refoufi, 2015 ; Euzenat *et al.*, 2007 ; Shvaiko et Euzenat, 2005) ou encore gérer l'évolution des alignements entre concepts (Dos Reis *et al.*, 2015). Dans cet article, nous proposons et évaluons trois métriques de similarité et un algorithme heuristique appliqués sur des données de DIV. En tenant compte des caractéristiques des SOC et l'emploi de ces alignements, nous établissons plusieurs critères d'évaluation pour discriminer ces méthodes. Les résultats ont été comparés à un alignement partiel entre LOINC® et SNOMED CT® réalisé par des experts et disponible sur le web[3].

La suite de cet article est organisée comme suit. La section 2 décrit un état de l'art sur la problématique d'intégration et liage de données dans le domaine du DIV. La section 3 est dédiée à la présentation des ressources utilisées et des méthodes que nous avons évaluées dans la section 4. La section 5 conclut cette étude et présente plusieurs pistes de recherche pour de futurs travaux.

2. Revue de littérature

Cette partie s'attache à présenter les principaux apports de la littérature pour résoudre les problématiques d'intégration de données du diagnostic *in vitro*. La première section introduit les caractéristiques générales des données du DIV. Nous détaillons dans la deuxième section la problématique d'hétérogénéité et présentons des solutions d'intégration de données en biologie et médecine. Dans la troisième section nous présentons une classification des méthodes d'alignements. Nous détaillons les métriques permettant de comparer deux termes (section 2.4) et les stratégies d'alignement permettant d'identifier les correspondances entre deux SOC (2.5).

3. http://www.loinc.org/news/alpha-phase-3-edition-of-draft-loinc-snomed-ct-mappings-and-expression-associations-now-available.html/

2.1. Caractéristiques des données du diagnostic in vitro

Le diagnostic *in vitro* est une discipline à l'interface entre le domaine de la biologie et de la médecine. Il hérite des problématiques et contraintes liées aux domaines cliniques et biologiques qui sont exacerbées de par sa nature transdisciplinaire. Les SOC représentant des données de DIV ont trois caractéristiques qui impactent directement les critères de sélection d'algorithmes et métriques pour réaliser une intégration de données.

Les SOC représentant les données du DIV ont un impact direct sur la qualité des données patients et suivent les normes préconisées par les instances de santé publique (*i.e.* ISO13450 relative à l'industrie pharmacologique). Les méthodes utilisées pour aligner les SOC doivent permettre l'obtention d'alignements de qualité, afin que ceux-ci soient réutilisables dans un contexte médical. Nous avons pris en compte le critère de qualité dans l'évaluation des méthodes en préférant analyser la précision et le rappel des alignements comme deux paramètres distincts.

Les SOC du DIV se caractérisent par un gros volume de données ainsi que des données très évolutives. LOINC® par exemple compte 79 000 termes et SNOMED CT® plus de 350 000 concepts. L'évolution des connaissances médicales et biologiques a un impact dans la mise à jour des SOC représentant des données du DIV. LOINC® et SNOMED CT® par exemple sont mis à jour deux fois par an, ce qui entraîne pour chaque mise à jour de SOC la recompilation d'un nouvel alignement. Ce paramètre nous incite à considérer des méthodes d'alignement performantes en temps de calcul qui sont détaillées en section 2.5.

La dernière caractéristique concerne l'hétérogénéité des données au sein du DIV qui est due à la fois 1) à un aspect terminologique et 2) à la nature multidisciplinaire des données.

2.2. Hétérogénéité des données du diagnostic in vitro

Dans un premier temps nous présentons des généralités sur la complexité terminologique, complétées par des exemples concrets du monde du DIV dans la deuxième partie. La troisième partie s'attache à présenter des implémentations de systèmes d'intégration de données dans le monde médical et biologique.

2.2.1. Complexité terminologique

La complexité terminologique est une problématique très bien décrite dans la littérature notamment à travers les disciplines de recherche ayant trait à l'indexation de textes (Ananiadou et McNaught, 2006 ; Krauthammer et Nenadic, 2004 ; Liu *et al.*, 2001). Nous distinguons deux problèmes : la variabilité des termes pour représenter un même concept (*synonymie*), et l'ambiguïté d'un terme pouvant représenter deux concepts différents (*polysémie*). Le problème de la synonymie peut être résolu par l'emploi de thesaurus biomédicaux tels que *l'Unified Medical*

Language System (UMLS®, Nelson *et al.*, 2006) ou le portail terminologique de santé (HeTOP, Grosjean *et al.*, 2011). Liu *et al.* décrivent la problématique d'ambiguïté des termes dans le domaine biomédical comme étant lié 1) à la polysémie des termes, 2) à l'utilisation d'acronymes et abréviations, 3) à l'ambiguïté des acronymes. Des dictionnaires d'abréviations existent, comme Allie (Yamamoto *et al.*, 2011) ou ADAM (Zhou *et al.*, 2006) permettant de retrouver la/les formes longues d'une abréviation.

Dans sa *key note* présentée à l'atelier IA Santé Kevin B. Cohen compare langage médical et scientifique. Il met en avant des différences syntaxiques (constructions de phrases différentes) et lexicales (l'utilisation de langages de spécialité différents augmentent le nombre de termes synonymes) au sein des corpus de textes. Ces différences entre langage médical et biologique se retrouvent au sein des différents SOC représentant le DIV (Ogren *et al.*, 2004) et ajoutent une dimension nouvelle à la problématique d'intégration de données dans ce domaine.

2.2.2. Exemples extraits du DIV

Le DIV est un domaine multidisciplinaire : les tests réalisés en laboratoire relèvent à la fois de la biochimie, de l'immunologie, de la microbiologie et de la biologie moléculaire. Les problématiques de variabilité de termes sont plus ou moins exacerbées en fonction du domaine de connaissance. Par exemple, les composés chimiques (biochimie) ou les noms des organismes (microbiologie) suivent des règles très strictes de nomenclature (de Morveau, 1787 ; Lapage *et al.*, 1992). Il n'y a donc pas ou très peu de variabilité terminologique (lexicale et syntaxique) possible pour représenter les concepts appartenant à ces domaines de connaissances. Dans les domaines de l'immunologie ou de la biologie moléculaire les concepts (*i.e.* anticorps anti HIV) sont représentés sous trois formes terminologiques :

– forme complète : Human Immunodeficiency Virus Anticorps ;

– semi-abrégée : Human Immunodeficiency Virus Ab ou HIV Anticorps ;

– totalement abrégée : HIV Ab.

Par ailleurs l'aspect transdisciplinaire du DIV accentue les variabilités lexicales type synonyme et abréviation. Une des terminologies qui illustre le mieux la variabilité terminologique entre langage clinique et langage scientifique est le thésaurus MeSH (*Medical Subject Headings*) qui a été développé pour l'indexation d'articles scientifiques dans MEDLINE accessibles *via* l'outil PubMed[4]. Si l'on cherche le concept représentant l'espèce *Escherichia coli* dans le thesaurus MeSH nous retrouvons plus d'une dizaine de termes synonymes associés au nom scientifique. Parmi les termes synonymes six sont des abréviations complètes, quatre des termes semi-abrégés, et 8 termes contextualisent la description de l'espèce avec une information pathologique (*i.e. E. coli enteroinvasive*). Nous avons donc une

4. http://www.ncbi.nlm.nih.gov/pubmed

variabilité des termes utilisés pour représenter le même concept entre deux SOC qui dépend du contexte d'utilisation, et s'adapte aux connaissances de l'utilisateur.

En plus de la variabilité terminologique, la conceptualisation de la connaissance est une problématique clé dans l'intégration de données du DIV. Si l'on s'intéresse à la description d'un *prélèvement*, cette information est retrouvée à la fois dans la terminologie LOINC® et SNOMED CT®. Au sein de la terminologie LOINC®, le *prélèvement* est une partie de l'information permettant de contextualiser un *test* réalisé ; il est modélisé comme étant une caractéristique de l'entité *test.* Dans l'ontologie SNOMED CT® la notion de *prélèvement* est modélisée comme une entité et peut être utilisée indépendamment de la description d'un *test*. Ces utilisations différentes amènent à une représentation terminologique et conceptuelle différente entre les deux SOC. Dans LOINC® la connaissance liée au prélèvement est sommaire : chaque prélèvement est défini par un terme abrégé et classé dans une hiérarchie avec une faible profondeur moyenne. *A contrario* au sein de l'ontologie SNOMED CT® l'information de *prélèvement* est structurée avec plusieurs dimensions de lecture. Le *prélèvement sang* dans LOINC® est représenté de manière unique par le terme *Bld*, alors que SNOMED CT® distingue cette information en deux concepts : le concept *87612001*| *Blood (substance)*| représentant l'aspect composition chimique du sang, et *119297000* |*Blood specimen (specimen)*| qui représente le sang comme une entité matérielle. D'un point de vue terminologique, les termes sont identiques mais la différence de conceptualisation entre les deux SOC soulève une problématique d'ambiguïté conceptuelle qui a un impact pour le liage des données.

2.2.3. Intégration des SOC biologiques ou médicaux : principales solutions

Les caractéristiques du DIV (qualité, quantité et hétérogénéité) ont un impact sur les stratégies d'intégration des SOC. Dans cette partie nous présentons des solutions retenues pour intégrer des SOC biologiques ou médicaux.

Dans le domaine médical, l'intégration des SOC est réalisée à travers la construction de modèles unifiés qui intègrent un ensemble de terminologies, classifications et ontologies représentant des concepts médicaux. L'*Unified Medical Language System* (UMLS®) Metathesaurus est un thesaurus développé par l'*US National Library of Medicine* et disponible sur le web[5]. Le Metathesaurus intègre plus de 195 ressources terminologiques et ontologiques principalement de langue anglaise qui sont des standards dédiés à la codification des données patients (ICD - *International Classification of Diseases*, SNOMED CT®) à l'expression de domaines biologiques (*Gene Ontology, Human Phenotype Ontology*) à l'indexation de textes (MeSH). Le Metathesaurus propose une unification des 13 millions de termes issus de ces ressources par un regroupement conceptuel (CUI-*Concept Unique Identifier*) propre au Metathesaurus. La version 2015 recense plus de 3 millions de concepts organisés thématiquement par une centaine de types

5. https://www.nlm.nih.gov/research/umls/

sémantiques (*Organism, Body System*) et reliés entre eux par un réseau sémantique décrit par plus de 50 relations (McCray, 1989). Le Metathesaurus de l'UMLS® est très souvent employé comme ressource pour indexer et aligner des SOC médicaux et notamment LOINC® et SNOMED CT® (Bodenreider, 2008 ; Dolin *et al.*, 1998) . Développé à l'origine pour mettre à disposition des SOC en français non disponibles dans l'UMLS, HeTOP (*Health Terminologies and Ontologies Portal*) est un portail multi-terminologique et multilingue développé par l'université et le CHU de Rouen. Il intègre plus de 63 SOC dans 23 langues (principalement le français).

La construction de systèmes intégrant des dizaines de SOC tels que le *Metathesaurus* ou HeTOP nécessite une approche automatisée pour identifier des termes, ou concepts partageant une même sémantique. Merabti *et al.* (2012) proposent en 2012 une revue des méthodes d'alignement terminologique appliquées aux SOC médicaux. De manière plus générale la problématique d'alignement des données entre deux SOC est étudiée pour des problématiques d'alignement d'ontologies (Euzenat *et al.*, 2007).

L'intégration de SOC dans le domaine de la biologie est une problématique qui doit prendre en compte à la fois les questions terminologiques mais également de conceptualisation de la connaissance. En 2015 Lapatas *et al.* (Lapatas *et al.*, 2015) proposent un état des lieux des solutions appliquées pour intégrer des connaissances biologiques qui fait écho à la revue proposée en 2007 par Louie *et al.* (Louie *et al.*, 2007) relative à la problématique d'intégration de données génomiques en médecine. Cet article illustre trois grandes approches pour intégrer des connaissances biologiques.

La première approche consiste à intégrer les connaissances durant les processus d'analyses de données. La deuxième approche utilise l'état de l'art sur l'intégration de bases de données relationnelles pour construire des vues unifiées type *data warehouse* (PathWay commons - Cerami *et al*., 2011) ou *federate databases* (projet ENCODE, 2004). La création de *data warehouse* et de bases de données fédérées sont des stratégies d'intégration matures. La littérature aborde ces problématiques sous différents angles tels que l'architecture des systèmes (Sahama et Croll, 2007) ou des stratégies d'implémentation. En 1986 Batini *et al.* (Batini *et al.*, 1986) présentent une analyse complète des problématiques et stratégies existantes liées à l'intégration de données relationnelles représentées par des schémas conceptuels de données différents. Par la suite, ces travaux ont été généralisés à la problématique d'intégration de schémas (Bellahsene *et al.*, 2011 ; Rahm et Bernstein, 2001). La troisième approche propose une intégration des données à travers un *Linked Open Data* ce qui est notamment l'objectif du projet Bio2RDF (Dumontier *et al*., 2014). Le projet Bio2RDF propose une suite d'outils et d'applications web pour intégrer des données biologiques afin de faciliter leur utilisation notamment dans le domaine de la bio-informatique. Bio2RDF est construit sur une vingtaine de SOC de référence (GO, MeSH, NCBI) qui sont stockés dans des *triple stores* locaux et reliés par un réseau sémantique. L'ajout de données par un utilisateur passe par l'utilisation de *parseur* transformant les données et un ensemble d'ontologies

permettant de faire la correspondance entre les données et les SOC de référence. L'intégration des données dans un *Linked Data* telle que le propose aujourd'hui le projet Bio2RDF se limite aux SOC de référence inclus et cette solution ne permet pas d'incrémenter le *Linked Data* avec de nouveaux domaines de connaissances.

2.3. Classification des méthodes d'alignement

Dans la section précédente nous avons identifié les caractéristiques des SOC du DIV. La section 2.3.3 illustre la problématique d'intégration de SOC à travers des exemples de solutions médicales et biologiques. Nous avons identifié deux axes de recherche permettant d'intégrer des SOC. La construction de thésauri médicaux s'inspire de la littérature sur les alignements d'ontologies pour intégrer les SOC. Les solutions de type *data warehouse* ou *federate databases* utilisent des stratégies d'intégrations basées sur des alignements de schémas de données. Cette partie présente une classification des méthodes d'alignement qui sera complétée dans les sections suivantes par l'étude des métriques d'alignement lexicales (02.4) et les stratégies d'alignements (2.5).

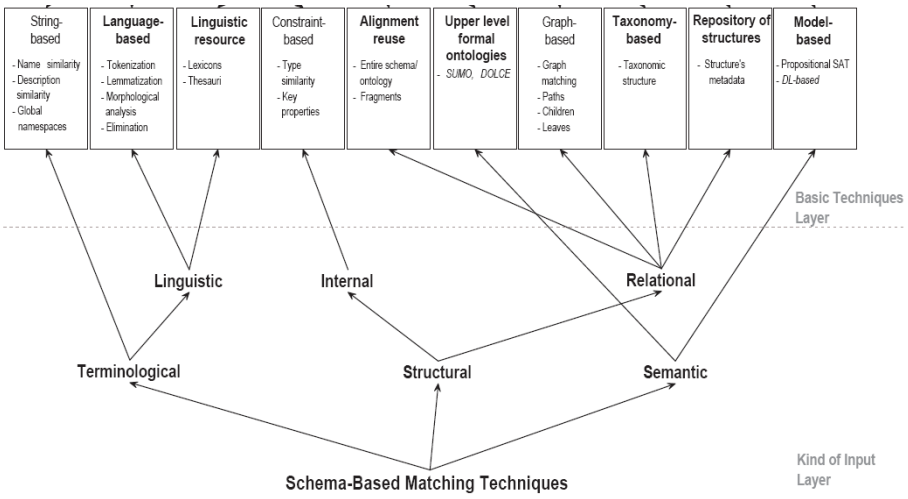

Figure 1. Classification des méthodes d'alignements d'ontologies (extrait de l'article Shvaiko et Euzenat, 2005)

Rahm et Bernstein (2001) ont proposé une classification des méthodes d'alignement pour l'intégration de schémas qui a été incrémentée dans différents articles d'Euzenat et Shvaiko (Shvaiko et Euzenat, 2005 ; Euzenat et Shvaiko, 2007) pour décrire les alignements entre ontologies (voir figure 1). Les méthodes d'alignement sont classées en fonction de trois dimensions. Les méthodes terminologiques utilisent les termes associés aux concepts pour réaliser les

alignements. L'organisation hiérarchique, les relations entre concepts et les contraintes sur les données constituent le point d'entrée pour établir des alignements structuraux entre deux ontologies. La dimension sémantique utilise les propriétés des logiques de description et la réutilisation de *top ontology* (comme la *Basic Formal Ontology*) pour établir des correspondances entre concepts.

Merabti *et al.* proposent en 2012 (Merabti *et al.*, 2012) une autre classification des algorithmes d'alignement pour les terminologies biomédicales. Les terminologies étant globalement moins structurées que les ontologies, cette classification distingue deux types d'algorithmes :

– les algorithmes lexicaux qui utilisent les propriétés linguistiques et lexicales des termes pour définir la similarité entre deux termes ;

– les méthodes structurelles qui utilisent l'organisation taxonomique des termes pour établir une similarité entre termes.

Les métriques lexicales (*String-based*, *Language-based* et *Linguisitic resources*) ont pour avantage de ne pas dépendre de la structure (organisation taxonomique ou schématique) des SOC ; elles peuvent donc être utilisées de manière systématique pour réaliser des alignements entre deux SOC. Dans la section suivante nous développons quelques métriques en proposant une classification.

2.4. Classification des métriques lexicales

On compte actuellement beaucoup de métriques lexicales publiées dans la littérature (Brahma et Refoufi, 2015 ; Euzenat *et al.*, 2007). Nous proposons dans cette section de décrire les principaux types de métriques lexicales en fonction des caractéristiques langagières (*Language*) ou des chaînes de caractères (*String*) utilisées pour comparer deux termes (figure 2).

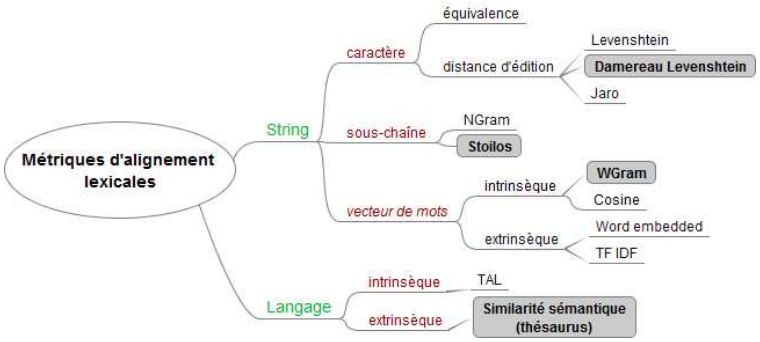

Figure 2. Classification des méthodes et métriques d'alignement lexical. Les métriques utilisées dans la suite de l'article sont encadrées en gris

Les méthodes basées sur les chaînes de caractères (*String*) peuvent être classées en fonction de l'élément atomique pris en compte dans l'établissement des similarités. Les méthodes basées sur les caractères utilisent des fonctions de coût entre deux caractères pouvant être plus ou moins élaborées. Les méthodes d'équivalence utilisent un coût binaire alors que la métrique de Damereau Levenshtein (DL, équation (1)) (Damerau, 1964 ; Levenshtein, 1966) prend en compte trois différences : 1) insertion/délétion, 2) substitution et 3) transposition ; chacune pouvant être paramétrer avec une fonction de coût.

$$sim_{DL}(t_1, t_2) = 1 - \frac{dist_{DL}(t_1, t_2)}{min(|t_1|, |t_2|)} \quad (1)$$

Dans les métriques de type sous-chaîne, on distingue celles dont la longueur de sous-chaîne est fixe et connue (NGram comme les digram ou trigram) et celles dont la longueur varie en fonction de l'alignement. La métrique proposée en 2005 par Stoilos fait partie de cette catégorie. La similarité de Stoilos est une métrique développée spécifiquement pour les libellés de concepts d'ontologies qui pondère positivement les termes ayant un préfixe commun (fonction winkler équation (5), Winkler, 1999). Elle repose sur l'idée que la similitude entre deux chaînes est liée à leurs points communs (fonction *common*, équation (3)) ainsi qu'à leurs différences (fonction *diff*, équation (4)). La distance de Stoilos entre deux chaînes de caractères t_1 et t_2 est définie par la formule suivante :

$$sim_{Stoilos}(t_1, t_2) = common(t_1, t_2) - diff(t_1, t_2) + winkler(t_1, t_2) \quad (2)$$

$$common(t_1, t_2) = \frac{2 * \sum_{i=1}^{n} |substring(t_1, t_2}{|t_1| + |t_2|} \quad (3)$$

$$diff(t_1, t_2) = \frac{diff_1 * diff_2}{p_{diff} + (1 - p_{diff})(|diff_1| + |diff_2| - |diff_1| * |d} \quad (4)$$

$$winkler(t_1, t_2) = p_{winkler} * (1 - common(t_1, t_2)) * min(4, |common_{prefix}|) \quad (5)$$

Les métriques basées sur les vecteurs de mots sont classées en fonction des informations requises pour leur mise en place. Nous distinguons les méthodes intrinsèques qui ne nécessitent aucune ressource externe des méthodes extrinsèques qui utilisent notamment des corpus de textes pour calculer la similarité. La métrique WGram (Mary *et al.*, 2016) que nous proposons (détaillée dans la section 3.1.1), est une méthode intrinsèque qui s'appuie sur la décomposition des termes en vecteur de mots. Les méthodes extrinsèques telles que le calcul de l'inverse de la fréquence d'un terme (TFIDF –- Robertson et Jones, 1976) et *Word Embedded* (Turian *et al.*, 2010) utilisent des corpus de texte pour établir et regrouper des mots en fonction de leur fréquence et cooccurrence.

Certaines métriques lexicales utilisent les propriétés inhérentes à une langue (*Language*) pour normaliser et évaluer la similarité entre deux termes. Nous pouvons également les catégoriser en méthodes intrinsèques et extrinsèques. La lemmatisation d'un terme permet d'affiner la comparaison de deux termes en utilisant leur forme normale ce qui permet de s'affranchir des problèmes liés au genre, pluriel et formes nominative ou verbale. Par exemple, la lemmatisation permet d'identifier les termes *normalisation* et *normalisée* comme étant très similaires de par leur racine commune *norm*. Ce type d'algorithme est très dépendant du langage utilisé dans les SOC. Des problématiques similaires sont abordées pour améliorer l'indexation de textes biomédicaux sur des Metathesaurus (Hettne *et al.*, 2010) ou améliorer l'alignement de noms d'éléments schématiques (Sorrentino *et al.*, 2010). Ils proposent un système de règles adaptées à leur contexte d'étude visant à normaliser les termes. Les méthodes langagières extrinsèques utilisent des thésaurus comme ressources externes pour calculer la similarité entre deux termes. Leur principal avantage est qu'elles permettent d'identifier des concepts identiques représentés par des termes synonymes sans racine commune. L'UMLS®, plus particulièrement le *Specialist Lexicon,* est la principale ressource extrinsèque utilisée dans le domaine biomédical. Dans cet article nous proposons une méthode réutilisant la ressource UMLS® pour calculer la similarité sémantique entre deux termes (détaillée dans la section 3.1.2).

2.5. Stratégies d'alignement et filtres

Dans la section précédente nous avons présenté de manière non exhaustive quelques métriques lexicales mentionnées dans la littérature. Ces métriques sont des indicateurs de la similarité entre deux concepts. Cette partie s'attache à présenter les stratégies d'alignement que nous définissons comme étant l'implémentation d'une combinaison de métriques de similarité et de filtres permettant d'aligner deux SOC.

La problématique de combinaison des méthodes est détaillée dans l'ouvrage *Schema Matching* (Bellahsene *et al.*, 2011) et *Ontology Matching* (Euzenat *et al.*, 2007). Les auteurs distinguent trois approches de combinaison de méthodes 1) l'approche séquentielle, 2) l'approche parallèle et 3) les approches mixtes. La combinaison séquentielle de méthodes d'alignement renvoie à un résultat d'alignement dépendant de l'ordre d'exécution des méthodes. L'approche de parallélisation des méthodes a pour avantage de permettre l'optimisation les temps de calculs d'un alignement mais s'accompagne d'une phase de réconciliation des résultats obtenus par les différentes méthodes.

L'établissement d'une stratégie d'alignement pour l'intégration de données doit tenir compte de la volumétrie des SOC à intégrer. L'exploration complète de l'espace d'alignement est envisageable pour des SOC avec un faible volume de données. Pour deux SOC de l'ordre du millier de termes chacun, le nombre de comparaisons est de l'ordre du million. Les SOC issus du DIV étant beaucoup plus volumétriques (350 000 concepts dans SNOMED CT®, 44 200 classes dans la *Gene*

Ontology), il est nécessaire de mettre en place des stratégies heuristiques pour diminuer l'espace de recherche et/ou le temps de calcul. La littérature propose deux approches pour réduire les temps de calcul pour des alignements volumétriques. La première approche consiste en la parallélisation de la recherche ; l'espace d'alignement est exploré de manière complète avec des calculs distribués. La seconde approche, consiste à réduire l'espace de recherche d'alignement. Cette problématique est abordée par les approches d'alignement *de novo* d'ontologie ou de schémas et de réalignement d'ontologie. L'approche *de novo* utilise le principe de partitionnement des SOC en clusters de concepts consistants. Hamdi *et al.* (2010) proposent de construire des blocs en utilisant les relations taxonomiques décrites au sein d'un SOC. Pour restreindre l'espace de recherche entre les partitions des deux SOC, les auteurs calculent dans un premier temps un alignement lexical des libellés qui, une fois filtrés, sont utilisés comme ancres pour générer des alignements avec des méthodes structurelles. Outre la restriction de l'espace de recherche, ces méthodes ont pour avantage de proposer des alignements consistants avec la structure hiérarchique des deux SOC. Pour répondre à la problématique de réalignement de SOC Dos Reis (2015) propose d'utiliser une version antérieure d'alignement et des algorithmes de calcul des différences entre deux versions d'un SOC pour mettre à jour des alignements. La méthode d'alignement de Seddiqui et Aono (2009) s'inspire des deux principes présentés précédemment. Les auteurs utilisent un alignement préexistant pour construire les ancres et d'étendre cet alignement en utilisant des méthodes de comparaison structurelles et lexicales entre concepts voisins. La stratégie des ancres que nous proposons et détaillons en section 3.1.3 s'inspire en partie de l'algorithme proposé par Seddiqui et Aono.

3. Matériel et méthode

3.1. Méthode

Les méthodes que nous avons évaluées dans cette étude ont été sélectionnées en fonction des caractéristiques des données du DIV. La stratégie des ancres permet de répondre aux problématiques liées au volume de données et à leurs évolutions régulières. La variabilité terminologique a été prise en compte par l'évaluation de métriques lexicales (syntaxiques et sémantique) représentatives des différentes approches de la littérature (voir section 2.4 et la figure 2). Les méthodes sont implémentées dans le langage R.

3.1.1. Similarités syntaxiques

Dans cette, étude nous avons comparé trois méthodes de scores, la similarité calculée à partir de la distance de Damereau Levenshtein, la similarité proposée par Stoilos (2005) et une métrique que nous avons développée (que nous intitulons WGram) basée sur le calcul de similarité sur les vecteurs de mots composant deux termes. La méthode WGram s'appuie sur la décomposition des termes en vecteurs

de mots (*w*) (algorithme 1) et produit pour un alignement deux scores de similarité ($t_1 \rightarrow t_2$ et $t_2 \rightarrow t_1$).

Algorithme 1. Processus d'alignement de la méthode WGram

```
1:  alignementWGram (SOC1, SOC2, filtre : bool, seuil∈[0,1]l)
2:  {
3:      alignement ← ∅
        // création de l'alignement des mots composant les termes des deux SOC
4:      index ← alignementMot(SOC1, SOC2)
        // sélection des termes t1, t2 possédant au moins un mot aligné dans l'index
5:      combineT1T2← alignementPotentiel(SOC1, SOC2, index)
6:      pour tout i ∈ combineT1T2 faire
7:          alignement ← alignement ∪ calculWGram(i,index)
8:      fin pour
9:    si filtre faire alignement←BestForBoth(alignement)
10:   alignement← superieurA(alignement, seuil)
11:   retourne(alignement)
12:   }
```

La similarité d'un terme par rapport à un autre est calculée (*calculWGram*) grâce aux similarités des mots qu'ils ont en communs (équation (6)). L'alignement des mots (w) composant les termes du SOC 1 et ceux du SOC 2 sont obtenus par calcul de similarités de Damereau Levenshtein décrite dans la section 2.4 (simDL).

$$sim_{WGram}(t_1 \rightarrow t_2) = \frac{\sum_{i=1}^{K} |w_1^i| * sim_{DL}(w_1^i, w_2^i)}{\sum_{i=1}^{N} |w_1^i|} \quad (6)$$

Où *K* représente le nombre de mots alignés (w^i) entre les termes t_1 et t_2, N représente le nombre de mots (w^i_1) composant t_1 et w^i_2 un mot composant le terme t_2.

3.1.2. Similarité sémantique

La similarité sémantique mesure l'adéquation de deux termes en fonction de leur signification. La méthode que nous proposons utilise le Metathesaurus de l'UMLS® pour identifier la sémantique de chaque terme. Le calcul de la similarité sémantique entre deux termes se décompose en deux étapes : la première consiste à identifier la (ou les) signification(s) de chacun des termes des deux SOC (0) et la seconde consiste à calculer la similarité entre la sémantique du terme t_1 et celle du terme t_2 (3.1.2.2).

3.1.2.1. Extraction des informations sémantiques d'un terme

Nous avons développé un algorithme pour déduire la (les) sémantique(s) d'un terme à partir des concepts identifiés (CUI) dans le Metathesaurus de l'UMLS®. Nous utilisons l'outil MetaMap initialement développé pour l'indexation de textes médicaux (Aronson, 2006).

Algorithme 2. Protocole d'obtention des termes sémantiques candidats

```
1:  créationTermeSemantique(id, term)
2:  {
3:      termeSemantique←∅
4:      listTSem ←initialiserListe(termCandidat)
5:      resultat ← appelMetaMap(id,terme)
6:      listPhrase← recupererPhrases(resultat)
        // Création des termes candidats
7:     pour tout p ∈ listPhrase faire
8:      oldTSem← listTSem
9:      listTSem ← ∅
10:       pour chaque mapping ∈ p.MetaMaping faire
11:           listTSem← listTSem ∪ ajoutInfoCandidat(oldTSem, (mapping))
12:     fin pour
13:       Si longeur(listTSem) == 0 faire
14:          listTSem ← ajoutInfoCandidat(oldTSem, termeSemantique)
15:       fin Si
16:     fin pour
17:   retourne(listTSem )
18:   }
```

Le principe d'indexation de MetaMap est le suivant. MetaMap segmente le *terme* en un vecteur de sous termes (*listPhrase*). Pour chacun des sous-termes (*p*) l'outil MetaMap identifie un (ou plusieurs) alignement(s) (*mapping*) sur un concept UMLS (CUI). Un *mapping* est composé d'un CUI et d'un score d'alignement. Chaque *mapping* entre un sous-terme *p* et un CUI représente une sémantique possible de *p*.

La fonction *créationTermeSemantique* (algorithme 2) interprète l'indexation du *terme* proposé par MetaMap (*resultat)* pour déduire la (les) sémantique(s) du *terme* (*listTSem*). Un terme est associé à plusieurs significations, s'il existe un sous-terme *p* impliqué dans plusieurs *mapping.* Dans la suite de cet article nous appelons un *termeSemantique* comme étant une signification possible du *terme*. Nous définissons un *termeSemantique* comme étant une combinaison de concepts UMLS® (CUI). Chaque CUI représente la signification d'un des sous-termes (*p*) du vecteur *listPhrase*. Nous mesurons l'adéquation entre un *termeSemantique* et le *terme* qu'il représente en calculant la moyenne des scores d'alignement (noté confiance sémantique ou $conf_{sem}$).

3.1.2.2. Similarité sémantique entre deux termes

Le calcul de la similarité sémantique entre deux termes (t_1 et t_2) se détermine de manière indirecte à travers l'alignement de leurs *termeSemantique* (t^i_{sem1}, t^j_{sem2}). Nous définissons la similarité sémantique (sim_{sem} équation (7)) entre deux termes t_1 et t_2 comme étant la similarité de hamming ($sim_{hamming}$) maximale calculée pour toutes les combinaisons de termes sémantiques (t^i_{sem1}, t^j_{sem2}) représentant respectivement les termes t_1 et t_2. Le calcul de la similarité de hamming est réalisé sur les vecteurs de CUI représentant les termes t^i_{sem1}, t^j_{sem2}.

$$sim_{sem}(t_1, t_2) = max(sim_{hamming}(t^i_{sem1}, t^j_{sem2})) \quad (7)$$

Nous mesurons également la confiance de l'alignement comme étant le minimum entre la confiance sémantique des *termeSemantique* (t^i_{sem1}, t^j_{sem2}) représentant respectivement t_1 et t_2 (t^i_{sem1}, t^j_{sem2}) étant les *termeSemantique* qui maximisent la similarité sémantique).

$$conf_{sim.sem}(t^i_{sem1}, t^j_{sem2}) = min(conf_{sem}(t^i_{sem1}), conf_{sem}(t^j_{sem2})) \quad (8)$$

3.1.2.3. Paramètres utilisés pour MetaMap

Pour cette étude nous effectuons la recherche de concepts sur une version modifiée du Metathesaurus® (2014AB) qui exclut les ressources LOINC® et SNOMED CT®. En excluant les SOC nous cherchons à estimer les performances de la méthode d'alignement sémantique entre deux ressources dans le cas où celles-ci ne seraient pas incluses dans le Metathesaurus®.

3.1.3. Alignement heuristique par la méthode des ancres

La méthode des ancres que nous proposons est une stratégie d'alignement heuristique, initiée à partir d'un alignement préexistant entre les deux SOC. Le principe de cette méthode s'inspire d'un algorithme développé pour résoudre les problématiques d'alignement d'ontologies volumineuses (Seddiqui et Aono, 2009). La méthode que nous proposons (algorithme 3) permet d'étendre un alignement initial entre deux ressources. Nous définissons cet alignement comme étant un *ObjetAncre* (figure 3). L'algorithme d'extension d'alignement part du postulat que les deux ressources (*ObjetAncre.SOC1* et *ObjetAncre.SOC2*) reposent sur une organisation hiérarchique des données similaires. Le principe de la méthode consiste à étendre de manière récursive les alignements initiaux (appelés par la suite les ancres initiales) en comparant trois sous-ensembles de termes : les termes parents (fonction *parents*), les termes enfants (fonction *enfants*) et les termes frères (fonction *freres*). La comparaison des termes (fonction *aligner*) peut être effectuée avec l'ensemble des métriques syntaxiques et sémantique présentées dans les sections précédentes grâce à la spécification des paramètres dans *ObjetAncre*. Dans cette étude nous utilisons la méthode (DL) pour comparer les termes, et nous filtrons les alignements en fonction d'un seuil de similarité.

Algorithme 3. Méthode des ancres

```
1:  trouverNouvellesAncres(i)
2:  {
      nouvellesAncres← ∅
3:    miseAJourStatutAnalyse(ObjetAncre,i)
4:    si ObjetAncre.provenance(i) <> 'enfant' faire
5:      parentsId1← ObjectAncre.parent(i, 'soc1')
6:      parentsId2← ObjectAncre.parent(i, 'soc2')
7:      nouvellesAncres← nouvellesAncres∪ObjectAncre.aligner(parentsId1, parentsId2,
                                                  provenance = 'parent')
8:    fin si
9:    si ObjetAncre.provenance(i) <> 'parent' faire
10:     enfantsId1← ObjectAncre.enfant(i, 'soc1')
11:     enfantsId2← ObjectAncre.enfant(i, 'soc2')
12:     nouvellesAncres← nouvellesAncres ∪ ObjectAncre.aligner(enfantsId1,
                                                  enfantsId2, provenance = 'enfant')
13:   fin si
14:   si ObjectAncre.provenance(i) <> 'frère' faire
15:     freresId1← ObjectAncre.frère(i, 'soc1')
16:     freresId2← ObjectAncre.frère(i, 'soc2')
17:     nouvellesAncres← nouvellesAncres ∪ ObjectAncre.aligner (freresId1, freresId2,
                                                  provenance = 'frere')
18:   fin si
19:   ObjectAncre.ajouter(nouvelleAncres)
20:   pour j ∈ nouvellesAncres faire
21:     ObjectAncre.trouverNouvellesAncres(j)
22:   fin pour
23: }
```

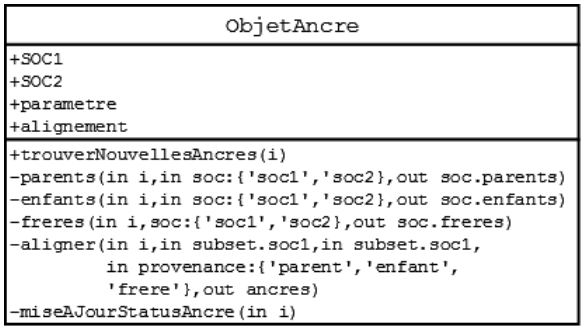

Figure 3. Diagramme de classe ObjetAncre

3.2. Normalisation des termes et filtre des alignements

3.2.1. Normalisation des termes

Nous avons défini une méthode de normalisation des termes par SOC. Dans les parties LOINC® les éléments de ponctuations sont supprimés. Les libellés des concepts SNOMED CT® sont normalisés au niveau de la ponctuation et le tag sémantique est supprimé.

3.2.2. Paramètres de filtre des alignements

Nous utilisons deux filtres pour étudier les alignements calculés à partir des similarités syntaxiques. Par défaut, un alignement entre deux SOC ($\mathcal{T}_1$et $\mathcal{T}_2$) est composé des meilleurs alignements par terme du SOC 1 et des meilleurs alignements par terme du SOC 2 (équation (9)).

$$\mathcal{A}(\mathcal{T}_1, \mathcal{T}_2) = \{\mathcal{A}(t_x, t_y) | sim(t_x, t_y) = max(sim(t_x, \mathcal{T}_2)) \vee sim(t_x, t_y) = max(sim(\mathcal{T}_1, t_y))\} \quad (9)$$

Le filtre `BestForBoth` (équation (10)) que nous proposons permet de sélectionner les alignements qui sont les meilleurs à la fois pour t_1 et pour t_2.

$$BestForBoth(\mathcal{T}_1, \mathcal{T}_2) = \{\mathcal{A}(t_x, t_y) | sim(t_x, t_y) = max(sim(t_x, \mathcal{T}_2)) = max(sim(\mathcal{T}_1, t_y))\} \quad (10)$$

Les résultats de ces deux filtres sont contextuels. Ils sont fonction à la fois de la métrique utilisée mais surtout des deux ensembles de termes utilisés pour l'alignement.

3.3. Matériel

3.3.1. Logical Observation Identifiers Names and Codes (LOINC®)

La terminologie LOINC®[1] a été construite en 1994 afin de standardiser la description des tests cliniques et de diagnostic *in vitro.* Elle est développée et mise à jour deux fois par an par le *Regenstrief Institute* (Sheide et Wilson, 2013). Un test codé en LOINC® se décompose en 6 dimensions. Les dimensions *Composant, Milieu, Technique* et *Temps* décrivent le principe du test de laboratoire alors que les dimensions *Échelles* et *Grandeur* caractérisent le type de résultat. Les valeurs permettant d'exprimer chacune de ces dimensions (nommées *parties* par la suite) sont organisées hiérarchiquement. Afin de compacter la description d'un test, les parties *Milieu*, *Échelles* et *Grandeurs* sont représentées par des codes mnémoniques. Par exemple le mot « *blood* » est représenté par le code mnémonique « bld » dans la partie *Milieu*. Les parties représentant des *Composant* ou des *Technique* peuvent contenir des abréviations (par exemple « Ab » pour « *Antibody* » ou « EIA » pour « *Elisa Immuno Assay* »). Dans cette étude nous utilisons la version 2.5 de LOINC®

décrivant plus de 68 000 tests composés par 40 000 parties. Les parties et leur hiérarchie sont extraites de la base de données utilisée par l'application RELMA[1].

3.3.2. Systematized Nomenclature Of MEDicine – Clinical Terms (SNOMED CT®)

La SNOMED CT® est une ontologie sous licence créée et maintenue deux fois par an par l'*International Health Terminology Standards Development Organisation* (IHTSDO)[2] (Cornet et de Keizer, 2008). L'ontologie SNOMED CT® permet de représenter le domaine clinique avec plus de 350 000 concepts organisés en 19 axes. Les concepts sont identifiés par un libellé unique (*Fully Specified Name*) qui se compose d'un terme spécifiant la sémantique du concept et d'un tag sémantique. Le tag sémantique est placé à la fin du libellé et apporte une information contextuelle sur la classification dans SNOMED CT® et l'utilisation du concept. Cette étude a été réalisée à l'aide de la version de SNOMED CT® de janvier 2015[6].

3.3.3. Alignement entre LOINC® et SNOMED CT®

Les alignements entre LOINC® et SNOMED CT® utilisés dans cette étude sont le résultat d'une collaboration initiée en 2013 entre l'IHTSDO et le *Regenstrief Institute* qui a pour objectif d'aligner la description des tests LOINC® sur des concepts SNOMED CT® pour améliorer l'agrégation de données dans les dossiers patients informatisés (Vreeman, 2015). L'alignement est réalisé à deux niveaux. Dans un premier temps les parties décrivant les dimensions d'un test ont été alignées sur des concepts SNOMED CT®. Par la suite les tests LOINC® sont décrits par des définitions formelles en SNOMED CT®. Dans cette étude nous utilisons la première version d'alignement publiée en septembre 2014[7], qui couvre 0,15 % des tests LOINC® et 2 115 parties. L'alignement partie-concepts n'est pas bijectif. Sur les 2 177 alignements, 62 parties LOINC® sont alignées sur plusieurs concepts et 92 concepts sont alignés avec plus d'une partie LOINC®.

4. Résultats et discussion

L'objectif de cette évaluation est d'identifier les méthodes et métriques les plus performantes pour l'alignement de termes appartenant au domaine du diagnostic *in vitro* (voir 0). Nous évaluons les méthodes d'alignement avec deux critères :

– Nous favorisons des méthodes précises (donnant peu de faux positif) pour respecter les exigences de qualité du monde médical. Nous distinguons donc la précision du rappel pour affiner nos conclusions ;

– Les métriques sont également jugées selon leur capacité à rendre des résultats pertinents quel que soit le degré d'hétérogénéité des sous-domaines du DIV (2.2). C'est ce que par la suite nous appelons la robustesse.

---

6. https://uts.nlm.nih.gov/home.html

7. https://loinc.org/news/draft-loinc-snomed-ct-mappings-and-expression-associations- now-available.html/

4.1. Évaluation des méthodes de similarité syntaxiques

4.1.1. Analyse des similarités syntaxiques

Le tableau 1 résume l'évaluation des alignements obtenus par les métriques de similarité syntaxique. Tout d'abord il faut noter que pour utiliser la méthode Damereau Levenshtein, les termes LOINC® et SNOMED CT® ont été normalisés. Cette étape de normalisation n'est pas nécessaire pour les méthodes Stoilos et WGram.

Tableau 1. Résultat de l'évaluation des métriques de similarité syntaxique. (Norm : données normalisées ; P : précision ; R : rappel)

Méthode	Norm.	Filtre	Nb alignements	P.	R.
DL	Oui		4 589	0,26	0,54
	Oui	`BestForBoth`	1 281	0,77	0,45
Stoilos (p_{diff} = 0,6 ; $p_{winkler}$ = 0,1)	Non		3 132	0,50	0,72
	Non	`BestForBoth`	**1 356**	0,92	0,57
WGram	Non		3 263	0,47	0,71
	Non	`BestForBoth`	2 322	0,63	0,67
WGram et Stoilos	Non	`Stoilos.BestForBoth`	1 202	**0,95**	0,53
WGram ou Stoilos	Non	`Stoilos.BestForBoth`	3 497	0,47	**0,76**

On observe une augmentation significative de la précision (+50 %) pour les résultats d'alignement DL et Stoilos après l'application du filtre `BestForBoth` sans pour autant diminuer de manière significative le rappel. Dans un second temps nous avons cherché à combiner les métriques pour maximiser la précision. La combinaison *WGram et Stoilos* permet d'obtenir des alignements avec une forte précision (95 %) ce qui est cohérent avec les exigences de qualité du monde du DIV.

4.1.2. Comparaison des temps de calcul entre les métriques

L'étude que nous avons réalisée n'est pas représentative de la volumétrie des données du DIV. Nous nous sommes donc intéressés à l'évolution des temps de calcul des différentes métriques en fonction de la volumétrie des SOC initiaux (figure 4). On observe des courbes de temps de calcul très différentes en fonction des métriques utilisées. Le temps de calcul des alignements Stoilos et WGram suit une courbe quadratique (voire exponentielle). Nous pouvons en conclure que les méthodes Stoilos et WGram bien que permettant d'obtenir des alignements de qualité ne sont cependant pas adaptées à l'alignement de SOC avec volume de données de l'ordre du 10^4-10^5.

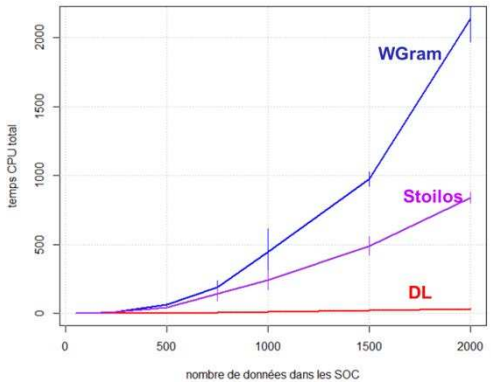

Figure 4. Temps d'exécution moyen (10 répétitions) des métriques DL, Stoilos et WGram en fonction de la volumétrie dans les SOC initiaux

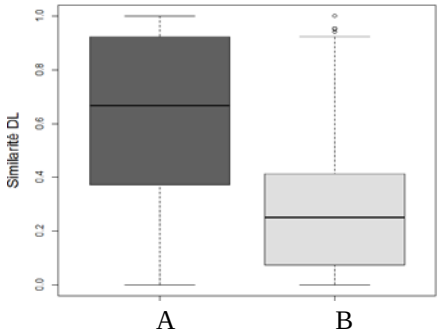

Figure 5. Distribution de la similarité calculée avec la métrique DL en fonction des dimensions LOINC. A : la dimension Composant, B : les autres dimensions

4.1.3. Comparaison des performances du filtre BestForBoth avec des seuils

Les performances des méthodes de similarité sont généralement étudiées avec un paramètre de seuil comme filtre d'alignement. Dans cette étude nous avons décidé de ne pas investiguer les performances associées à ce type de filtre. Ce choix est motivé par une question majeure : Comment déterminer sans *a priori* le seuil de filtrage ?

Comme expliqué dans la section 2.2.2, la variabilité de la représentation terminologique des concepts dépend fortement du sous-domaine étudié. On observe que le filtre `BestForBoth` permet une meilleure discrimination des vrais alignements. Par exemple pour le sous-domaine des organismes, les termes *Babesia bovis* et *Babesia ovis* ne représentent pas les mêmes taxa malgré une forte similarité (0,92 ; `BestForBoth` = faux) alors que les termes *Leukocytes* et *Leukocyte*

représentent le même concept avec une similarité plus faible (0,89 ; `BestForBoth` = vrai). La figure 5 illustre la distribution de la similarité en fonction des dimensions LOINC®. On observe un écart de moyenne de 0,3 entre les alignements obtenus pour des composants (termes peu abrégés dans LOINC®) et les autres alignements souvent abrégés ou représentés sous forme mnémonique. L'emploi d'un seuil de similarité pour filtrer les alignements sur les similarités syntaxiques doit être paramétré en fonction de la variabilité lexicale du sous-domaine étudié. Avec un paramétrage *a priori* du seuil en fonction des sous-domaines étudiés, le critère de robustesse n'est pas respecté. Il faudrait donc être capable de paramétrer le seuil sans *a priori*, en utilisant par exemple des statistiques bayésiennes ou des analyses de variabilité intra SOC.

4.2. Évaluation de la stratégie des ancres

L'évaluation des similarités syntaxiques a permis de démontrer deux principaux freins à leur utilisation pour l'alignement des SOC du DIV :

– les métriques syntaxiques ne permettent pas d'aligner des termes abrégés sur leur forme complète ;

– le temps de calcul des métriques Stoilos et WGram n'est pas adapté au volume de données des SOC du DIV.

Dans cette partie nous étudions la stratégie des ancres pour générer des alignements et à un filtre utilisant la similarité sémantique pour répondre au problème de variabilité des termes (synonymes et forme complètes et abrégées).

4.2.1. Évaluation des termes sémantiques candidats

Avant d'utiliser le calcul de similarité sémantique pour filtrer *a posteriori* les données obtenues par la méthode des ancres, nous avons vérifié la cohérence des termes sémantiques que nous obtenons par l'indexation de MetaMap (03.1.2.1). Nous observons que 24 % des parties LOINC® (688) et 17 % des concepts SNOMED CT® impliqués dans l'alignement n'ont pas d'équivalent (terme sémantique) dans le Metathesaurus. Nous remarquons également que les proportions de parties sans terme sémantique sont plus élevées dans les dimensions *Milieu* (60 %), *Unité* (83 %) et *Méthode* (40 %) dont les libellés sont sous une forme soit abrégée, soit de codes mnémoniques. Nous pouvons en conclure que l'utilisation de l'UMLS pour identifier la sémantique d'un terme ne permet pas de résoudre les problématiques liées aux codes mnémoniques pour désigner un concept. Nous avons également observé que 45 % des alignements de la collaboration impliquent un terme LOINC® ou SNOMED CT® représenté par aucun terme sémantique candidat. Ce résultat explique en partie le faible rappel obtenu par l'application du filtre sémantique sur les ancres initiales (voir tableau 2). L'utilisation de l'UMLS comme thésaurus ne permet pas de résoudre de manière systématique un alignement sémantique entre deux termes.

Cependant, on observe que plus de 50 % des termes sont représentés par un unique terme sémantique candidat, ce qui nous conforte dans l'idée que l'UMLS est une ressource suffisamment précise pour calculer la similarité sémantique entre deux termes du DIV. Pour confirmer la pertinence des termes sémantiques candidats, nous avons calculé les fréquences des types sémantiques associés aux termes sémantiques candidats pour chaque dimension de LOINC® et chaque axe SNOMED CT® représentés. Nous observons que les types sémantiques les plus fréquemment retrouvés sont cohérents avec les sous-domaines représentés par les dimensions et les axes. Par exemple les concepts de l'axe *410607006|Organism* sont indexés par les types sémantiques représentant soit les procaryotes (*T007|Bacterium*), les eucaryotes (*T204|Eukaryote*) ou les virus (*T005|Virus*).

4.2.2. Performances de la stratégie des ancres et du filtre sémantique a posteriori

La méthode des ancres utilise un alignement existant pour déterminer les ancres initiales. Dans cette article nous avons utilisé le jeu d'alignement fournit par la collaboration, comme alignement initial (ancres initiales) que nous avons cherché à étendre avec la méthode des ancres.

Tableau 2. Résultats de l'évaluation des méthodes des ancres. La précision et le rappel des ancres générées sont calculés à partir de la curation manuelle des données non filtrées. 1) L'alignement est un sous-ensemble du jeu de données initial, la précision est donc de 1. 2) Aucun rappel ne peut être calculé car l'alignement de référence est obtenu par curation à partir de ces données

Méthode	Nb al.	Paramètres de filtre	Précision	Rappel
	2 177 ancres initiales normalisées	($sim_{Semantique} = 1 \wedge conf_{Semantique} > 800$)	NA (1)	0,37
Ancres DL à 0,80	1 833 ancres générées	NA	0,33	NA (2)
		`BestForBoth`	0,58	0,97
		$sim_{Semantique} = 1$	0,69	0,85
		$sim_{Semantique} = 1 \wedge conf_{Semantique} > 800$	0,81	0,83
		($sim_{Semantique} = 1 \wedge conf_{Semantique} > 800) \vee sim_{DL} = 1$	**0,82**	**0,87**
		$sim_{Semantique} = 1 \wedge conf_{Semantique} > 800 \wedge$ `BestForBoth`	0,82	0,83

Nous avons choisi de réaliser le calcul des nouvelles ancres par la méthode DL en appliquant un seuil de 0,8 pour filtrer les alignements générés (tableau 2). Le paramètre de seuil 0,8 permet d'obtenir une précision supérieure à 0,75 pour l'alignement réalisé à partir de alignements LOINC® SNOMED CT® issus de la collaboration avec un nombre d'alignements cohérent par rapport au nombre de vrais alignements (< 1 200 alignements). L'utilisation ici du filtre *BestForBoth* pendant la génération des ancres n'est pas judicieuse. En effet, *BestForBoth* est un filtre contextuel (dont le résultat varie en fonction de l'ensemble des alignements) et dont la pertinence dépend du nombre total d'alignements. Les résultats obtenus avec le paramètre 0,8 ont une précision très faible (0,33). L'application *a posteriori* des filtres de similarité sémantique ($sim_{Semantique}$) et BestForBoth permettent de doubler la précision de l'alignement sur les ancres générées. Les meilleures performances sont obtenues par combinaison des informations sémantiques et syntaxiques. En effet, les filtres (i) $(sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800) \vee sim_{DL} = 1$ et (ii) $sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800 \wedge$ BestForBoth permettent d'obtenir 80 % de précision pour 80 % de rappel.

5. Conclusion

La problématique d'intégration de données a été largement étudiée dans la littérature que ce soit par la proposition de méthodes génériques ou appliquées à un domaine particulier. Lors de cette étude nous avons cherché à identifier les méthodes les plus appropriées pour lier des données du DIV. Cet article compare des méthodes existantes (Stoilos, DL) à de nouvelles méthodes inspirées de la littérature (WGram, Algorithme des ancres, similarité sémantique).

Nous avons montré que l'utilisation d'un filtre basé sur les meilleurs alignements par terme (BestForBoth) donne des résultats plus précis et complets que l'utilisation d'un seuil de similarité sans pour autant augmenter significativement le temps d'exécution des algorithmes. Nous avons également montré que la combinaison de métriques syntaxiques *Stoilos et WGram* permet d'augmenter les performances mais ne sont pas adaptées pour réaliser des alignements entre deux SOC volumineux. Pour optimiser les temps de calcul de ces algorithmes nous envisageons de paralléliser les processus d'alignement.

Nous avons par la suite étudié les résultats issus d'un algorithme heuristique qui, couplé avec un filtre sémantique *a posteriori*, permet d'améliorer la précision et le rappel. Un des avantages de la méthode des ancres est qu'elle garantit des alignements cohérents avec l'organisation hiérarchique des concepts dans les deux SOC ; les alignements obtenus sont donc directement utilisables pour enrichir un SOC avec de nouveaux concepts ou pour réaliser des études complémentaires entre ces deux SOC (Mary *et al.*, 2016). Cette méthode sera complétée par l'implémentation de la combinaison de métriques *Stoilos et WGram* et l'intégration du calcul de similarité sémantique pendant la génération de l'alignement.

Le calcul de similarité sémantique est une méthode extrinsèque dont les performances varient en fonction de la présence d'information non spécifique au concept (tel le tag sémantique). Nous avons également montré que le Metathesaurus® bien qu'intégrant près de 200 ressources ne permet pas de garantir l'identification d'une sémantique (*termeSemantique*) pour l'ensemble des parties LOINC®, ou concepts SNOMED CT®. Nous préconisons d'utiliser la similarité sémantique (et de manière plus générale toutes les métriques extrinsèques) avec des méthodes ne dépendant pas de ressources externes (comme les métriques lexicales). L'un des enjeux dans l'alignement entre LOINC® et SNOMED CT®, ou de manière plus générale entre les SOC, concerne l'alignement d'abréviations. Les métriques syntaxiques sont par essence incapables de reconnaître les abréviations. Nous pensions que l'utilisation du Metathesaurus® améliorerait l'alignement entre termes abrégés. L'étude que nous avons réalisée n'a pas permis de confirmer cette hypothèse. La gestion des abréviations demeure donc une problématique que nous envisageons de résoudre grâce à l'utilisation de dictionnaires d'abréviations tels qu'ADAM (Zhou *et al*., 2006) ou Allie (Yamamoto *et al*., 2011).

En conclusion, cette étude nous a permis d'identifier les forces et faiblesses de chaque algorithme et métrique d'alignement. L'ensemble de ces travaux nous a permis d'établir la stratégie d'alignement que nous allons réutiliser sur d'autres SOC représentant le domaine du diagnostic *in vitro.* Il nous semble important de souligner que quelle que soit la méthode utilisée, il est fondamental d'impliquer des experts dans le processus d'alignement entre deux SOC. Les méthodes de filtres que nous proposons garantissent au mieux 80 % de précision et 80 % de rappel. Les similarités calculées doivent être considérées comme des indicateurs d'alignements plus que comme une vérité systématique. Ce constat nous pousse à élargir notre conception d'un processus d'alignement. En plus de l'aspect algorithmique, le processus doit également intégrer une part d'expertise humaine par une dimension visuelle et explicative de l'alignement. Ces points sont argumentés dans le domaine de recherche sur les alignements d'ontologies (Shvaiko et Euzenat, 2008).

Bibliographie

Ananiadou S., McNaught J. (2006). *Text Mining for Biology and Biomedicine*. Citeseer.

Aronson A. R. (2006). Metamap: Mapping text to the umls metathesaurus. Bethesda, MD: NLM, NIH, DHHS, p. 1-26.

Bellahsene Z., Bonifati A., Rahm E. (2011). *Schema Matching and Mapping* (vol. 57). Springer.

Blumenthal D. (2010). Launching HITECH. *New England Journal of Medicine*, vol. 362, n° 5, p. 382-385.

Bodenreider O. (2008). Issues in Mapping LOINC Laboratory Tests to SNOMED CT. *AMIA Annual Symposium Proceedings*, p. 51-55.

Brahma B., Refoufi A. (2015). Ontology Matching Algorithms. Communication présentée au Proceedings of the International Conference on Intelligent Information Processing, *Security and Advanced Communication*, New York, NY, USA : ACM, p. 89:1-89:5.

Cerami E. G., Gross B. E., Demi E., Rodchenkov I., Babur Ö., Anwa N., Shcultz N., Bader G. D., Sander C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1), D685-D690.

Cohen K. B. (2016, juin). Clinical language and scientific language: linguistic contrasts and ontological similarities. Communication présentée au Atelier IA & Santé, *27e journées francophones d'Ingénierie des Connaissances*.

Encode Project Consortium (2004). The ENCODE (ENCyclopedia of DNA elements) project. Science, 306, n° 5696, p. 636-640.

Cornet R., de Keizer N. (2008). Forty years of SNOMED: a literature review. *BMC Medical Informatics and Decision Making*, 8 (Suppl 1), S2.

Damerau F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, vol. 7, n° 3, p. 171-176.

De Morveau L.-B. G. (1787). *Méthode de nomenclature chimique.*

Dolin R. H., Huff S. M., Rocha R. A., Spackman K. A., Campbell K. E. (1998). Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies. *Journal of the American Medical Informatics Association*, vol. 5, n° 2, p. 203-213.

Dos Reis J. C., Pruski C., Reynaud-Delaître C. (2015). State-of-the-art on mapping maintenance and challenges towards a fully automatic approach. *Expert Systems with Applications*, vol. 42, n° 3, p. 1465-1478.

Dumontier M., Callahan A., Cruz-Toledo J., Ansell P., Emonet, V., Belleau F., Droit A. (2014). Bio2RDF Release 3: A Larger Connected Network of Linked Data for the Life Sciences. Communication présentée au *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, vol. 1272 Aachen, Germany, p. 401-404.

Euzenat J. et Shvaiko P. (2007). *Ontology matching* (vol. 333). Springer.

Fieschi M. (2009). La gouvernance de l'interopérabilité sémantique est au coeur du développement des systèmes d'information en santé (rapport public Publication 094000394).

Grosjean J., Merabti T., Dahamna B., Kergourlay I., Thirion B., Soualmia L. F., Darmoni S. J. (2011). Health multi-terminology portal: a semantic added-value for patient safety. *Stud Health Technol Inform*, 166, p. 129-138.

Hamdi F., Safar B., Reynaud C., Zargayouna H. (2010). Alignment-based partitioning of large-scale ontologies. *Advances in Knowledge Discovery and Management* Springer, p. 251-269.

Hettne K. M., van Mulligen E. M., Schuemie M. J., Schijvenaars B. J., Kors J. A. (2010). Rewriting and suppressing UMLS terms for improved biomedical term identification. *Journal of Biomedical Semantics*, vol. 1, n° 1, p. 5.

Hodge G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files.* ERIC.

IHTSDO et Regenstrief Institute. (juillet 2013). Regenstrief and the IHTSDO are working together to link LOINC and SNOMED CT. Repéré à https://loinc.org/collaboration/ihtsdo

Krauthammer M., Nenadic G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, vol. 37, n° 6, p. 512-526.

Lapage S. P., Sneath P. H. A., Lessel E. F., Skerman V. B. D., Seeliger H. P. R., Clark W. A. (Dir.). (1992). International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. Washington (DC) : ASM Press.

Lapatas V., Stefanidakis M., Jimenez R. C., Via A., Schneider M. V. (2015). Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki*, vol. 22, n° 1, p. 9.

Levenshtein V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Communication présentée au Soviet Physics Doklady, vol. 10, p. 707-710.

Liu H., Lussier Y. A., Friedman C. (2001). Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: An Unsupervised Method. J*ournal of Biomedical Informatics*, vol. 34, n° 4, p. 249-261.

Louie B., Mork P., Martin-Sanchez F., Halevy A., Tarczy-Hornoch P. (2007). Data integration and genomic medicine. *Journal of Biomedical Informatics*, vol. 40, n° 1, p. 5-16.

Macary F. (2007). IHDE, CDA et LOINC : des composants d'interopérabilité au service du partage des résultats de biologie médicale. *Spectra biologie*, vol. 26, n° 158, p. 51-57.

Mary M., Soualmia, L. F., Gansel, X. (2016, juin). Evaluation de la qualité des liens sémantique entre vocabulaires contrôlés. Communication présentée au Atelier SoWeDo, *27e Journées francophones d'Ingénierie des Connaissances*, Montpellier.

Mary M., Soualmia L. F., Gansel X. (2016, octobre). Projection des propriétés d'une ontologie pour la classification d'une ressource terminologique. Communication présentée aux *6e Journées francophones sur les ontologies*, Bordeaux.

McCray A. T. (1989). The UMLS Semantic Network. Communication. *Annual Symposium Proceedings on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*. American Medical Informatics Association, p. 503-507.

Merabti T., Grosjean J., Soualmia L. F., Joubert M., Darmoni S. J. (2012). Aligning biomedical terminologies in French: towards semantic interoperability in medical applications. INTECH Open Access Publisher.

Nelson S. J., Powell T., Humphreys L. B. (2006). The Unified Medical Language System (UMLS) of the National Library of Medicine, 61, p. 40-42.

Ogren P. V., Cohen K. B., Acquaah-Menah G. V., Eberlein J., Hunter L. (2004). The compositional structure of Gene Ontology terms. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 214-225.

Rahm E., Bernstein P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, vol. 10, n° 4, p. 334-350.

Robertson S. E., Jones K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, vol. 27, n° 3, p. 129-146.

Sahama T. R., Croll P. R. (2007). A data warehouse architecture for clinical data warehousing. Communication présentée au *Proceedings of the fifth Australasian symposium on ACSW frontiers*, vol. 68, Australian Computer Society, Inc., p. 227-232.

Sheide A., Wilson P. S. (2013). Reading up on LOINC. *Journal of AHIMA/American Health Information Management Association*, vol. 84, n° 4, p. 58-60.

Shvaiko P., Euzenat J. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics IV*. Springer, p. 146-171.

Shvaiko P., Euzenat J. (2008). Ten challenges for ontology matching. *On the Move to Meaningful Internet Systems: OTM 2008*, p. 1164-1182.

Sorrentino S., Bergamaschi S., Gawinecki M., Po L. (2010). Schema label normalization for improving schema matching. *Data & Knowledge Engineering*, vol. 69, n° 12, p. 1254-1273.

Stroetmann V. (2009). Semantic Interoperability for Better Health and Safer Healthcare. European Communities.

Turian J., Ratinov L., Bengio Y. (2010). Word representations: a simple and general method for semi-supervised learning. *Communication présentée au Proceedings of the 48th annual meeting of the association for computational linguistics* (p. 384–394), Association for Computational Linguistics.

Vreeman D. (7 novembre 2015). Guidelines for using LOINC and SNOMED CT Together. Daniel Vreeman. https://danielvreeman.com/guidelines-for-using-loinc-and-snomed-ct-together-without-overlap/

Winkler W. E. (1999). The state of record linkage and current research problems. Communication présentée au Statistical Research Division, US Census Bureau, Citeseer.

Yamamoto Y., Yamaguchi A., Bono H., Takagi T. (2011). Allie: a database and a search service of abbreviations and long forms. *Database: The Journal of Biological Databases and Curation*.

Zhou W., Torvik V. I., Smalheise N. R. (2006). ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, vol. 22, n° 22, p. 2813-2818.