Vers l'Alignement Complexe d'Ontologies par Grands Modèles de Langage : Modularisation Topologique et Verbalisation de Graphes

Igor Nascimento[1,4], Rinaldo Lima[2], Cassia Trojahn[3,4]

[1] Univ. Federal de Pernambuco, CIn, Recife, Brésil
[2] Universidade Federal Rural de Pernambuco, Brésil
[3] Univ. Grenoble Alpes, INRIA, CNRS, Grenoble INP, France
[4] Institut de Recherche en Informatique de Toulouse, France

ivln@cin.ufpe.br, rinaldo.jose@ufrpe.br, cassia.trojahn-dos-santos@univ-grenoble-alpes.fr

Résumé

L'alignement complexe d'ontologies demeure un défi majeur en raison de l'hétérogénéité structurelle et sémantique des ontologies du monde réel. Alors que les approches traditionnelles reposent principalement sur des règles ou des similarités lexicales, les approches récentes fondées sur des Grands Modèles de Langage démontrent un fort potentiel. Cet article propose une approche pour la génération automatique d'alignements complexes. Nous introduisons une chaîne de traitement d'alignement intégrant deux composants clés : (i) une stratégie de modularisation fondée sur des heuristiques topologiques combinant les algorithmes PageRank *et* Personalized PageRank *afin de réduire l'espace de recherche et de fournir un contexte sémantiquement pertinent, et (ii) un module de verbalisation transformant des modules ontologiques en descriptions textuelles exploitables par des modèles de langage. L'approche est évaluée sur plusieurs jeux de données de la piste complexe de l'Ontology Alignment Evaluation Initiative. Les résultats expérimentaux montrent que la combinaison des heuristiques topologiques avec la verbalisation permet d'obtenir des performances surpassant des systèmes de l'état de l'art.*

Mots-clés

Alignement d'ontologies complexes, Modularisation ontologique, Verbalisation ontologique.

Abstract

Complex ontology alignment remains one of the major challenges in Ontology Matching, due to the structural and semantic heterogeneity of real-world ontologies. While traditional approaches mainly rely on rules or lexical similarity measures, recent approaches based on large language models have shown strong potential. This paper proposes an approach for the automatic generation of complex ontology alignments. We introduce a pipeline integrating two key components : (i) a modularization strategy based on topological heuristics combining the PageRank and Personalized PageRank algorithms, aimed at reducing the search space while providing semantically relevant context ; and (ii) a verbalization module that transforms ontology modules into textual descriptions exploitable by language models. The proposed approach is evaluated on several datasets from the Ontology Alignment Evaluation Initiative Complex track. The experimental results show that integrating topological heuristics with verbalization techniques achieves performance that outperforms state-of-the-art Complex Ontology Matching systems.

Keywords

Complex Ontology Alignment, Ontology Modularization, Ontology Verbalization.

1 Introduction

L'alignement d'ontologies est la tâche consistant à identifier des correspondances entre des entités sémantiquement liées appartenant à différentes ontologies, facilitant ainsi l'interopérabilité et l'intégration de connaissances [37]. Traditionnellement, la majorité des systèmes d'alignement se sont concentrés sur des alignements simples, reliant une entité d'une ontologie source à une entité d'une ontologie cible. Cependant, les ontologies du monde réel présentent fréquemment une hétérogénéité riche dans la manière dont elles modélisent les connaissances (en termes de granularité, de relations ou de structures) qui ne peut pas toujours être conciliée par des correspondances un-à-un [30]. Les alignements complexes (correspondances 1 :N, N :1, et N :N avec opérateurs logiques et des transformations) offrent une représentation plus précise des connaissances, mais leur découverte automatique reste difficile. Ces alignements sont plus expressifs, car ils permettent de capturer des nuances telles que des chaînes de propriétés, des unions de classes ou des transformations de valeurs entre ontologies [22]. Par exemple, au lieu de tout simplement aligner des classes individuelles, une correspondance complexe peut établir qu'une classe d'une ontologie correspond à une restriction ou à une combinaison de classes dans une autre ontologie. Pour illustrer, étant données deux ontologies O_1 et O_2, voici quelques exemples :

$$\forall x,\ O_1 : \mathit{AcceptedPaper}(x) \equiv \exists y,\ O_2 : \mathit{acceptedBy}(x, y)$$

constitue une *correspondance complexe* faisant intervenir un *constructeur logique* (restriction de propriété) ;

$$\forall x, y,\ O_1 : \textit{authorOf}(x, y) \equiv \forall x, y,\ O_2 : \textit{writtenBy}(y, x)$$

constitue une *correspondance complexe* faisant intervenir un *constructeur logique* (propriété inverse) ;
3. Les valeurs de la propriété O_1 : *fullName* peuvent être obtenues par la concaténation des valeurs de O_2 : *firstName* et O_2 : *lastName* ; il s'agit d'une *correspondance complexe* reposant sur une *fonction de transformation de valeurs littérales* [22].
Dans ce contexte, cet article propose une nouvelle approche pour la génération d'alignements ontologiques complexes, fondée sur la *modularisation* et la *verbalisation* d'ontologies. La modularisation et la verbalisation constituent des composants centraux pour rendre l'alignement d'ontologies à la fois extensible, précis et sémantiquement informé. La modularisation permet de réduire l'espace de recherche en extrayant des sous-ontologies pertinentes autour d'entités candidates, tout en préservant le contexte structurel essentiel et en atténuant le bruit, ce qui est indispensable face à des ontologies de grande taille et fortement hétérogènes. La verbalisation, quant à elle, traduit les structures formelles (classes, propriétés et axiomes) en descriptions textuelles, permettant aux méthodes fondées sur la similarité sémantique et aux modèles de langage de capturer des nuances conceptuelles fines. Ces deux composants sont exploités à la fois lors de l'étape de génération de candidates et lors de la phase d'alignement dans les approches basées sur des Grands Modèles de Langage (*Large Language Models* - LLMs).
Les questions de recherche (QR) auxquelles cet article vise à répondre sont les suivantes :
QR1 : Dans quelle mesure l'utilisation de techniques de modularisation d'ontologies peuvent-elles améliorer les performances du système ?
QR2 : La verbalisation de modules en langage naturel améliore-t-elle la qualité des alignements (F1-score), lorsque ses sorties sont utilisées pour la génération de plongements et comme contexte dans des prompts destinés aux LLMs, par rapport à l'utilisation directe des étiquettes et URI des ontologies ?
Les **contributions principales** de l'article peuvent être résumées comme suit :
1. Une nouvelle approche pour l'alignement d'ontologies complexes fondée sur des LLMs, combinant modularisation topologique et verbalisation ontologique au sein d'une chaîne de traitement unifié, surpassant les systèmes de l'état de l'art sur plusieurs jeux de données de la piste *Complex de l'Ontology Alignment Evaluation Initiative* (OAEI).
2. une technique de modularisation automatique intégrant les algorithmes PageRank et Personalized PageRank (PPR) [2], permettant de réduire l'espace de recherche tout en fournissant un contexte sémantique ciblé et pertinent pour la tâche d'alignement complexe ;
3. une méthode de verbalisation d'ontologies en langage naturel qui transforme des modules ontologiques formels en descriptions textuelles, exploitables à la fois pour la génération de plongements et comme contexte dans des prompts destinés aux LLMs, améliorant ainsi la qualité des alignements complexes générés.
Le reste de l'article est organisé comme suit. La Section 2 présente les travaux connexes, avec un accent particulier sur les alignements complexes et les méthodes fondées sur des modèles de langage. La Section 3 décrit l'approche proposée. La Section 4 est consacrée à l'évaluation expérimentale. La Section 5 rapporte les résultats obtenus et analyse les performances du système proposé en comparaison avec les approches de l'état de l'art. La Section 6 discute ces résultats au regard des questions de recherche. Enfin, la Section 7 conclut l'article et ouvre des perspectives pour les travaux futurs.

2 Travaux connexes

La recherche en alignement d'ontologies a avancé grâce à des techniques efficaces pour les alignements simples, fondées sur la similarité lexicale, la similarité structurelle et l'apprentissage automatique [37]. Toutefois, l'identification de correspondances complexes demeure le sous-domaine le plus difficile nécessitant fréquemment des stratégies plus avancées [6, 38, 32].

2.1 Approches classiques d'Alignement complexe d'ontologies

CANARD [32] est un système d'alignement complexe guidé par des *Competency Questions for Alignment* (CQA), soumises sous forme de requêtes SPARQL basées sur l'ontologie source. Les correspondances sont générées en projetant le sous-graphe de la CQA source vers des voisinages lexicalement similaires des instances correspondantes dans l'ontologie cible. Les évaluations montrent que CANARD favorise la couverture au détriment de la précision, tout en étant l'un des rares systèmes capables de traiter des bases de connaissances à grande échelle. Ses performances restent néanmoins fortement dépendantes de la densité des données et de la qualité des liens *sameAs* entre instances. Les auteurs notent que l'intégration de contre-exemples améliore significativement la précision, au prix d'un temps d'exécution plus élevé.
AROA [38] génère automatiquement des alignements simples et complexes entre ontologies partageant des instances communes. Le système extrait des triplets pour construire une base de transactions typées, sur laquelle l'algorithme FP-growth fouille des règles d'association, ensuite filtrées par des patrons de correspondance simples et complexes (tels que l'équivalence de classes et le patron *Class by Attribute Type*).

2.2 Approches basées sur les modèles de langage

Matcha [6] est un système d'alignement d'ontologies étendant l'architecture d'*AgreementMakerLight* (AML) [5] pour traiter les correspondances complexes et holistiques. Sa contribution centrale repose sur l'intégration du modèle pré-entraîné *sentence-BERT* [25], qui génère des plon-

gements à partir des étiquettes et synonymes des entités, permettant de capturer un contexte sémantique au-delà des déclarations explicites de l'ontologie. L'alignement est ensuite réalisé par calcul de similarité cosinus.

Les premières explorations fondées sur les architectures *Transformer* ont mis en évidence leur potentiel. Le système DITTO [19] s'est distingué comme l'un des premiers à exploiter des modèles basés sur *Bidirectional Encoder Representations from Transformers* (BERT) [3] pour déterminer les correspondances entre entités ontologiques. De son côté, *BERTMap* [15] a intégré l'ajustement fin (*fine-tuning*) d'encodeurs *Transformer* dans des tâches d'alignement, en le combinant à des mécanismes de réparation des alignements afin d'assurer leur cohérence.

Des travaux ont ensuite exploré l'utilisation de LLMs génératifs en mode *zero-shot* et *few-shot*. Les auteurs de [24] ont notamment appliqué ChatGPT[1] à une paire d'ontologies de petite taille (piste *Conference* de l'OAEI), obtenant un rappel acceptable mais une précision insuffisante, avec un *F1-score* inférieur à celui d'un *système de base* fondé sur de simples similarités de chaînes de caractères. Ce résultat illustre la tendance des LLMs non contraints à générer un excès de faux positifs [24, 16].

Des travaux plus récents montrent que les LLMs, lorsqu'ils sont correctement guidés, peuvent atteindre des performances remarquables. Les auteurs de [36] ont ainsi démontré que GPT-3.5 [4] et GPT-4 [23], combinés à des stratégies de récupération par plongements, surpassent l'état de l'art en alignement d'ontologies biomédicales.

De même, LLMs4OM [11] a montré que GPT-3.5 et LLaMA-2 [33] peuvent égaler, voire dépasser, les systèmes d'alignement traditionnels sur plusieurs pistes de l'OAEI grâce au *retrieval-augmented prompting* [18].

Les LLMs, associés à une conception rigoureuse des *prompts* et à des stratégies de recherche efficaces, s'imposent comme des outils prometteurs pour l'alignement d'ontologies. Dans cette perspective, [1] exploite la méthodologie MOMo (*Modular Ontology Modeling*) [28], qui intègre des modules ontologiques pour fournir un contexte sémantique structuré. La démarche consiste à charger l'ontologie source comme *prompt* initial, puis à introduire les entités cibles pour vérifier les alignements, en enrichissant le *prompt* par des descriptions *few-shot*, des stratégies de *Chain-of-Thought* et des axiomes centraux des modules. Cette approche atteint de bons niveaux de précision et surpasse les tentatives de *prompting* direct sans modularisation. En revanche, elle présente deux limitations principales : la modularisation reste manuelle et le chargement d'ontologies entières dans les *prompts* engendre un temps de traitement plus élevé.

CMatch [27] améliore l'approche de [1] en introduisant une modularisation automatique par l'algorithme PageRank [2], améliorant ainsi le passage à l'échelle du processus d'alignement. Le système s'appuie sur une stratégie few-shot et sélectionne, par similarité cosinus, les k modules cibles les plus pertinents comme candidats à la génération.

1. https ://openai.com/index/chatgpt/

Le présent travail s'inscrit dans cette continuité en enrichissant la modularisation par le PPR, afin de fournir un contexte plus ciblé et de capturer des entités plus pertinentes pour *l'appariement*. Un module de verbalisation des triplets, inspiré de [7], est également intégré afin de fournir aux modèles de langage une représentation textuelle explicite des graphes ontologiques, favorisant une meilleure compréhension sémantique des ontologies.

3 Approche proposée

Selon [8], la majorité des systèmes d'alignement suivent une chaîne de traitement similaire composée de trois étapes : prétraitement, génération de candidats et post-traitement. L'approche proposée enrichit cette architecture par deux composants supplémentaires : la modularisation et la verbalisation. La Figure 1 présente une vue d'ensemble de l'architecture, dont chaque composant est détaillé ci-dessous.

3.1 Composant de prétraitement

Lors du prétraitement, les ontologies sont chargées puis nettoyées afin d'affiner les données utilisées lors de l'alignement. Les opérations effectuées sont : la suppression des instances (l'alignement étant réalisé uniquement au niveau du schéma des ontologies), la suppression de noeuds anonymes (*BNodes*) pour simplifier la représentation de l'ontologie, et la suppression des relations *disjointWith*. La dernière opération est motivée par l'étape de modularisation : les relations *disjointWith* peuvent nuire aux performances de PageRank en conduisant à la construction de modules contenant des entités sémantiquement disjointes. Par exemple, si A est déclaré *disjointWith* B, ces deux entités ne devraient pas coexister au sein d'un même module. Dans certaines ontologies, des entités peuvent être associées à un grand nombre de relations *disjointWith*, ce qui peut amener PageRank à leur attribuer une importance excessive. Or, selon la définition d'un module proposée par [28], seules les entités sémantiquement liées à l'entité centrale du module sont pertinentes.

3.2 Composant de modularisation

L'étape de modularisation s'inspire des travaux de [1] et [30]. [1] exploitent les modules ontologiques [28] pour réaliser des alignements complexes à l'aide de LLMs, en supposant que la modularisation facilite leur génération. En revanche, leur approche nécessite le chargement de l'intégralité de l'ontologie cible pour identifier les alignements à partir des modules. À l'inverse, [30] réduisent l'espace de recherche en s'appuyant sur PageRank [2] pour identifier les entités candidates représentant les centres des modules (centroïdes), combiné à un parcours en largeur (*Breadth-First Search* - BFS) pour sélectionner les voisins les plus proches de ces centroïdes jusqu'à une profondeur de 1.

Dans notre approche, l'algorithme PageRank est utilisé comme une heuristique d'importance globale : les concepts présentant un score de PageRank élevé tendent à être connectés à de nombreux autres concepts ou à des concepts eux-mêmes centraux, jouant ainsi le rôle de noeuds cen-

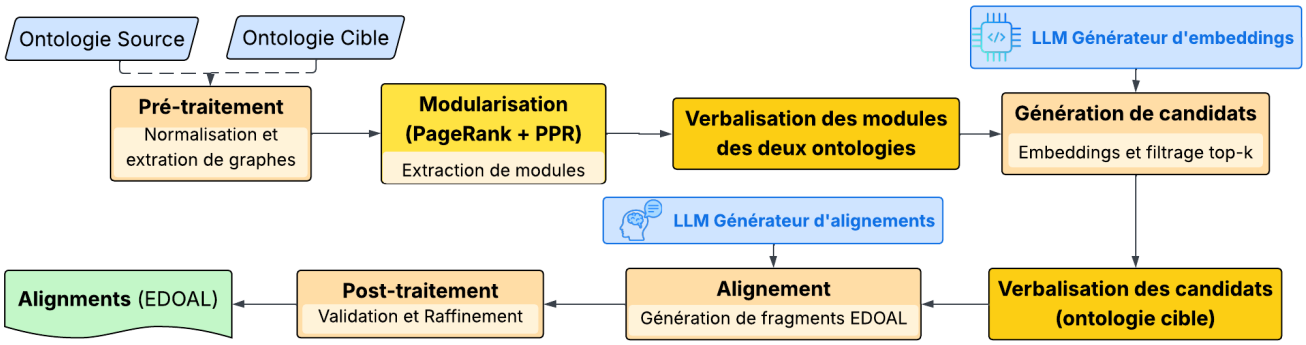

FIGURE 1 – Vue d'ensemble de la chaîne de traitement d'alignement d'ontologies complexes proposée.

traux sémantiques au sein de l'ontologie. Ces valeurs de centralité globale peuvent guider le processus d'alignement en mettant en évidence des entités potentiellement plus pertinentes ou plus influentes. Par exemple, des classes fortement centrales dans les deux ontologies peuvent être considérées comme des points de départ pour la recherche de correspondances. De même, les scores de PageRank peuvent être exploités comme des poids dans la combinaison de mesures de similarité, en accordant davantage d'importance aux correspondances impliquant des concepts centraux.

Proposition et justifications. Une des contributions de cet article repose sur deux axes principaux : (i) la suppression des relations *disjointWith* dans les ontologies, et (ii) l'enrichissement du contexte associé à une entité source par l'utilisation de l'algorithme de PPR [2].

Alors que le PageRank répond à la question « quelle est l'importance du nœud x dans l'ensemble du graphe ? », le PPR permet de répondre à « quelle est la pertinence du nœud x par rapport à un nœud y donné ? ». Dans le contexte de l'alignement d'ontologies, l'objectif est que le module extrait pour décrire un concept contienne des informations capables de le discriminer sémantiquement. Cette approche permet ainsi que le « contexte » fourni au LLM ne soit pas pollué par des concepts génériques, mais qu'il soit principalement composé de voisins qui contribuent fortement à la définition sémantique de l'entité considérée. Par exemple, pour le concept « Person » dans le domaine de la conférence, le PPR privilégierait des concepts tels que « Author » ou « Reviewer » plutôt que des concepts plus généraux comme « Entité physique ». De cette manière, la construction des modules devient plus riche et plus pertinente sur le plan sémantique.

Le PPR vise à améliorer l'étape de modularisation de l'ontologie source en le combinant au PageRank. Alors que le PageRank classique attribue un score de centralité fondé sur des parcours aléatoires uniformes à travers l'ontologie, le PPR oriente ces parcours afin de privilégier certaines parties du graphe, à partir d'un ensemble de nœuds d'intérêt, appelés *seeds* ou nœuds semences [17]. Autrement dit, le PPR calcule une pertinence contextuelle des nœuds : ceux qui sont le plus fortement connectés, directement ou indirectement, aux nœuds centraux reçoivent un score plus élevé, reflétant leur importance relative au sein de ce contexte spécifique. Dans la littérature, les méthodes fondées sur le PPR se sont révélées efficaces pour prioriser des informations pertinentes dans des graphes de connaissances. Ainsi, [17] ont intégré le PPR à des mesures de similarité sémantique pour réaliser des tâches d'*entity linking* dans des ontologies biomédicales, obtenant des gains significatifs en termes de précision.

La Figure 2 illustre le fonctionnement du composant de modularisation au sein de l'architecture proposée.

3.3 Composant de verbalisation

La seconde contribution introduit un module de verbalisation dont l'objectif est de traduire certaines parties sélectionnées des ontologies, les modules extraits lors de l'étape précédente, en descriptions en langage naturel. Dans le travail de [30], le module extrait est représenté et manipulé en syntaxe *Turtle*. Bien que ces formats formels soient exploitables par les machines, ils ne capturent pas directement une signification textuelle exploitable comme entrée par des modèles de langage [9]. En ce sens, nous proposons de transformer ces modules en un ensemble de phrases ou de paragraphes exprimant explicitement les faits et axiomes qu'ils contiennent.

À titre d'exemple, supposons qu'un module contienne un concept *Patient*, défini comme une sous-classe de *Person* et possédant une propriété *hasDisease* le reliant à des instances de *Disease*. Une verbalisation appropriée pourrait être : « A Patient is a kind of Person that has at least one Disease (through the relation 'hasDisease') ». De la même manière, lorsque des restrictions ou des équivalences sont présentes, elles peuvent être transformées en phrases suivant des patrons linguistiques prédéfinis, par exemple : « Every Cardiac Disease is a Disease that affects the heart », afin de représenter une définition textuelle.

Il existe dans la littérature des méthodologies établies pour la verbalisation d'ontologies, reposant sur l'utilisation de *templates* prédéfinis pour les structures *Web Ontology Language* (OWL), telles que les classes équivalentes ou les restrictions [34]. Des outils et des langages contrôlés, comme *Attempto Controlled English* (ACE) [10], ou des approches fondées sur les *Lexicon Models for Ontologies* (LEMON) [35], constituent également des sources d'inspiration pour la génération de phrases grammaticalement correctes à partir d'axiomes ontologiques.

Le module de verbalisation proposé apporte deux bénéfices

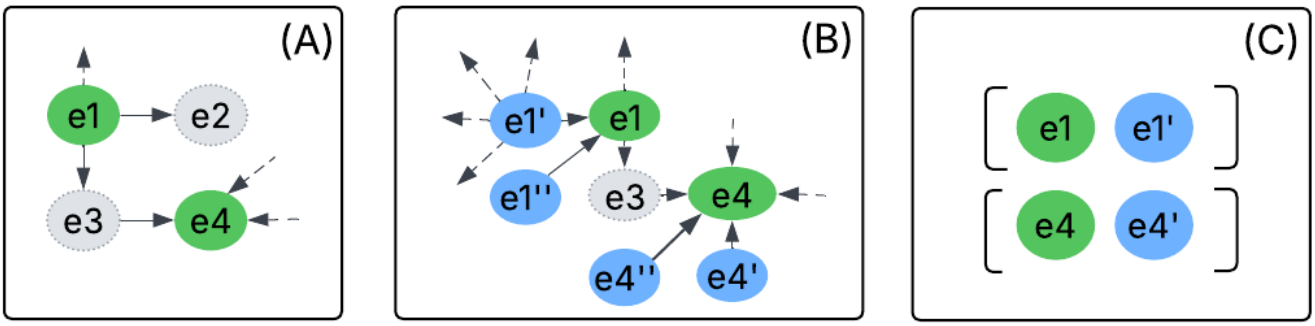

FIGURE 2 – Le flux (A) correspond à la capture des centroïdes (en vert). Le flux (B) représente la capture des nœuds les plus importants associés à ces centroïdes (en bleu). Le flux (C) correspond à la sélection finale des nœuds à fort degré (nombreuses arêtes incidentes), utilisés pour la modularisation de l'ontologie source.

principaux au chaîne de traitement d'alignement :

Génération de plongements textuels enrichis. Les descriptions en langage naturel des entités et de leur contexte peuvent être exploitées pour produire des plongements sémantiques plus expressifs. Contrairement à une représentation fondée uniquement sur le nom d'une classe, l'ensemble du texte descriptif peut être fourni à un modèle de langage, par exemple, un encodeur *Transformer* de type BERT afin d'obtenir une représentation vectorielle dense capturant des aspects subtils du sens [25]. On s'attend ainsi à ce que deux concepts équivalents appartenant à des ontologies différentes présentent des descriptions textuelles similaires, conduisant à des vecteurs proches dans l'espace des plongements. La comparaison de plongements issus de verbalisation augmente donc la probabilité d'identifier des correspondances qui ne seraient pas détectables par une simple comparaison lexicale.

Contextualisation des prompts pour les LLMs. La verbalisation sert également à constituer le contexte fourni au modèle de langage lors des étapes ultérieures. L'idée est d'inclure des descriptions en langage naturel des concepts et relations les plus pertinents, ce qui améliore les performances des LLMs dans les tâches d'alignement [16, 29, 31]. Par exemple, fournir au LLM des phrases telles que « Patient is a kind of Person that has a disease » peut l'aider à reconnaître qu'un concept nommé *Patient* dans une autre ontologie, décrit comme « Patient is a kind of Person who suffers from some disease », correspond très probablement au même concept, et non à l'adjectif *patient*, réduisant ainsi les ambiguïtés sémantiques.

Proposition et justifications. Comme illustré dans la Figure 1, le composant de verbalisation intervient sur les sorties des composants de modularisation et de génération de candidats. Lorsque les modules sont produits par le composant de modularisation, ceux-ci sont verbalisés afin de servir d'entrée au composant de génération de candidats, où ils sont utilisés comme texte pour la génération de plongements, puis pour le calcul de similarité. Pour chaque module source, un ensemble de modules candidats issus de l'ontologie cible est généré ; ces modules candidats sont ensuite regroupés en un module unique, à partir duquel une nouvelle étape de verbalisation est effectuée.

```
http://confOf#Reviewing_event type owl:Class
http://confOf#Reviewing_event subClassOf http://confOf#Administrative_event
                                (a)
Reviewing_event type Class
Reviewing_event subClassOf Administrative_event
                                (b)
reviewing event is type of class
reviewing event subclass of administrative event
                                (c)
```

FIGURE 3 – Types de verbalisation utilisés.

Par ailleurs, l'utilisation des LLMs pour des tâches sémantiques a favorisé l'émergence d'approches dans lesquelles les ontologies sont décrites sous forme textuelle à destination du modèle. Des prototypes récents, tels que OLaLa [16], ont appliqué des LLMs en modes *zero-shot* et *few-shot*, en fournissant uniquement des étiquettes et des descriptions simples des classes, obtenant ainsi des résultats prometteurs. Notre proposition s'inscrit dans cette continuité en générant automatiquement des descriptions plus riches et plus contextualisées.

Ainsi, la verbalisation agit comme un pont entre le formalisme symbolique des ontologies et les capacités linguistiques des modèles de langage. Le présent travail s'inspire en particulier de l'approche proposée dans DeepOnto [14]. DeepOnto repose sur un algorithme basé sur des *templates* qui convertit des construits OWL en phrases naturelles en anglais à l'aide de patrons linguistiques prédéfinis. Chaque triplet ⟨subject, predicate, object⟩ est ainsi associé à une structure linguistique spécifique, que nous désignons comme un *template*. Par exemple, dans DeepOnto, la propriété *SubClassOf* est traduite par l'énoncé « A is a type of B ».

La Figure 3 illustre les trois stratégies de verbalisation proposées. La verbalisation *base* (a) exprime l'ontologie sous forme de triplets en conservant les *Internationalized Resource Identifier* (IRI) complets de chaque entité. La verbalisation *label* (b) remplace les IRI par les étiquettes qui en sont dérivés. Enfin, la verbalisation *naturelle* (c) reformule les triplets en langage naturel, en normalisant les noms composés en minuscules et en séparant leurs termes par des espaces.

Dans le contexte de l'alignement d'ontologies, des recherches antérieures suggèrent que l'enrichissement des

ontologies par des descriptions textuelles (par exemple des définitions extraites d'encyclopédies ou de dictionnaires, ou encore des descriptions générées automatiquement) peut accroître la couverture et la précision de l'alignement, en particulier lorsqu'il est combiné à des modèles de traitement automatique du langage naturel [13, 29].

3.4 Composant de génération de candidats

Ce composant s'appuie sur un modèle LLM pour générer des plongements des modules issus des ontologies source et cible, dont les similarités sont calculées par similarité cosinus, conduisant à la construction d'une matrice de similarité. Pour chaque centroïde d'un module de l'ontologie source identifié par PageRank, un ensemble de nœuds de l'ontologie cible est sélectionné puis verbalisé. Les k modules cibles les plus similaires au module source, selon la similarité cosinus, sont retenus comme candidats à l'alignement. Les modules candidats ainsi récupérés sont ensuite fournis comme contexte au modèle de langage, qui génère les alignements correspondants. Le paramètre k est un hyperparamètre contrôlant le compromis entre couverture des candidats et complexité computationnelle.

3.5 Composant d'alignement

Le composant d'alignement repose sur un LLM de raisonnement, capable de décomposer des problèmes complexes en étapes intermédiaires, chargé de générer les alignements à partir des informations fournies dans le *prompt* au format *Expressive and Declarative Ontology Alignment Language* (EDOAL). Une stratégie *few-shot* est adoptée, fournissant un nombre limité d'exemples pour guider le modèle. La stratégie de *prompting* suit celle de [30], où deux exemples de paires de modules ontologiques sont fournis au format *Turtle*, adaptés selon le type de verbalisation employé. La structure générale du *prompt* suit le travail de [27].

3.6 Composant de post-traitement

Le composant de post-traitement agrège les sorties du LLM au format EDOAL en un document unique contenant l'ensemble des alignements générés, puis applique une phase de raffinement afin de corriger les erreurs introduites par le modèle.

Il a été observé que, dans les configurations utilisant les verbalisations *label* et *naturelle*, le LLM générait fréquemment des entités avec des IRI reprenant le préfixe générique `http://example.org/ontology1/`, identique à celui utilisé dans les exemples *few-shot* du prompt. Pour remédier à ce problème, deux dictionnaires associant les étiquettes aux IRI d'origine sont générés lors de l'étape de verbalisation, un pour chaque ontologie. Après génération du document EDOAL, les étiquettes produites par le LLM sont remplacées par les IRI correspondants à l'aide de ces dictionnaires. La Figure 4 illustre ce processus de raffinement.

4 Évaluation expérimentale

Cette section présente les expérimentations menées afin d'évaluer les questions de recherche **QR1** et **QR2**.

4.1 Configuration expérimentale des modules

Les configurations expérimentales de chaque composant de l'architecture proposée en Figure 1 sont détaillées ci-dessous.

Composant de prétraitement. Les ontologies ont été chargées à l'aide des bibliothèques *rdflib* [2] et *networkx* [3]. L'ontologie est d'abord chargée dans *rdflib*, puis traduite dans une représentation de graphe *networkx*. Cette étape de traduction est motivée par le fait que *networkx* met à disposition un ensemble d'algorithmes particulièrement utiles pour le traitement et l'analyse de graphes.

Composant de modularisation. Les algorithmes PageRank et PPR ont été configurés avec un facteur d'amortissement (*damping factor*) de 0,8 et 100 itérations de calcul.

Composant de verbalisation. Le composant de verbalisation s'appuie sur OntoAligner [12], qui génère des triplets à partir des ontologies ainsi que les étiquettes associées aux entités ontologiques.

Composant de génération de candidats. Les paramètres du composant de génération de candidats sont identiques à ceux de CMatch. Le modèle de plongements utilisé est Qwen3-Embedding-8B [4], avec un *batch size* de 2 et une longueur maximale de contexte de 8192 jetons.

Composant d'alignement. Le modèle Qwen3-14B [5] est utilisé pour la phase d'alignement (matcher), avec une longueur maximale de contexte de 8192 jetons, une température de 1,0 et un *seed* fixe garantissant la reproductibilité des sorties.

4.2 Jeux de données

Nous utilisons plusieurs jeux de données de l'OAEI issus de la piste *Complex*, notamment *Conference*, *Hydrography*, *GeoLink* et *Taxon* [6].

Au total, 15 ontologies ont été utilisées, correspondant à 17 paires d'ontologies considérées pour les expérimentations [7] : 10 paires issues de *Conference* (*cmt–conference*, *cmt–confOf*, *cmt–edas*, *cmt–ekaw*, *conference–confOf*, *conference–edas*, *conference–ekaw*, *confOf–edas*, *confOf–ekaw*, *edas–ekaw*) ; 1 paire issue de *GeoLink* (*gbo–gmo*) ; 3 paires issues de *Hydrography* (*hydro3–swo*, *hydrOntology–swo*, *cree–swo*) ; et 3 paires issues de *Taxon* (*taxon–agrovoc*, *taxon–dbpedia* et *taxon–taxref*).

Le Tableau 1 présente les statistiques relatives au nombre d'ontologies, de triplets, ainsi que d'alignements simples et complexes pour chacun des jeux de données considérés.

Les combinaisons expérimentales suivantes ont été évaluées : **Topologie** : avec ou sans PPR ; **Profondeur de voisinage** : 1 et 2 ; **Nombre de candidats générés** : 2, 3 et 5 ; **Verbalisation** : *base*, *label* et *naturelle*.

2. https ://rdflib.readthedocs.io/en/stable/
3. https ://networkx.org/en/
4. https ://huggingface.co/Qwen/Qwen3-Embedding-8B
5. https ://huggingface.co/Qwen/Qwen3-14B
6. https ://oaei.ontologymatching.org/2025/complex/index.html
7. https ://github.com/liseda-lab/complex-OM-benchmark/tree/main

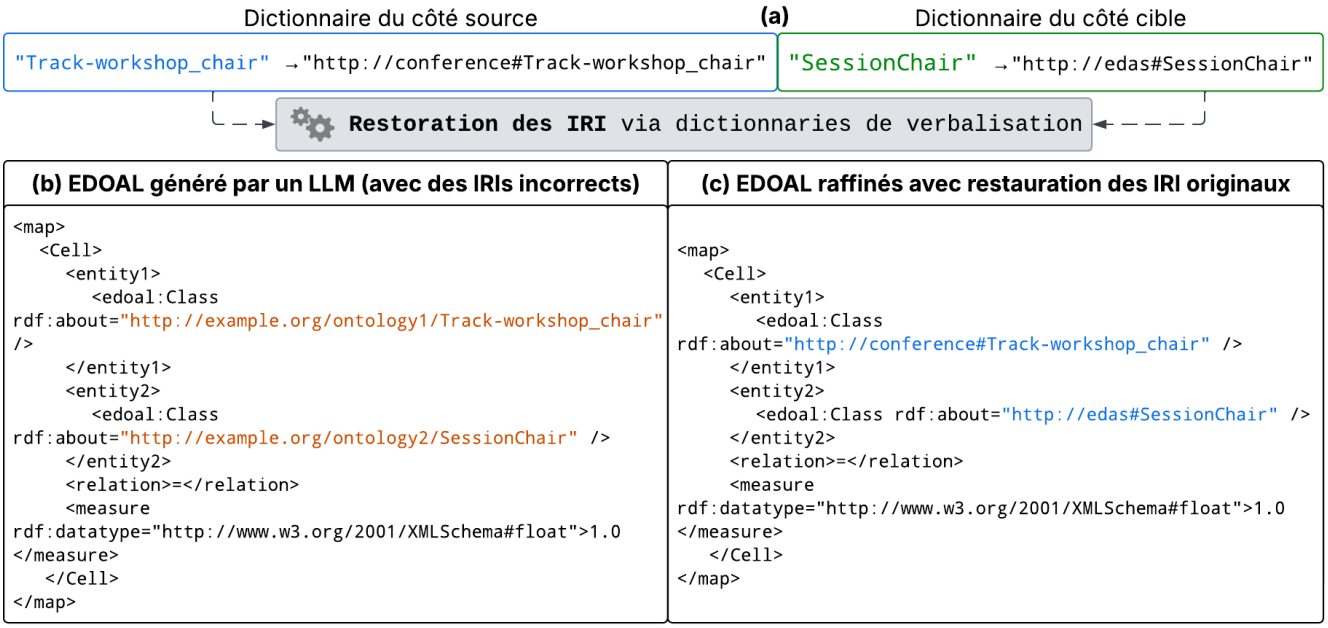

FIGURE 4 – Deux dictionnaires (un par ontologie) associent les verbalisations aux IRI d'origine (a) ; chaque verbalisation est recherchée dans le document généré (b), puis remplacée par son IRI d'origine (c).

TABLE 1 – Résumé des jeux de données utilisés

Jeu de données	Conf.	Geo	Hydro	Taxon
Ontologies	5	2	5	4
Triples	3867	2720	17196	776690
Simples	111	19	113	6
Complexes	184	48	84	20

Le système proposé est comparé aux systèmes participants de la piste *Complex* de l'OAEI : CMatch (Conference, Hydrography, GeoLink et Taxon), CANARD(Conference, GeoLink et Taxon), Matcha (Hydrography) et AROA (GeoLink).

4.3 Modèles de language et métriques d'évaluation

Modèles de language. Les modèles de plongements ont été sélectionnés sur la base du *Massive Text Embedding Benchmark* (MTEB) [21]. Le modèle Qwen3-Embedding-8B a été retenu, figurant parmi les trois meilleurs modèles à code libre selon ce benchmark. Le modèle de raisonnement utilisé pour la tâche d'alignement est Qwen3-14B, chargé via l'API *HuggingFace*[8].

Métriques d'évaluation. Les métriques d'évaluation sont celles adoptées par l'OAEI pour les jeux de données considérés, en particulier la *Tree Edit Distance* (TED) [26].

5 Résultats expérimentaux

Les résultats sont présentés comme suit : la combinaison obtenant le meilleur *F1-score* est d'abord sélectionnée, puis sa contrepartie avec modularisation PPR est examinée (ou, si cette stratégie est déjà intégrée, la version sans PPR lui est substituée).

8. https ://huggingface.co/models

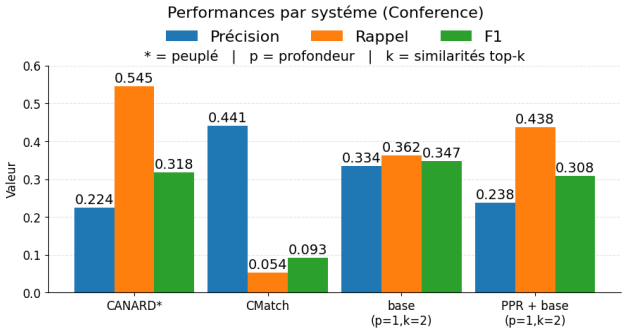

FIGURE 5 – Résultats sur le jeu de données *Conference*. *=peuplé, indique que le système a utilisé les instances des classes et propriétés

Pour chaque jeu de données, les résultats sont moyennés sur l'ensemble des paires d'ontologies le composant : la précision et le rappel sont calculés par moyenne arithmétique, puis le *F1-score* est dérivé comme leur moyenne harmonique.

La Figure 5 présente les résultats sur le jeu de données *Conference*, comparés aux systèmes CMatch et CANARD. La configuration la plus performante combine la verbalisation *base*, une exploration BFS de profondeur 1, sans PPR, surpassant CANARD et CMatch sur ce jeu de données.

La Figure 6 présente les résultats sur le jeu de données *Hydrography*, comparés aux performances des systèmes CMatch et Matcha. Le système proposé surpasse ces deux références avec des gains de 15 à 20 points de pourcentage, obtenus avec la verbalisation *base*, une exploration BFS de profondeur 2 et sans PPR.

La Figure 7 présente les résultats sur le jeu de données *GeoLink*, comparés aux performances des systèmes CANARD, CMatch et AROA. Le système proposé dépasse CMatch de

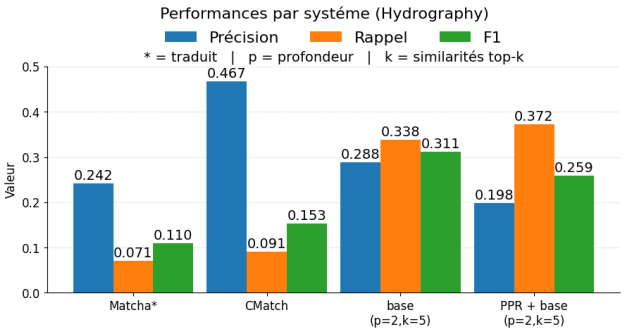

FIGURE 6 – Résultats sur le jeu de données *Hydrography* *=traduit, indique que le système utilise une traduction en anglais des entités de l'ontologie.

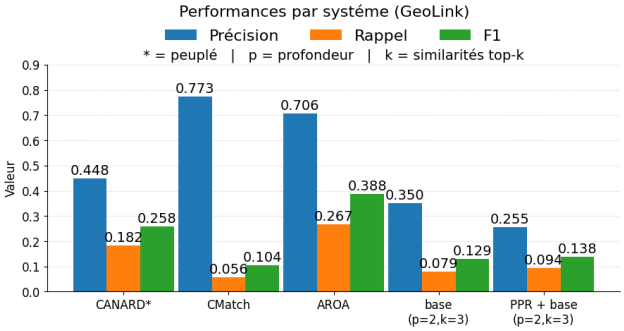

FIGURE 7 – Résultats sur le jeu de données *GeoLink*

.

3 à 4 points de pourcentage, mais reste en retrait par rapport à CANARD et AROA, avec la même configuration : verbalisation *base*, BFS de profondeur 2 et sans PPR.

La Figure 8 présente les résultats sur le jeu de données *Taxon*, comparés aux performances des systèmes CMatch et CANARD. Notre système surpasse CMatch et améliore d'environ deux points de pourcentage les performances de CANARD, avec la verbalisation *base*, une profondeur de voisinage 1 et sans PPR.

Dans la Figure 9, la moyenne des résultats sur l'ensemble des jeux de données (*Conference*, *Hydrography*, *GeoLink* et *Taxon*) est comparée entre les systèmes. Notre approche dépasse CMatch en moyenne de plus de 20 points de pourcentage, avec la verbalisation *base*, une exploration BFS de profondeur 1 et sans PPR.

La Figure 10 présente la moyenne des résultats sur le jeux de données *Conference*, *GeoLink* et *Taxon*, comparée aux performances du système CANARD.

Malgré un *F1-score* moyen inférieur à CANARD sur les jeux de données *Conference*, *GeoLink* et *Taxon*, il convient de souligner que CANARD requiert la présence d'instances pour effectuer l'alignement. À l'inverse, le système proposé s'appuie uniquement sur les informations de schéma des ontologies, sans recourir aux instances, ce qui lui confère un avantage notable en termes de généricité et d'applicabilité, imposant moins de prérequis sur les ontologies d'entrée. Par ailleurs, le système proposé obtient de meilleurs *F1-scores* sur les jeux de données *Conference* et *Taxon* lorsqu'ils sont considérés individuellement.

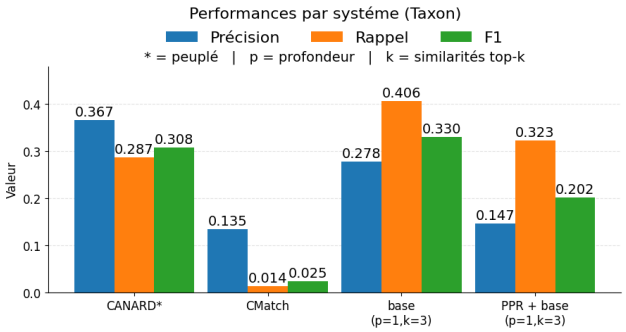

FIGURE 8 – Résultats sur le jeu de données *Taxon*

.

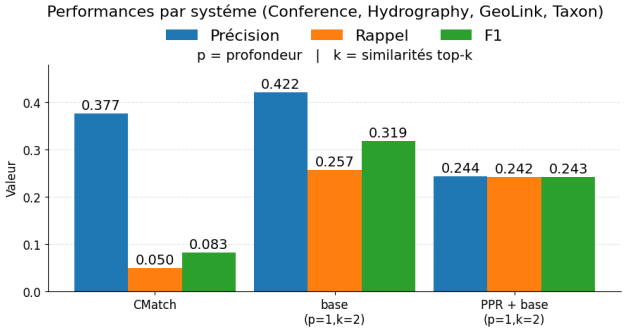

FIGURE 9 – Comparaison des performances moyennes avec CMatch (*Conf.*, *Hydrog.*, *GeoL.* et *Taxon*).

Les résultats montrent que la verbalisation *base* offre les meilleures performances globales, en combinaison avec une exploration BFS de profondeur 1, et sans PPR. Concernant la verbalisation, la configuration *base* produit de meilleurs résultats car le LLM tend à reproduire fidèlement la structure fournie dans les exemples du *prompt*. Cet aspect est déterminant, l'évaluation reposant sur une comparaison stricte entre les alignements produits au format EDOAL et les alignements de référence, qui exige une correspondance exacte des entités alignées.

Concernant la modularisation, l'intégration du PPR produit un *F1-score* inférieur à la configuration sans PPR, probablement en raison du volume accru d'informations soumises au LLM, ce qui dégrade la qualité des alignements. En effet, les modèles de langage tendent à perdre le focus sur les informations centrales face à des *prompts* excessivement longs ou denses (phénomène *Lost in the Middle*) [20]. Des travaux futurs porteront sur les stratégies de *prompting* appliquées au modèle de langage.

6 Discussion

Les résultats montrent que la verbalisation *base* obtient les meilleures moyennes de *F1-score* sur l'ensemble des paires d'ontologies évaluées. Les configurations intégrant le *PPR* présentent quant à elles un rappel élevé au détriment de la précision, ce qui s'explique par plusieurs facteurs : premièrement, le contexte supplémentaire induit par le PPR peut provoquer une surcharge informationnelle du *prompt*, amenant le LLM à atteindre ou dépasser sa fenêtre de contexte maximale ; deuxièmement, face à un volume d'information

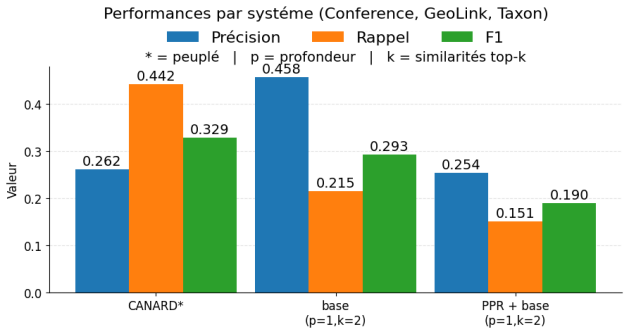

FIGURE 10 – Comparaison des performances moyennes avec CANARD (jeux de données *Conf.*, *GeoL.* et *Taxon*)

trop important, le modèle tend à ne pas exploiter efficacement tous les éléments pertinents, entraînant une perte d'informations.

En réponse à **QR1**, le PPR exerce un double effet : il enrichit le contexte fourni au LLM, mais peut simultanément nuire à la précision des alignements en introduisant un excès d'information. Le seul jeu de données pour lequel la configuration avec PPR s'avère bénéfique est *GeoLink*, probablement en raison de sa forte spécificité et de sa structuration particulière.

Concernant **QR2**, les stratégies de verbalisation se sont révélées globalement prometteuses, surpassant les systèmes de l´état de l´art dans plusieurs scénarios. La verbalisation *base* présente en particulier une amélioration cohérente du rappel, attribuable à l'hypothèse que ce comportement soit lié à la présence d'un préfixe d'IRI commun qui augmente les scores de similarité textuelle. Les IRI partagent des jetons communs, tels que « https » ou d'autres éléments propres à la structure des IRI, qui ne sont pas pertinents pour l'alignement ontologique. Cette similarité tend à augmenter le rappel, non pas en raison d'une véritable proximité sémantique entre les concepts, mais du fait de similarités entre les chaînes de caractères, au détriment d'un alignement fondé sur les relations conceptuelles des entités ontologiques. Les performances inférieures de la verbalisation *label* ne reflètent donc pas nécessairement une moindre correspondance sémantique, mais plutôt l'absence de cet avantage structurel.

7 Conclusion

Cet article a présenté une nouvelle approche pour l'alignement complexe d'ontologies fondée sur des LLMs, intégrant deux composants complémentaires : une modularisation topologique et un module de verbalisation transformant des structures ontologiques formelles en descriptions en langage naturel. L'évaluation menée sur plusieurs jeux de données de la piste Complex de l'OAEI montre que le système proposé surpasse, dans la majorité des scénarios, les systèmes de l'état de l'art. Ces résultats indiquent que la combinaison de la modularisation et de la verbalisation constitue une stratégie efficace pour améliorer la qualité des alignements complexes générés par des LLMs.

Dans la continuité de ce travail, plusieurs axes de recherche seront explorés. En particulier, nous envisageons d'étudier des stratégies de *prompting* plus sophistiquées, notamment des approches hybrides combinant différents types de verbalisation à différentes étapes de la chaîne de traitement, afin d'atténuer les limitations observées pour les verbalisations étiquette et naturelle. Des algorithmes de modularisation plus efficaces seront également explorés, notamment pour mieux contrôler le volume d'informations fournies au LLM et ainsi réduire l'effet de *Lost in the Middle*. Enfin, l'impact de l'utilisation de différents modèles de plongements et de raisonnement sur les performances du système fera l'objet d'expérimentations approfondies.

Remerciements

Cassia Trojahn est partiellement financée par le projet ANR DACE-LD (ANR-21-CE23-0019-02).

Références

[1] R. Amini et al. Towards complex ontology alignment using large language models. In *Knowledge Graphs and Semantic Web - 6th International Conference, KGSWC 2024*, volume 15459 of *Lecture Notes in Computer Science*, pages 17–31. Springer, 2024.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks*, 30(1-7) :107–117, 1998.

[3] J. Devlin et al. BERT : pre-training of deep bidirectional transformers for language understanding. In *Proc. of the NAACL-HLT 2019*, pages 4171–4186. ACL, 2019.

[4] T. B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, 2020.

[5] D. Faria et al. Agreementmakerlight. *Semantic Web*, 16(2) :SW–233304, 2022.

[6] D. Faria et al. Results in OAEI 2024 for matcha. In *Proc. of the 19th Int. Workshop on Ontology Matching (ISWC 2024)*, volume 3897 of *CEUR Workshop Proceedings*, pages 110–117. CEUR-WS.org, 2024.

[7] B. Fatemi, J. Halcrow, and B. Perozzi. Talk like a graph : Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024.

[8] Lucas Ferraz, Pedro Giesteira Cotovio, and Catia Pesquita. Can language models align biomedical ontologies ? : Evaluating retrieval-augmented prompt strategies in bio-ml. In *CEUR Workshop Proceedings*, volume 4001. CEUR-WS, 2025.

[9] J. Frey, L.-P. Meyer, N. Arndt, F. Brei, and K. Bulert. Benchmarking the abilities of large language models for RDF knowledge graph creation and comprehension : How well do llms speak turtle ? In *Proc. of the Workshop on Deep Learning for KGs (DL4KG 2023) co-located with the 21th Int. SW Conf. (ISWC 2023), Athens, 6-10, 2023*, volume 3559 of *CEUR Workshop Proceedings*, 2023.

[10] N. E. Fuchs et al. *Attempto Controlled English for Knowledge Representation*, pages 104–124. Springer Berlin Heidelberg, 2008.

[11] H. B. Giglou et al. Llms4om : Matching ontologies with large language models. In *The Semantic Web : ESWC 2024 Satellite Events*, volume 15344 of *LNCS*, pages 25–35. Springer, 2024.

[12] H. B. Giglou et al. Ontoaligner : A comprehensive modular and robust python toolkit for ontology alignment. In *The Semantic Web - 22nd European Semantic Web Conference, ESWC 2025*, 2025.

[13] Y. He, J. Chen, H. Dong, and I. Horrocks. Exploring large language models for ontology alignment. In *Proc. of the ISWC 2023 : From Novel Ideas to Industrial Practice co-located with 22nd Int. Semantic Web Conference (ISWC 2023), Athens, Greece, November 6-10, 2023*, volume 3632 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

[14] Y. He, J. Chen, H. Dong, I. Horrocks, C. Allocca, T. Kim, and B. Sapkota. Deeponto : A python package for ontology engineering with deep learning. *ArXiv*, abs/2307.03067, 2023.

[15] Y. He et al. Bertmap : A bert-based ontology alignment system. In *Thirty-Sixth AAAI Conf. on AI, AAAI 2022*, pages 5684–5691. AAAI Press, 2022.

[16] S. Hertling and H. Paulheim. Olala : Ontology matching with large language models. In *Proc. of the Knowledge Capture Conference 2023 (K-CAP '23)*. ACM, 2023.

[17] A. Lamurias, P. Ruas, and F. M. Couto. Ppr-ssm : personalized pagerank and semantic similarity measures for entity linking. *BMC Bioinformatics*, 20(1) :534, 2019.

[18] P. et al. Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, 2020.

[19] Y. Li et al. Deep entity matching with pre-trained language models. *Proc. VLDB Endow.*, 14(1) :50–60, 2020.

[20] N. F. Liu et al. Lost in the middle : How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12 :157–173, 2024.

[21] N. Muennighoff et al. MTEB : massive text embedding benchmark. In *Proc. of EACL 2023*, pages 2006–2029. ACL, 2023.

[22] OAEI. Complex track, 2024.

[23] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[24] R. Peeters and C. Bizer. Using chatgpt for entity matching. In *New Trends in Database and Information Systems - ADBIS 2023*, volume 1850 of *Communications in Computer and Information Science*, pages 221–230. Springer, 2023.

[25] N. Reimers and I. Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proc. of EMNLP-IJCNLP 2019*, pages 3980–3990. ACL, 2019.

[26] G. S. Sousa, R. Lima, and C. Trojahn. On evaluation metrics for complex matching based on reference alignments. In *The Semantic Web*, pages 77–93. Springer Nature Switzerland, 2025.

[27] G. S. Sousa, R. Lima, and C. Trojahn. Results of CMatch in OAEI 2025. In *Proc. of the 20th Int. Workshop on Ontology Matching co-located with the 24rd International Semantic Web Conference*, Nara, Japan, November 2025.

[28] C. Shimizu, Q. Hirt, and P. Hitzler. MODL : A modular ontology design library. In *Proc. of the 10th Workshop on Ontology Design and Patterns (WOP 2019)*, volume 2459 of *CEUR Workshop Proceedings*, pages 47–58. CEUR-WS.org, 2019.

[29] Y. Song, J. Chen, and R. A. Schmidt. Genom : Ontology matching with description generation and large language model. *CoRR*, abs/2508.10703, 2025.

[30] G. Sousa, R. Lima, and C. Trojahn. Towards generating complex alignments with large language models via prompt engineering. In *Proc. of the 19th Int. Workshop on Ontology Matching (OM-2024)*, CEUR Workshop Proceedings, 2024.

[31] M. Taboada, D. Martinez, M. Arideh, and Rosa Mosquera. Ontology matching with large language models and prioritized depth-first search, 2025.

[32] E. Thiéblin, G. Sousa, and C. Trojahn. Canard : An approach for generating expressive correspondences based on competency questions for alignment. *Semantic Web*, 15 :1–33, 2024.

[33] H. Touvron et al. Llama 2 : Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

[34] E. V. Vinu and P. S. Kumar. Ontology verbalization using semantic-refinement. *CoRR*, abs/1610.09964, 2016.

[35] W3C. Lexicon model for ontologies : Community report, 2016.

[36] Q. Wang, Z. Gao, and R. Xu. Exploring the in-context learning ability of large language model for biomedical concept linking. *CoRR*, abs/2307.01137, 2023.

[37] M. Zhao, S. Zhang, W. Li, and G. Chen. Matching biomedical ontologies based on formal concept analysis. *J. Biomed. Semant.*, 9(1) :11 :1–11 :27, 2018.

[38] L. Zhou and P. Hitzler. AROA results for OAEI 2020. In *Proc. of the 15th Int. Workshop on Ontology Matching (ISWC 2020)*, volume 2788 of *CEUR Workshop Proceedings*, pages 161–167. CEUR-WS.org, 2020.