# SUJET DE THÈSE : Système d'Évaluation de la ABox d'une Ontologie par Rapport à un Gold Standard (SEABORGS)

Lieu : Université de Caen-Normandie / Laboratoire GREYC
Direction : Yann Mathet (yann.mathet@unicaen.fr)
Co-encadrement : Céline Alec (celine.alec@unicaen.fr), Esteban Marquer (esteban.marquer@unicaen.fr) et Antoine Widlöcher (antoine.widlocher@unicaen.fr)
Dates : 01/10/26 – 30/09/29

## RÉSUMÉ DU PROJET

Version Française

Dans le domaine de l'intelligence artificielle, les connaissances peuvent être organisées au sein d'« ontologies », des structures qui distinguent les règles abstraites (le modèle) des données concrètes (les faits). Lorsqu'un logiciel ou un humain remplit ces structures avec des données concrètes (on parle de peuplement d'une ontologie), il est crucial de vérifier la qualité du résultat. Par exemple, on peut comparer les données produites à une version idéale sur des exemples validés par des experts. Cette évaluation permet entre autres de déterminer quelle est la meilleure méthode de peuplement parmi plusieurs approches concurrentes. Cependant, les méthodes d'évaluation actuelles s'avèrent insuffisantes : trop rigides, les mesures utilisées sanctionnent la moindre différence technique comme une erreur totale, sans percevoir les nuances de sens ou les réponses partiellement justes. Ainsi, il est possible qu'une approche à peu près correcte mais un peu approximative soit très mal évaluée, à cause de cette rigidité. L'objectif de cette thèse est de concevoir une méthode innovante pour mesurer avec finesse dans ce contexte la distance entre les données produites et la vérité attendue. Pour cela, une piste envisagée est de s'inspirer et tirer bénéfice des travaux existants sur la qualité des annotations et les mesures d'accord, qui reposent eux aussi sur la comparaison de structures complexes.

Version Anglaise

In the field of artificial intelligence, knowledge can be organized within "ontologies"—structures that distinguish between abstract rules (the model) and concrete data (the facts). When software or a human fills these structures with concrete data (known as ontology population), it is crucial to verify the quality of the result. For example, the produced data can be compared to a perfect version on examples validated by experts. Among other things, this evaluation makes it possible to determine the best population method among several competing approaches. However, current evaluation methods prove insufficient: the measures used are too rigid and penalize the slightest technical difference as a total error, failing to perceive semantic nuances or partially correct answers. Consequently, a substantially correct but slightly imprecise approach might be very poorly evaluated due to this rigidity. The objective of this thesis is to design an innovative method to measure the distance between the produced data and the expected truth with greater nuance. To achieve this, one approach being considered is to draw inspiration from existing research on annotation quality and agreement measures, which also rely on the comparison of complex structures.

**Mots clés :** Ontologies, Validité des données, Mesures d'évaluation, Gold Standard, ABox

## PRÉSENTATION DU PROJET :

**Contexte et Objectifs :**

**Contexte général**

Une **ontologie** [1] est une structure formelle utilisée pour modéliser et représenter de manière explicite les connaissances dans un domaine particulier. Elle définit les concepts, les relations entre ces concepts, et les propriétés associées. Ce sont des outils essentiels pour structurer et organiser des connaissances dans des domaines variés tels que la biologie, la médecine, et les données ouvertes liées (Linked Open Data).

Une ontologie est généralement divisée en deux composantes principales [2] :

1. **TBox** (Terminological Box) : Elle décrit les aspects terminologiques, c'est-à-dire la structure conceptuelle et les relations générales entre les concepts. Par exemple, dans une ontologie pour le domaine médical, la TBox pourrait définir que *Médicament* est une sous-classe de *Produit* ou qu'un *Patient* peut recevoir une *Prescript*ion.
2. **ABox** (Assertional Box) : Elle contient les instances qui exploitent les concepts et relations définis dans la TBox. Par exemple, elle pourrait inclure des informations concrètes comme « *Paracétamol* est un médicament » ou « *Alice* est une patiente » ou encore le fait qu'*Alice* reçoit la *prescription1* qui contient du *Paracétamol*.

Les éléments ontologiques (cf. Figure 1) sont représentés sous la forme de **triplets** $(sujet, prédicat, objet)$. Pour la ABox, on parle d'assertions de classes, par exemple, $(Paracétamol, type, Médicament)$ ou $(Alice, type, Patiente)$, et d'assertions de propriétés, comme $(Alice, reçoit, prescription1)$ ou $(prescription1, contient, Paracétamol)$.

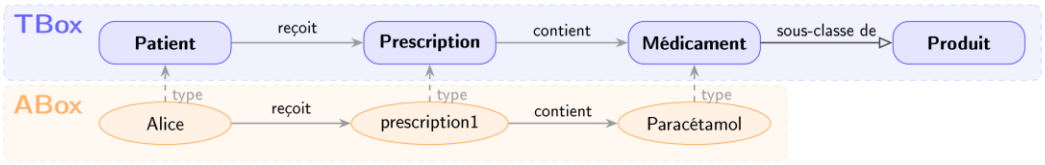


*Figure 1 : Exemple simplifié d'une ontologie*

La qualité d'une ontologie est cruciale pour garantir son efficacité dans des applications pratiques. L'évaluation des ontologies vise à mesurer des caractéristiques telles que leur cohérence, leur complétude, leur exactitude et leur pertinence par rapport à un domaine donné. Dans des applications critiques comme la gestion de données biomédicales ou le suivi d'actifs industriels, il est essentiel d'évaluer à quel point une ABox reflète fidèlement une vérité terrain : le **Gold Standard (GS)**. Il s'agit d'une référence idéale ou un modèle de vérité établi dans un domaine particulier, souvent validé par des experts.

**Objectifs**

L'objectif de ce travail de thèse est de proposer un système d'évaluation de la ABox d'une ontologie par rapport à une ABox Gold Standard. Dans le cadre de travaux antérieurs [3], nous avons proposé, appliqué et évalué des algorithmes de peuplement d'ontologie, i.e., permettant d'ajouter une ABox sur la base d'une TBox préexistante. Sur les 3 domaines étudiés, nous disposons d'une ontologie GS. Nous souhaitons donc comparer des ontologies pour lesquelles la TBox est la même mais la ABox diffère. Nous avons constaté que les méthodes d'évaluation actuellement utilisées dans l'état de l'art ne nous

permettaient pas d'établir une évaluation suffisamment fine et automatisée dans ce contexte. Il s'agit donc d'une proposition de projet de **recherche innovante**.

**État de l'art**

Les études [4] et [5] présentent des outils et des méthodes d'évaluation d'ontologies. On peut y constater que l'accent est porté sur l'évaluation de la TBox des ontologies. En termes de qualité, les aspects considérés sont structurels (taxonomie, lien entre les concepts), fonctionnels (à quel point l'ontologie est adaptée à diverses tâches) et d'utilisabilité (ontologie cohérente, concise, complète, correcte, etc.). Très peu d'articles s'intéressent réellement à l'évaluation de la ABox d'une ontologie et les méthodes pour comparer les faits dans les ontologies n'ont pas encore été investiguées [6].

Un cadre est proposé dans [7] pour comparer et évaluer quatre ontologies du domaine du bâtiment en termes de TBox et de ABox. Pour l'évaluation de la ABox, une étude empirique est menée. Celle-ci se concentre sur des cas pratiques impliquant des données d'instance. On cherche à mesurer la facilité avec laquelle des instances spécifiques peuvent être encodées, à tester la capacité des ontologies à répondre à des requêtes, et à simuler des mises à jour dynamiques. De ce fait, l'évaluation propre à la ABox ne concerne au final pas directement la ABox, mais plutôt la structuration, gestion et exploitation des données.

Dans [8], les auteurs cherchent à évaluer la qualité de bases de connaissances connues du Web des données (DBpedia, Freebase, YAGO, etc.). Ils définissent des métriques d'évaluation, portant sur la TBox et la ABox de ces ontologies. À titre d'exemples, il y a deux métriques évaluant la complétude en termes de ABox. L'une d'entre elles concerne les individus et ne s'applique qu'aux assertions de classes pour lesquelles on peut construire un Gold Standard (par exemple, les instances de la classe *Ville* sont comparées à une liste de villes créées à partir de plusieurs ressources). L'autre concerne les assertions de propriétés et ne s'applique que sur les propriétés pour lesquelles on est sûr que chaque individu du domaine doit avoir une valeur associée. Par exemple, c'est le cas de la *date de naissance* pour la classe *Personne*, mais pas le cas de la *date de décès*.

L'article [9] s'intéresse à des approches d'apprentissage d'ontologies et en évaluant les résultats par rapport à un GS. Des métriques s'apparentant à la précision, au rappel et à la F-mesure sont utilisées. L'évaluation concerne les classes, propriétés, relations de subsomption et individus. Les ontologies GS possèdent respectivement 5 et 0 individus, et aucune n'a d'assertions de propriétés. L'évaluation de la ABox est donc minime.

Dans [6], un cadre théorique est proposé pour comparer semi-automatiquement des ABox. L'auteur s'intéresse particulièrement à la justesse et à la complétude des faits. La justesse se caractérise par l'absence de contradiction entre ABox et TBox et par le respect des règles du domaine (par exemple, un numéro de téléphone doit contenir le bon nombre de chiffres). La complétude se définit par certaines règles (au moins un label pour chaque individu, chaque individu fait partie d'au moins une relation, etc.). À la suite de ce travail, une évaluation de plusieurs ontologies dans le domaine des unités de mesure a été proposée [10]. Les alignements entre les individus des ontologies sont basés sur des liens d'équivalence prédéfinis et sur des alignements obtenus via les labels des individus. Le développement [11] et perfectionnement [12] d'un outil, ABECTO (An ABox Evaluation and Comparison Tool for Ontologies) a été initié afin de disposer d'un outil d'évaluation pour les ABox ontologiques. Cet outil, se voulant le plus modulaire possible, nécessite une forte implication de l'utilisateur, pour paramétrer toutes les caractéristiques souhaitées.

En résumé, l'évaluation des ABox est un sujet relativement peu étudié. Les métriques utilisées actuellement sont à gros ou moyen grain. Des métriques plus précises pourraient être prises en compte, surtout lorsqu'on dispose d'un GS. De plus, l'identification des individus équivalents entre les deux ABox considérées est supposée simple (même label), et les individus équivalents au sein d'une même ABox ne sont pas considérés.

**Projet détaillé** :

Ce travail s'inscrit dans la continuité de travaux menés au GREYC dans deux branches jusqu'ici distinctes : l'ingénierie des connaissances d'une part, les mesures d'évaluation et d'accord d'autre part. C'est donc à la fois un travail incrémental et innovant.

L'objectif de cette thèse est de concevoir des méthodes et des outils robustes pour évaluer une ABox par rapport à un Gold Standard (GS). Ainsi, les objectifs sont de fournir :

- un cadre méthodologique complet pour l'évaluation des ABox par rapport à un GS ;
- un outil pratique implémentant ce cadre ;
- des contributions scientifiques sous forme de nouvelles métriques, algorithmes et expérimentations comparatives.

**Verrous scientifiques et techniques**

Le verrou scientifique majeur est l'absence de métriques suffisamment fines pour évaluer une ABox candidate par rapport à une ABox de référence (GS). En effet, les métriques standard utilisées en l'absence de GS sont à gros grain (cohérence générale de la ABox, présence d'assertions). En présence d'un GS, celles-ci sont à grain moyen : elles exploitent les notions de triplets vrais positifs (VP), faux positifs (FP) et faux négatifs (FN). Aucun système n'admet qu'un triplet pourrait être partiellement vrai. Ainsi, si obtient le triplet $X = (Pierre, frèreDe, Paul)$ au lieu de $Y = (Pierre, frèreDe, Jacques)$, on comptabilise 0 VP, 1 FP (X) et 1 FN (Y). Or, même si X est erroné, il contient tout de même une information juste (*Pierre* est le frère de quelqu'un). Nous aimerions identifier des moyens de tenir compte de ce genre de problème et proposer des métriques plus adaptées.

De plus, les langages ontologiques ne respectent pas l'hypothèse de nom unique. Cela signifie que plusieurs individus équivalents peuvent être présents dans une même ontologie sans aucune trace explicite de leur équivalence. Avant de pouvoir confronter deux ABox, il faudra il faudra donc déceler ces équivalences internes. Dans le cas où celles-ci ne sont pas directes (équivalences non explicitées et aucun label en commun), ce genre de problème reste un défi [10] qui n'est pas pris en compte dans les évaluations actuelles.

Ces **verrous scientifiques** constituent l'**originalité** de la thèse. Notons que nous n'étudierons que des cas d'utilisation sur des sujets pour lesquels un GS est disponible ou constructible par des experts du domaine.

Par ailleurs, afin de pouvoir comparer deux ABox, il faut être capable de trouver les couples d'individus équivalents dans l'une et dans l'autre (par exemple, trouver que « *JohnDoe* » dans la source *A* correspond à « *J_Doe* » dans la source *B)*. La plupart des travaux d'évaluation de ABox, comme [8], [9] et [10], ne font pas vraiment face à ce problème car

les individus à aligner ont soit le même identifiant, soit des labels identiques ou proches, ce qui n'est pas forcément le cas dans les données dont nous disposons. L'alignement d'ontologies, en particulier la correspondance d'instances, est un domaine de recherche actif avec une variété d'approches. Lorsque la TBox est la même dans les ontologies à aligner, cela simplifie le problème et ouvre la voie à des approches plus ciblées.

**Méthodes envisagées**

Le principe consistant à s'appuyer sur un « alignement » des éléments à comparer entre le candidat et le GS qui est à l'œuvre dans certaines mesures d'accord inter-annotateurs [13] pourrait être adapté au cas de la ABox candidate vs GS. Il s'agirait ici d'envisager quel élément de la Abox candidate (par exemple, « Alice ») correspond, éventuellement, à quel élément du GS (« patiente 1 »), et de s'appuyer sur cet alignement pour effectuer la mesure de validité. Si chaque élément du candidat correspond parfaitement à un élément du GS, le score est maximum, et inversement si aucun élément ne correspond, le score est nul. Entre ces deux extrêmes, les alignements contiendront généralement certains éléments qui ne se correspondent parfaitement, certains autres partiellement, et d'autres enfin qui restent sans aucune correspondance. L'alignement global retenu parmi l'ensemble des alignements possibles est celui qui minimise la mesure résultante. Aligner et mesurer se font donc conjointement.

D'autres méthodes d'alignement seront également à étudier :

- les méthodes basées sur la similarité, qui exploitent la similarité lexicale (labels, notions de distance entre chaînes de caractères) [14] et structurelle (relations communes) [15], ou bien calculent une similarité en utilisant des représentations vectorielles des instances, par exemple, au moyen de OWL2Vec* [16] ;
- les méthodes probabilistes, qui utilisent certains axiomes ontologiques (comme la fonctionnalité [17]) ;
- les méthodes à base d'apprentissage automatique [18], qui apprennent des modèles de correspondance basés sur les caractéristiques des instances et leurs relations dans la TBox commune ;
- les méthodes qui utilisent un processus itératif pour affiner les correspondances en exploitant les informations extraites des appariements précédents, comme le fait RiMOM-IM [19].

Enfin, la comparaison de deux ABox pourrait s'inspirer des mesures d'édition entre des graphes [20] représentant chacun une ABox. La distance entre elles correspondant alors au coût de la séquence d'opérations d'édition élémentaires (insertion d'un sommet, insertion d'une arête, ...) ayant le coût minimal parmi les différentes séquences de transformation possibles. Le paramétrage des coûts associés aux opérations d'édition devrait tenir compte des spécificités des nœuds et relations manipulés, en lien avec leur sémantique précise. Ainsi, par exemple, le coût de l'insertion d'une assertion de classe pourrait être plus élevé que celui d'une assertion de propriété. Notons tout de même qu'il est essentiel d'intégrer des critères qui vont au-delà des simples coûts structurels. En effet, une transformation qui, du point de vue structurel, semble être un changement plus important peut en réalité conserver un sens sémantique similaire. Soit par exemple un GS indiquant que Pierre est le frère de Paul. Si un candidat affirme que Pierre est le frère de Jacques, deux opérations seront nécessaires (suppression + insertion). Si un candidat ne se prononce pas, une seule opération sera nécessaire (insertion). Une pondération sera donc nécessaire pour que la différence de coût, dans de telles situations, soit justifiée.

Par ailleurs, ce travail de recherche pourra prolonger et étendre les travaux entrepris dans la thèse d'Antoine Boiteau, dans la mesure où l'on est confronté dans les deux cas à la comparaison de connaissances exprimées sous forme de graphes.

## Évaluation de la pertinence de la mesure proposée

Il est important de savoir et de comprendre dans quelle mesure la méthode d'évaluation proposée est pertinente, mais aussi comment elle se comporte par rapport aux autres méthodes d'évaluation. Une appréciation qualitative n'étant pas suffisante, nous prévoyons de nous appuyer sur l'outil Corpus Shuffling Tool [21] en l'adaptant aux structures ontologiques. Cet outil consiste en quelque sorte à inverser les rôles : tandis que c'est habituellement à une mesure d'évaluer la qualité de jeu de données, le Shuffling Tool crée des jeux de données spécifiques qui vont permettre d'évaluer le comportement d'une mesure. Il s'appuie sur un jeu de données initial considéré comme GS, qu'il va décliner en une série de jeux de données de plus en plus dégradés, via des processus de dégradation contrôlés de différents types. On lance alors la mesure sur ces données de plus en plus dégradées et on observe comment la mesure réagit. Idéalement, la mesure devrait être strictement décroissante. Cet outil a permis de comparer finement le comportement de diverses mesures dans le domaine de la catégorisation, et d'élaborer et de comparer les mesures Gamma dédiées à l'unitizing. Nous pourrons nous appuyer sur les travaux actuels d'Antoine Boiteau qui travaille lui aussi sur l'adaptation de cet outil à des structures relationnelles relativement proches de celles qui nous concernent.

**Références**


[1] N. Guarino, D. Oberle, et S. Staab, « What Is an Ontology? », in *Handbook on Ontologies*, S. Staab et R. Studer, Éd., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, p. 1-17.

[2] T. R. Gruber, « A translation approach to portable ontology specifications », *Knowledge Acquisition*, vol. 5, nº 2, p. 199-220, juin 1993.

[3] A. Sahbi, C. Alec, et P. Beust, « Semantic vs. LLM-based approach: A case study of KOnPoTe vs. Claude for ontology population from French advertisements », *Data & Knowledge Engineering*, vol. 156, p. 102392, 2025.

[4] M. Amirhosseini et J. Salim, « A Synthesis Survey of Ontology Evaluation Tools, Applications and Methods to Propose a Novel Branch in Evaluating the Structure of Ontologies: Graph-Independent Approach », *IJC*, vol. 33, nº 1, p. 46-68, mai 2019.

[5] J. Raad et C. Cruz, « A Survey on Ontology Evaluation Methods », in *International Conference on Knowledge Engineering and Ontology Development*, 2015.

[6] J. M. Keil, « Ontology ABox Comparison », in *The Semantic Web: ESWC 2018 Satellite Events*, A. Gangemi, A. L. Gentile, A. G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J. Z. Pan, et M. Alam, Éd., Cham: Springer International Publishing, 2018, p. 240-250.

[7] Z. Qiang *et al.*, « A systematic comparison and evaluation of building ontologies for deploying data-driven analytics in smart buildings », *Energy and Buildings*, vol. 292, p. 113054, août 2023.

[8] M. Färber, F. Bartscherer, C. Menne, et A. Rettinger, « Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO », *Semantic Web*, vol. 9, p. 77-129, 2017.

[9] M. Schaeffer, M. Sesboüé, L. Charbonnier, N. Delestre, J.-P. Kotowicz, et C. Zanni-Merk, « On the pertinence of LLMs for ontology learning », présenté à 3rd NLP4KGC International Workshop on Natural Language Processing for Knowledge Graph Creation at 20th International Conference on Semantic Systems (Semantics), Amsterdam, sept. 2024.

[10] B. Brodaric, J. M. Keil, et S. Schindler, « Comparison and evaluation of ontologies for units of measurement », *Semant. web*, vol. 10, nº 1, p. 33-51, 2019.

[11] J. M. Keil, « ABECTO: An ABox Evaluation and Comparison Tool for Ontologies », in *The Semantic Web: ESWC 2020 Satellite Events*, Cham: Springer International Publishing, 2020, p. 140-145.

[12] J. M. Keil, « Continuous Knowledge Graph Quality Assessment through Comparison using ABECTO. », in *Joint Proceedings of Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems co-located with 20th International Conference on Semantic Systems (SEMANTiCS 2024), Amsterdam, The Netherlands, September 17-19, 2024.*, 2024.

[13] Y. Mathet, A. Widlöcher, et J.-P. Métivier, « The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment », *Computational Linguistics*, vol. 41, nº 3, p. 437-479, sept. 2015.

[14] A.-C. Ngonga Ngomo *et al.*, « LIMES: A Framework for Link Discovery on the Semantic Web », *KI - Künstliche Intelligenz*, vol. 35, nº 3, p. 413-423, nov. 2021.

[15] A. Barbosa, I. I. Bittencourt, S. W. Siqueira, D. Dermeval, et N. J. T. Cruz, « A Context-Independent Ontological Linked Data Alignment Approach to Instance Matching », *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 18, nº 1, p. 1-29, 2022.
[16] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, et I. Horrocks, « OWL2Vec*: embedding of OWL ontologies », *Machine Learning*, vol. 110, nº 7, p. 1813-1845, juill. 2021.
[17] F. M. Suchanek, S. Abiteboul, et P. Senellart, « PARIS: Probabilistic Alignment of Relations, Instances, and Schema. », *Proc. VLDB Endow.*, vol. 5, nº 3, p. 157-168, 2011.
[18] M. Zhang et Y. Chen, « Link Prediction Based on Graph Neural Networks », in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, et R. Garnett, Éd., Curran Associates, Inc., 2018.
[19] C. Shao, L.-M. Hu, J.-Z. Li, Z.-C. Wang, T. Chung, et J.-B. Xia, « RiMOM-IM: A Novel Iterative Framework for Instance Matching », *Journal of Computer Science and Technology*, vol. 31, nº 1, p. 185-197, janv. 2016.
[20] X. Chen, H. Huo, J. Huan, et J. S. Vitter, « An efficient algorithm for graph edit distance computation », *Knowledge-Based Systems*, vol. 163, p. 762-775, 2019.
[21] Y. Mathet *et al.*, « Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics », in *Proceedings of COLING 2012: Posters*, M. Kay et C. Boitet, Éd., Mumbai, India: The COLING 2012 Organizing Committee, déc. 2012, p. 809-818. [En ligne]. Disponible sur: https://aclanthology.org/C12-2079/

Pour candidater, merci d'envoyer (au format pdf) un CV, une lettre de motivation et vos relevés de notes (de licence 3, 1ère année de master et les notes de 2ème année de master disponibles ; ou équivalent pour les écoles d'ingénieurs) à :

- Céline Alec <celine.alec@unicaen.fr> ;
- Esteban Marquer <esteban.marquer@unicaen.fr> ;
- Yann Mathet <yann.mathet@unicaen.fr> ;
- Antoine Widlöcher <antoine.widlocher@unicaen.fr>.