CIGI QUALITA MOSIM 2023
Enrichissement d'ontologies par la reconnaissance sémantique des données pour l'Internet Industriel des Objets

SIMON BAUER[1], CHRISTOPHE MERLO[2], ZINA BOUSSAADA[3]

[1] Univ. Bordeaux, ESTIA INSTITUTE OF TECHNOLOGY
F-64210 Bidart, France
s.bauer@estia.fr

[2] Univ. Bordeaux, ESTIA INSTITUTE OF TECHNOLOGY
F-64210 Bidart, France
c.merlo@estia.fr

[3] Univ. Bordeaux, ESTIA INSTITUTE OF TECHNOLOGY
F-64210 Bidart, France
z.boussaada@estia.fr

Résumé **- Les évolutions intrinsèques liées à l'industrie du futur mènent à l'intégration de nombreux systèmes d'information métiers au sein de l'industrie. La gestion du cycle de vie du produit, et plus récemment, la mise en œuvre du concept de jumeau numérique, amènent à se poser la question du partage des données métiers, à travers des modèles, des représentations et des significations qui ne sont pas nécessairement identifiés, compris et maitrisés par les différents acteurs métiers. Les ontologies, issues du web sémantique, sont l'un des outils utilisés pour faciliter cette intégration.**

L'omniprésence de l'Internet des objets et la multiplication des silos de données apporte également une hétérogénéité à de multiples niveaux qui rend complexe l'intégration de données issues de mesures à des modèles numériques issus de la conception d'un produit et présents dans les systèmes d'information, et qui empêche de tirer pleinement profit des capacités de ces outils à tisser des liens sémantiques entre les concepts de différents modèles et entre les données et les concepts.

Ce travail proposera une approche semi supervisée permettant de créer une ontologie d'interface, permettant l'enrichissement de la ABox en faisant le lien entre des données et des ontologies métier.

Abstract **- The intrinsic evolutions related to the industry of the future lead to the integration of many business information systems within the industry. The product life cycle management, and more recently, the implementation of the concept of digital twin, raises the question of the sharing of business data, through models, representations and meanings that are not necessarily identified, understood, and mastered by the various business actors. Ontologies, from the semantic web, are one of the tools used to facilitate this integration.**

The omnipresence of the Internet of Things and the multiplication of data silos also brings heterogeneity at multiple levels which makes it complex to integrate data from measurements with digital models from the design of a product and present in information systems, and which prevents taking full advantage of the capabilities of these tools to weave semantic links between the concepts of different models and between data and concepts.

This work will propose a semi-supervised approach to create an interface ontology, allowing the enrichment of the ABox by making the link between data and business ontologies.

Mots clés **– interopérabilité, ontologie, Internet des objets, approche MBSE, jumeau numérique**
Keywords **– interoperability, ontology, Internet of Things, MBSE approach, digital twin**

1 INTRODUCTION

La quatrième révolution industrielle (Xu et al., 2018) apporte des changements importants, et l'un de ces changements est l'intégration de l'Internet des Objets (IoT) dans presque tous les processus existants. Dans les entreprises industrielles, tout peut désormais être surveillé, produisant une quantité massive de données qui peuvent ensuite être utilisées pour la supervision, l'analyse et la prédiction. Ces capteurs se trouvent à la fois partout du fait de leur nombre, et nulle part du fait de leur taille. Chaque outil, chaque pièce est un élément au sein d'un tout, et les données qu'un tel élément fourni fait que le tout est plus grand que la somme de ses parties. Suivre et identifier la relation entre une décision et les changements qui en résultent dans les indicateurs clés de performance est l'une des ambitions possibles de cette évolution.

Si pendant longtemps, des documents textes, sujets à interprétations, étaient utilisés pour la réalisation de tels produits, à cause de la complexité croissante des systèmes et poussé par le besoin d'innovation, l'ingénierie des systèmes basée sur des modèles, aussi nommée Model-Based Systems Engineering (MBSE) tend à être utilisée de plus en plus

(Shevchenko, 2020). Le MBSE est alors l'«application formalisée de la modélisation pour prendre en charge les exigences du système, la conception, l'analyse, la vérification et les activités de validation commençant dans la phase de conception et se poursuivant tout au long du développement et des phases ultérieures du cycle de vie.» (INCOSE, 2007) Les modèles sont alors placés au centre du processus de design, remplaçant les documents jusque-là utilisés dans l'ingénierie.
L'approche MBSE peut être étendue via le concept de jumeau numérique (Digital Twin) (Phanden et al., 2021). Ce dernier correspond à l'avatar numérique d'un objet physique, qui le suit non seulement au cours de son cycle de conception et de fabrication, mais également de l'ensemble de son cycle de vie. En se basant sur l'expertise métier liée à l'objet, mais aussi sur les informations recueillies par des capteurs, il est ainsi possible de vérifier les modèles définis pendant le développement du produit par rapport à des mesures réelles, superviser les performances, et également réactualiser ces modèles ainsi que les actions à réaliser en fonction de ces données pour prolonger son maintien en conditions opérationnelles par exemple, ou pour planifier au mieux les opérations de maintenance, etc.
Le développement du concept de jumeau numérique, un avatar virtuel d'une entité physique, se sert des données collectées par l'entité physique afin d'établir un lien entre monde numérique et monde physique (*What Is a Digital Twin?*, s. d.). Toutes les pièces évoquées précédemment sont à la fois identifiées physiquement et numériquement.
Un exemple d'application pourrait être la traçabilité des composants tout au long du cycle de vie d'un produit spécifique (*3 Great Examples of Digital Twin Technology In Action*, s. d.). En connaissant la façon dont les composants sont liés les uns aux autres et ont été modifiés, il est possible de savoir exactement ce qui est arrivé au produit physique, et d'anticiper sur certains évènements ou actions que le produit physique subira dans le futur.
Un autre cas d'utilisation est l'inférence d'équations complexes à l'aide de relations plus simples (Jia et al., 2022). Un exemple trivial est le suivant : connaissant le lien entre poids et émission de CO_2 pendant le transport, et le lien entre matériau utilisé et poids, le lien entre matériau et émission de CO_2 pendant le transport pourrait être déduit.

Dans cet article, nous abordons la question de la gestion de l'information relative au produit, le caractérisant physiquement comme numériquement, et en particulier de la multiplicité des sources d'information qui génère des problèmes de duplication, de redondance, de versions mais aussi tout simplement de mise à disposition, que vont rencontrer les différents acteurs métiers. Nous présentons dans cet article les premiers résultats de nos travaux, qui demandent encore des expérimentations en grandeur réelle.
La prochaine section décrit plus précisément la problématique que nous étudions, et l'approche que nous avons privilégiée pour cela. La section 3 fait le point sur les techniques utilisées dans ce type de d'approche. Notre proposition méthodologique outillée est présentée en section 4, et se termine sur nos premiers résultats. La discussion en section 5 détaille les problèmes rencontrés lors de l'implémentation, les limites constatées à ce stade ainsi que les travaux nombreux restants à mener.

2 PROBLEMATIQUE

2.1 Unicité de la donnée

L'objectif central de notre travail consiste à permettre aux différents acteurs métiers de maitriser les données qu'ils manipulent ainsi qu'à celles qu'ils peuvent manipuler, générées par d'autres métiers, et dans des contextes différents. Cela pose la question de l'unicité de la donnée, et par là de la consolidation des différentes sources de données, issues de modèles ou de mesures IoT. Une information, issue d'un ensemble de données, doit être transférée de manière transparente sur le réseau et aucune donnée ne doit jamais être perdue.
Ce principe d'unicité de la donnée entraine la question de l'interopérabilité. L'interopérabilité est définie comme la capacité de communiquer, d'exécuter des programmes ou de transférer des données entre diverses unités fonctionnelles d'une manière qui oblige l'utilisateur à avoir peu ou pas de connaissance des caractéristiques uniques de ces unités. (ISO/IEC 2382:2015, 2015)
En pratique, l'unité de la donnée est souvent cassée. Les contraintes structurelles, telles que le fait d'avoir plusieurs sous-organisations dans l'entreprise et différents outils logiciels métiers, les problèmes de confidentialité, tels que les restrictions légales sur l'accès aux données personnelles et les problèmes de sécurité, tels que la nécessité de ne pas transférer des informations sensibles ou des secrets commerciaux à des sous-traitants, ont conduit à la création de nombreux silos de données. Les silos de données se caractérisent par le stockage de quantités massives de données souvent hétérogènes, isolées, parfois redondantes, voire contradictoires (Patel, 2019).
Un silo de données peut aussi parfois être créé afin de réduire les coûts. Il peut être plus facile et moins coûteux de répliquer des données dans un stockage local que d'avoir à gérer un moyen de rendre ces données globalement accessibles uniquement aux bonnes personnes, tenues à jour à l'aide d'un transfert de données massif et au bon format. Le nombre de silos de données augmente donc aussi vite que les sources de données et l'utilisation des données.

2.2 Approche ontologique

Une façon de répondre à ce problème des silos de données est l'utilisation d'ontologies. Une ontologie est une base de connaissances (knowledge base, KB) qui se compose de deux éléments. Les connaissances générales sur un domaine, appelées base de règles (terminological component, TBox) qui sont un ensemble fini d'axiomes. L'autre élément est constitué des connaissances sur des individus particuliers, appelées base de faits (assertion component, ABox), qui sont des assertions (Ben Mahria et al., 2021). Le mot ontologie désigne souvent la base de règles, mais peut parfois être utilisé pour des bases de connaissances entières. Les ontologies sont un outil venant du web sémantique, une vision de l'Internet par la sémantique plutôt que par les données brutes.
L'usage du web sémantique au service de l'industrie permet d'apporter les connaissances et techniques liées. Cette mise en pratique n'est pas récente. Dans (Léger et al., 2005) existent des cas d'usages du web sémantique dans un contexte industriel dès 2003. Par exemple, (Maier et al., 2003) décrit l'usage d'ontologies dans l'industrie automobile, en utilisant les ontologies comme une forme d'abstraction d'un système physique qui fournit des données, pour permettre le changement ou l'ajout de sources de données.
De nombreux domaines ont voulu bénéficier du web sémantique. Les ontologies sont utilisées dans le monde industriel pour stocker des connaissances et créer des liens entre des silos de données, et l'industrie s'est alors retrouvée avec une multitude d'ontologies, construits sur des domaines comme les standards industriels sur les données produits (Fraga et al., 2018), le management de la supply chain (Wallace, 2021) ou l'I40 (Sampath Kumar et al., 2019).

Avec cet accroissement de l'usage des ontologies, on a vu apparaitre plusieurs définitions (Euzenat & Shvaiko, 2013) en rapport avec l'interopérabilité sémantique :

- Un alignement d'ontologies (ontology alignment) est un jeu de correspondances entre plusieurs ontologies. Ces correspondances peuvent être simples, c'est à dire entre des entités, ou complexes, c'est à dire entre des groupes d'entités et des sous-structures.
- Une correspondance d'ontologies (ontology matching) peut contenir des prédicats sur la similarité, appelés matching, ou un axiome logique, appelé mapping. Les termes mapping et matching sont souvent utilisés indistinctement dans la littérature et de façon interchangeable.
- Une fusion d'ontologies (ontology merging) est la combinaison de plusieurs ontologies en une seule, qui contient la connaissance de toutes celles qui ont fusionné. Les fusions les plus avancées contiennent des axiomes supplémentaires qui définissent comment les ontologies relatent les unes aux autres. Ces axiomes sont souvent le résultat d'alignements.
- Une traduction d'ontologies (ontology translation) est le processus de changer la sémantique sous-jacente à une connaissance afin de passer de la sémantique d'une ontologie source à celle d'une ontologie cible. Une traduction est bonne lorsqu'aucune connaissance n'est perdue en faisant cette traduction.

De plus, construire et maintenir une ontologie est complexe. Il y a souvent besoin de plusieurs experts du domaine pour la maintenir à jour, et cela prend du temps, car il est nécessaire de s'assurer que les connaissances stockées dans l'ontologie restent à jour.

De plus, même si la TBox de la base de connaissances peut être maintenue à jour par des experts car elle a tendance à être statique, la ABox est beaucoup plus dynamique car elle est basée sur des données réelles et a besoin de pouvoir s'adapter rapidement lorsque les sources de données changent.

Notre travail a pour but de contribuer à faciliter ce maintien de l'ontologie, en réduisant la quantité de travail nécessaire pour la mise à jour de l'ABox dans le cadre des ontologies industrielles.

Un système d'information industriel peut être représenté par une ontologie avec sa TBox et ses silos de données. Dans la figure 1, l'ontologie est constituée des concepts A, B, C et les silos de données sont les tableaux colorés où les concepts associés à chaque colonne de données sont les entêtes du tableau. Après traitement, les concepts identiques entre l'ontologie et les silos de données ont été liés, tandis que les données sont jointes.

Les cas qui nécessitent d'être considérés par notre méthode ont été identifiés comme présenté dans le tableau I :

- la source de donnée peut être connue ou non par rapport à l'état de l'ontologie de destination avant traitement ;
- les données de ces sources peuvent également être connues (déjà vu) ou non ;
- enfin l'ontologie cible peut déjà ou non contenir la donnée considérée.

Le tableau illustre la multitude de cas qui peuvent être rencontrés quand nous souhaitons enrichir l'ontologie de destination.

Les spécificités des données industrielles doivent être prises en compte dans la méthode proposée : les données provenant de sources de l'Internet Industriel des Objets (IIOT) sont souvent impropres et peuvent contenir, soit des entrées manquantes, soit des entrées erronées, soit des entrées redondantes.

Comme les ontologies industrielles sont des entités évolutives, le fait que l'ensemble de données puisse ne pas être entièrement à jour et que le système doive en apprendre davantage pour pouvoir remplir correctement sa tâche doit également être pris en compte. Un nouveau type de source de données peut être ajouté sans que l'information ne soit transférée vers le système. Il est nécessaire de savoir que le modèle est peut-être obsolète et de demander à l'utilisateur s'il souhaite toujours l'exploiter. Enfin, il est important d'éviter d'avoir à stocker le jeu de données complet à chaque fois que le système doit être entraîné.

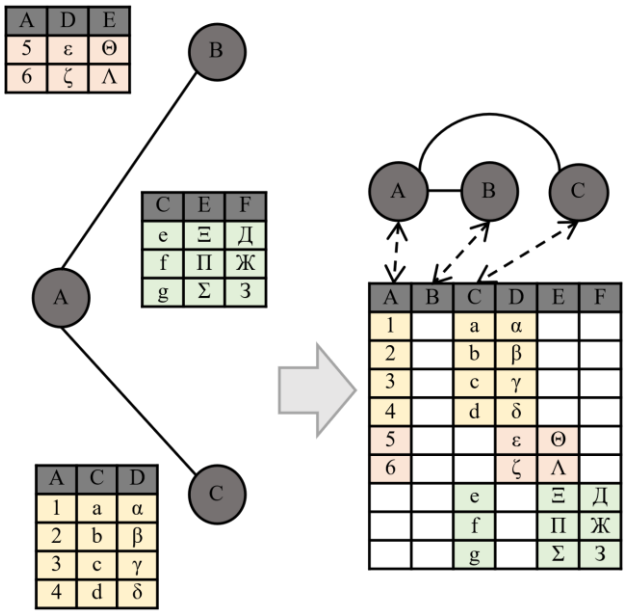


Figure 1. Création du lien entre ontologie et données

Tableau 1. Classification des données dans l'ontologie de destination

Cas	Source de donnée	Statut du type de donnée	Statut du type de donnée dans l'ontologie cible
1	Connue	Déjà vu	Présente
2	Connue	Déjà vu	Absente
3	Nouvelle	Déjà vu	Présente
4	Nouvelle	Déjà vu	Absente
5	Nouvelle	Jamais vu	Absente
6	Nouvelle	Jamais vu	Absente mais faussement identifiée

3 ÉTAT DE L'ART

Une approche basée sur l'interopérabilité sémantique n'est pas envisageable. Comme nous ne disposons que d'une ontologie écrite par un expert et que les autres ontologies sont extraites d'outils industriels, les alignements d'ontologies ou les correspondances d'ontologies ne seront pas une solution convenable pour notre problème car nécessitant un temps de traitement humain extrêmement important, entièrement manuel. La littérature est riche en outils permettant d'automatiser la correspondance de schémas (schema matching), tels que COMA (Rahm & Do, 2002). Ces techniques utilisent souvent des noms d'éléments, des types de données, des propriétés structurelles et des caractéristiques des instances de données. Mais dans notre cas, comme les données provenant de sources IoT sont utilisées avec peu ou pas de prétraitement et pas forcément de schéma connu pour les faire correspondre, un tel outil ne pourrait pas être appliqué.

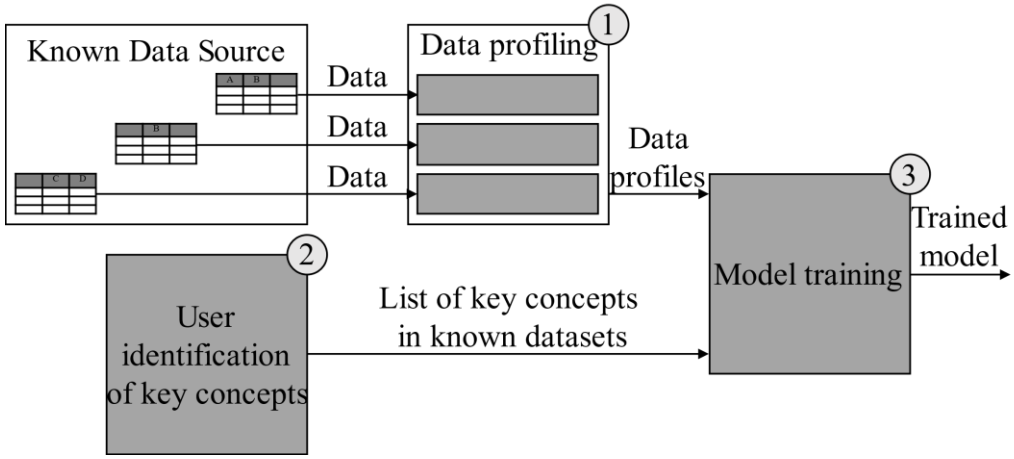


Figure 2. Schéma de la phase d'entrainement

Des outils tels que SiMA (Koutras et al., 2022) sont assez pertinents et efficaces pour faire correspondre le stockage de données dans le silo de données. Son utilisation des réseaux de neurones en graphe (Graph Neural Networks) fournit une approche intéressante, et ils répondent au besoin de "profils de données" pour réduire la quantité d'informations stockées dans chaque colonne. Cependant, cet outil est principalement axé sur l'utilisation de silos de données "propres", qui n'est pas adapté à notre utilisation sur des données brutes IoT.

Enfin, le domaine de la détection des types sémantiques de données (Semantic Data Type Detection) fournit des résultats intéressants pour résoudre les cas 1, 2, 3 et 4, où chaque donnée visible est connue à l'avance. Néanmoins, il manque les cas 5 et 6 où des données inconnues pourraient être rencontrées par le système. Utiliser des outils comme Sherlock (Hulsebos et al., 2019) ou Sato (Zhang et al., 2020) s'avère alors intéressant pour notre système : la capacité de ces outils à identifier le type de données sans avoir à stocker l'ensemble des données d'entraînement dans un silo est de plus l'un des points clés que nous recherchons.

4 METHODE

Dans la section suivante, nous présentons la méthode proposée, qui s'appuie sur les approches présentées en section 3, afin d'adapter ces outils aux données industrielles et de les rendre compatibles avec une utilisation conjointe d'une ontologie. Cette approche permet de surmonter les limitations des méthodes précédentes, principalement appliquées sur des ensembles de données idéaux et prétraités, en les rendant aptes à traiter des scénarios réels et complexes d'Internet des objets.

La méthode proposée, intègre des améliorations méthodologiques pour tenir compte des défis spécifiques liés à l'IoT, tels que la variabilité des données, les contraintes de ressources et les problèmes d'évolutivité. En associant ces outils à une ontologie intermédiaire, nous offrons de nouvelles perspectives et possibilités pour leur exploitation dans des contextes industriels, notamment en facilitant l'analyse et l'interprétation des données industrielles.

Cette méthode a été conçue pour éviter autant que possible la nécessité de transférer une grande quantité de données sur le réseau de l'entreprise. Il y a plusieurs raisons à ce choix : premièrement, cela peut entraîner à la fois des problèmes de sécurité, le risque d'une fuite de données augmentant lorsque le nombre d'ordinateurs sur lesquels elles sont stockées augmente, et des problèmes de confidentialité, les données peuvent devoir passer sur des systèmes qui pourraient appartenir à différents sous-traitants.

Une autre raison est une raison de bande passante : l'IoT a tendance à produire beaucoup de données et les transférer sur le réseau serait un gaspillage de bande passante. Les appareils IoT ont souvent une bande passante réduite en raison de leur localisation à l'extrémité du réseau.

Notre outil est basé sur un processus en 2 phases :

- Une phase d'entrainement semi-supervisé, utilisant les retours des utilisateurs pour sélectionner les colonnes importantes qui seront ensuite conservées pour l'entrainement. Le modèle est alors entrainé et peut être déployé.
- Une phase de détection, utilisant le modèle précédemment entrainé pour détecter les types de données et les convertir en ontologie.

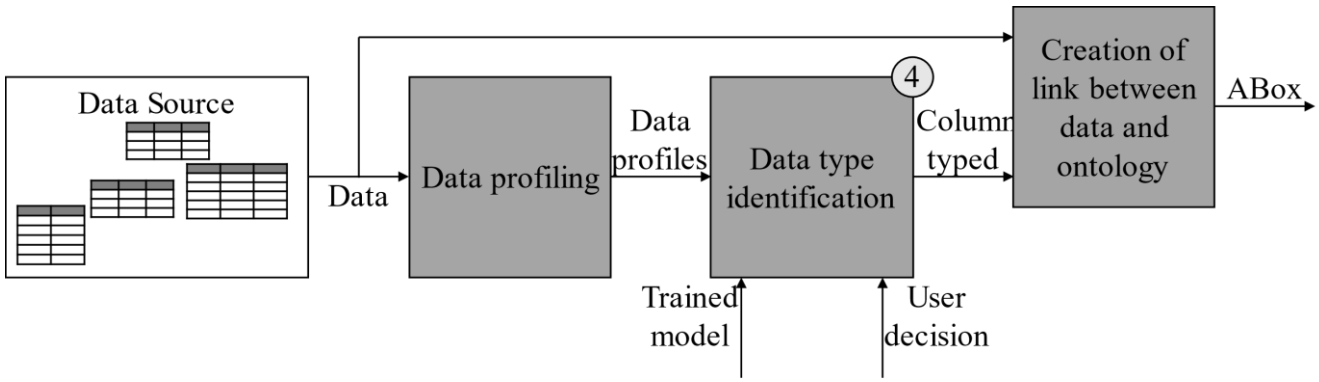


Figure 3. Schéma de la phase de détection

4.1 Phase d'entrainement

La première étape (**1**) est la **réduction** de chaque colonne de données provenant de la source IoT à ses caractéristiques principales. Dans notre implémentation, Sherlock a été utilisé pour construire un profil de 1588 données décrivant les caractéristiques de chaque colonne, telles que "les propriétés statistiques, la distribution des caractères, la vectorisation de mots (word embedding) et la vectorisation de paragraphes des valeurs des colonnes".

Cette étape de profilage est utilisée pour résoudre les problèmes soulevés précédemment sur la confidentialité des données et la bande passante. Ce processus de profilage est suffisamment léger pour être effectué sur un appareil embarqué.

Une autre étape (**2**) effectuée simultanément est **l'identification** de "concepts clés" dans le jeu de données utilisé pour l'entrainement. L'utilisateur doit identifier manuellement les colonnes qui seront utilisées pour l'entrainement du modèle et le concept qui peut être trouvé dans cette colonne.

Les raisons de ce choix de conception sont que, dans un ensemble de données industrielles, de nombreuses colonnes sont souvent remplies de données inutiles ou vides, les données de colonnes de type similaire peuvent avoir des noms différents et cette étape permet à un utilisateur ayant des connaissances métier de sélectionner ce qui est important et qui doit être conservé dans l'ontologie.

L'étape suivante (**3**) consiste à entrainer le modèle en utilisant les profils des données de (**1**) comme entrée et les concepts de (**2**) comme classes. Comme l'extraction de caractéristiques de Sherlock est utilisée pour (**1**), notre premier modèle est basé sur l'architecture de 'deep learning' utilisée par Sherlock.

Après l'entrainement, le modèle est prêt à être déployé pour la détection.

4.2 Phase de détection

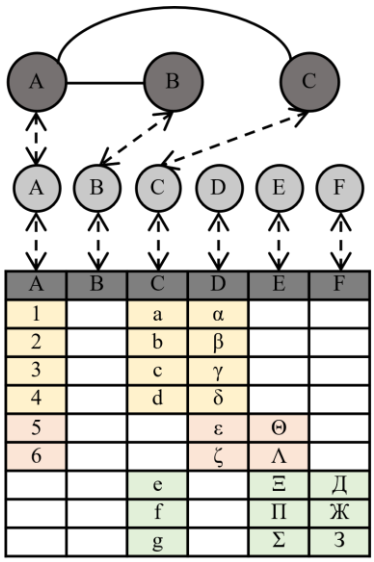
Pour identifier les données provenant d'une nouvelle source, la même technique de profilage qu'en (**1**) est utilisée pour maintenir le profil de données cohérent entre la phase d'entrainement et la phase de détection.

L'étape d'identification (**4**) utilise le modèle entraîné en (**3**) et se base sur un seuil de détection. Plutôt que d'utiliser uniquement la sortie avec la possibilité la plus élevée, un score de prédiction minimal pour conserver l'identification pour l'étape suivante a été mis en œuvre. Le niveau de confiance peut être ajusté en fonction du cas d'utilisation.

Cette approche est justifiée par des besoins industriels : il est plus risqué de créer un lien qui ne devrait pas être créé que de manquer un lien là où il devrait y en avoir un. Le premier cas peut conduire à un mauvais alignement des données, ce qui peut entraîner l'invalidité de l'ensemble de l'ontologie. D'autre part, la création d'un lien par la suite n'entraînera pas d'effets sur l'ontologie.

Suivant cette idée, une autre fonctionnalité offerte par la méthode proposée est la possibilité pour un utilisateur de prendre une décision lorsque le réseau de neurones offre un résultat suffisamment bon pour passer un certain seuil, mais pas assez pour atteindre notre seuil de confiance élevé.

Une ontologie d'interface constituée de chaque nom de colonne en tant que TBox et des données associées à chaque colonne en tant qu'ABox, comme illustré à la figure 4, est ensuite générée. Lors de l'étape d'entrainement, une ontologie composée de chaque concept sélectionné par l'utilisateur a déjà été créée. Les liens entre les colonnes identifiées par le réseau et les concepts sélectionnés peuvent enfin être créés.



A	B	C	D	E	F
1		a	α		
2		b	β		
3		c	γ		
4		d	δ		
5			ε	Θ	
6			ζ	Λ	
		e		Ξ	Д
		f		П	Ж
		g		Σ	З

Figure 4. Création du lien entre l'ontologie d'interface et l'ontologie industrielle

4.3 Implémentation

Nous avons implémenté cette méthode avec Sherlock en tant que profileur de données et en utilisant 2 ensembles de données IoT pour évaluer notre méthode sur un ensemble de données limité. Pour augmenter le nombre d'échantillons d'entrainement, nous avons fait le choix dans nos tests de scinder artificiellement notre jeu de données en plusieurs jeux de données plus petits. Au lieu d'avoir un profil représentant des milliers de lignes, nous en avons une douzaine, chacun représentant quelques centaines de lignes.

La précision atteint 98%. L'étape la plus longue du processus est la conversion des données en profil, environ 30 secondes. L'entrainement se fait généralement en moins de 10 epochs en raison de la petite taille de l'échantillon.

En conclusion de cette section dédiée aux résultats, il est nécessaire de souligner que l'évaluation de notre travail s'est avérée être difficile. Il est complexe de réaliser des comparaisons directes avec d'autres approches, car elles ne sont pas nécessairement adaptées aux mêmes conditions et objectifs que notre méthode.

C'est la raison pour laquelle nous nous sommes appuyés sur les mesures quantitatives présentées ici, plutôt que de comparer notre approche avec d'autres outils existants.

Dans cette étude, nous avons également rencontré des problèmes liés à la confidentialité des données. Les données IoT réelles, collectées et valorisées pour mettre en œuvre et évaluer notre méthode, proviennent souvent de sources industrielles et commerciales sensibles, qui sont soumises à des restrictions de confidentialité. Par conséquent, il n'est pas possible de diffuser ces données en raison de leur caractère confidentiel.

5 PROBLEMES, LIMITES ET AMELIORATIONS POSSIBLES

Ces premiers tests ayant donné des résultats satisfaisants avec le jeu de données limité dont nous disposons, nous prévoyons, dans le cadre des prochaines étapes de notre travail, de confronter les résultats de l'utilisation de la méthode proposée à un cas industriel complexe : un système PLM gérant des données produit et de fabrication, afin de comprendre ses avantages et ses limites.

Certains problèmes et améliorations possibles qui pourraient profiter à notre système ont déjà été identifiés.

L'un des inconvénients majeurs de l'approche basée sur le type sémantique est le risque de voir des colonnes de données différentes avec des types identiques se confondre les unes avec les autres. Par exemple, avec à la fois une "Date de début" et une "Date de fin" dans l'ensemble des données, le système devra s'appuyer uniquement sur la sémantique pour distinguer l'une de l'autre. Et si les dates sont "Date 1" et "Date 2", il sera difficile, voire impossible, de résoudre cette tâche sans apport externe.
Une amélioration consiste à prendre en compte les retours de l'utilisateur lors de l'exécution de l'identification du type de données pour une identification ultérieure. Cette étape est assez simple, car elle ne nécessite que l'envoi des profils de données actuelles dans l'étape d'entrainement. Cependant, cela nécessite à nouveau l'entraînement de l'ensemble du modèle sur l'ensemble de données complet. Pouvoir mettre en œuvre un apprentissage continu (Hadsell et al., 2020) pour prendre en compte de nouvelles sources de données au fur et à mesure qu'elles sont identifiées par l'utilisateur pourrait être utile. Néanmoins, comme la formation du modèle est pour l'instant une tâche rapide qui peut être effectuée en moins d'une minute sur un ordinateur portable, elle n'est pas l'une de nos priorités.

Une autre amélioration est l'ajout d'un sous-système dédié à la reconnaissance des sources de données, plutôt que des types de données. Dans les systèmes d'information industriels, il est courant d'avoir un même type de source utilisé plusieurs fois, ce qui peut être le cas par exemple pour un capteur utilisé à plusieurs endroits. La possibilité de créer un modèle source pourrait améliorer la précision en reconnaissant l'intégralité de la source plutôt que des colonnes individuelles. Cependant, il existe un risque qu'une même source utilise une sortie de données modifiée, par exemple si les données sont filtrées au préalable ou si le logiciel source est mis à jour.

Une dernière amélioration pourrait être implémentée au niveau de l'ontologie de l'interface. Actuellement, l'ensemble de données est stocké en tant qu'ABox et les concepts en tant que TBox, mais il n'y a aucune tentative de trouver un lien qui pourrait être contenu dans l'ABox. Par conséquent, les concepts contenus dans la TBox n'ont pas de liens ou de relations qui sont présents dans une ontologie normale.
Essayer d'exploiter les connaissances sous-jacentes dans l'ABox pourrait aider à trouver suffisamment de relations pour créer des liens reliant les concepts de la TBox. Un outil comme SiMA, exploitant également des profils de données pour construire un Graph Neural Networks peut être une approche intéressante pour commencer.
Comparer la TBox construite à partir des données et la TBox construite par un ingénieur avec des connaissances métier et corriger l'une des TBox grâce à l'autre ou les utiliser ensemble pour combiner les connaissances des deux sources pourrait fournir des résultats intéressants.

Néanmoins, il est crucial de considérer les aspects environnementaux comme une limite. Malgré le potentiel du jumeau numérique pour améliorer l'efficacité énergétique, optimiser l'utilisation des ressources et favoriser des pratiques durables sur le cycle de vie des objets, il est essentiel d'examiner cet aspect avec prudence. L'IA et l'IoT peuvent également engendrer des conséquences environnementales. L'IA entraîne une consommation énergétique croissante en raison des calculs intensifs, tandis que l'IoT génère une demande accrue de dispositifs connectés, ayant un impact sur l'extraction de ressources et la gestion des déchets électroniques. L'analyse doit tenir compte des incertitudes et des limites des données disponibles pour offrir une perspective équilibrée et nuancée de l'impact environnemental potentiel de l'IA, des jumeaux numériques et de l'IoT.

6 Conclusion

Un cadre a ainsi été proposé sur la base d'outils sémantiques pour créer des liens entre plusieurs silos de données et une ontologie. Grâce à nos propositions, les experts métiers pourront manipuler des ensembles de données sémantiquement consolidées, par le biais d'une « réunification » de différentes sources de données issues de logiciels de modélisation comme de mesures (IoT), en application du concept de Jumeau Numérique, tout au long du cycle de vie d'un système physique. Ce cadre utilise le profilage des données existantes et la capacité du deep learning pour apprendre et reconnaitre les caractéristiques significatives de ces profils. L'un des points clés du profilage est sa capacité à être effectué directement sur le terminal, réduisant ainsi le besoin de transférer des données.
Ce cadre a été mis en œuvre et les résultats obtenus lors de nos tests sont acceptables. Certains axes d'amélioration, mentionnés dans la partie 5, ont été détectés.

7 Remerciements

Nous remercions notre partenaire industriel avec lequel nous travaillons dans un contexte de thèse CIFRE.

8 References

Euzenat, J., & Shvaiko, P. (2013). *Ontology Matching.* Heidelberg: Springer Berlin. https://doi.org/10.1007/978-3-642-38721-0

Fraga, A., Vegetti, M., & Leone, H. (2018). Semantic Interoperability among Industrial Product Data Standards using an Ontology Network. *Proceedings of the 20th International Conference on Enterprise Information Systems*, *2*, pp. 328-335. https://doi.org/10.5220/0006783303280335

Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. (2020, November 03). Embracing Change: Continual Learning in Deep Neural Networks. *Trends in Cognitive Sciences*, *24*(12), 1028-1040. https://doi.org/10.1016/j.tics.2020.09.004

Hulsebos, M., Hu, K., Bakker, M., Zgraggen, E., Satyanarayan, A., Kraska, T., Demiralp, Ç., & Hidalgo, C. (2019, August 4-8). Sherlock: A Deep Learning Approach to Semantic Data Type Detection. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery \&\#38; Data Mining* (pp. 1500–1508). Anchorage: ACM. https://doi.org/10.1145/3292500.3330993

INCOSE. (2007). *Systems Engineering Vision 2020.* International Council on Systems Engineering, Seattle.

ISO/IEC 2382:2015. (2015, 5). *Information technology — Vocabulary*. https://www.iso.org/obp/ui/#iso:std:iso-iec:2382

Koutras, C., Hai, R., Psarakis, K., Fragkoulis, M., & Katsifodimos, A. (2022). SiMa: Effective and Efficient Data Silo Federation Using Graph Neural Networks. arXiv. https://doi.org/10.48550/ARXIV.2206.12733

Léger, A., Nixon, L. J., Shvaiko, P., & Charlet, J. (2005). Semantic Web applications: Fields and Business cases. The Industry challenges the research. Dans M. Bramer, & V. Terziyan (Éd.), *Industrial Applications of Semantic Web* (pp. 27-46). Boston, MA: Springer US. https://doi.org/10.1007/0-387-29248-9_2

Maier, A., Schnurr, H.-P., & Sure, Y. (2003). Ontology-Based Information Integration in the Automotive Industry. Dans D. Fensel, K. Sycara, & J. Mylopoulos (Éd.), *The Semantic Web - ISWC 2003* (pp. 897-912). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-39718-2_57

Patel, J. (2019, May). Overcoming Data Silos through Big Data Integration. *International Journal of Computer Science and Technology*, *3*(1). https://doi.org/10.5121/IJDMS.2019.1301

Phanden, R. K., Sharma, P., & Dubey, A. (2021). A review on simulation in digital twin for aerospace, manufacturing and robotics. *Materials Today, 38*(1), 174-178. https://doi.org/https://doi.org/10.1016/j.matpr.2020.06.446

Rahm, E., & Do, H.-H. (2002). COMA - A System for Flexible Combination of Schema Matching Approaches. *VLDB '02: Proceedings of the 28th International Conference on Very Large Databases,*, (pp. 610-621). Hong Kong. https://doi.org/10.1016/B978-155860869-6/50060-3

Sampath Kumar, V. R., Khamis, A., Fiorini, S., Carbonera, J. L., Olivares Alarcos, A., Habib, M., Goncalves, P., Li, H., & Olszewska, J. I. (2019, 11 22). Ontologies for Industry 4.0. *The Knowledge Engineering Review*. https://doi.org/10.1017/S0269888919000109

Wallace, E. (2021). Towards a Reference Ontology for Supply Chain Management. *10th International Conference on Interoperability for Enterprise Systems and Applications "Interoperability in the Era of Artificial Intelligence"*. Tarbes. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=929874

Xu, M., David, J. M., & Kim, S. H. (2018). The Fourth Industrial Revolution: Opportunities and Challenges. *International Journal of Financial Research, 9*(2). https://doi.org/10.5430/ijfr.v9n2p90

Zhang, D., Hulsebos, M., Suhara, Y., Demiralp, Ç., Li, J., & Tan, W.-C. (2020, August). Sato: contextual semantic type detection in tables. *Proceedings of the VLDB Endowment, 13*(11), 1835-1848. https://doi.org/10.14778/3407790.3407793