Contribution à la modélisation de réseaux de spécifications sémantiques et tissage des liens entre les données “ produit ” tout au long de son cycle de vie

Simon Bauer

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

PRÉPARÉE À L'ESTIA

ÉCOLE DOCTORALE : SCIENCES PHYSIQUES ET DE L'INGÉNIEUR

SPÉCIALITÉ : AUTOMATIQUE, PRODUCTIQUE, SIGNAL ET IMAGE, INGÉNIERIE COGNITIQUE

Par Simon BAUER

Contribution à la modélisation de réseaux de spécifications sémantiques et tissage des liens entre les données « produit » tout au long de son cycle de vie

Sous la direction de Christophe MERLO

Soutenue le 09/12/2024

Membres du jury :

M. Vincent CHEUTET	Pr.	INSA Lyon	Président
Mme. Roberta COSTA-AFFONSO	Pr.	ISAE-SUPMECA	Examinatrice
Mme. Rebeca ARISTA RANGEL	Dr.	Universidad de Sevilla	Examinatrice
M. Benoit EYNARD	Pr.	Université de Technologie de Compiègne	Rapporteur
M. Raymond HOUÉ NGOUNA	MCF	Université de Technologie de Tarbes	Rapporteur
M. Christophe MERLO	Pr.	ESTIA Recherche	Directeur de thèse

Membres invités :

Mme. Zina BOUSSAADA	MCF	ESTIA Recherche	Co-encadrante
M. Simon RINCÉ		Airbus	Invité

Résumés en français et en anglais

Titre de la thèse

Contribution à la modélisation de réseaux de spécifications sémantiques et tissage des liens entre les données « produit » tout au long de son cycle de vie

Résumé

Cette thèse s'inscrit dans le contexte de l'Industrie 4.0 et aborde les défis liés à l'enrichissement des jumeaux numériques dans des environnements industriels complexes. Elle se concentre particulièrement sur l'intégration de données hétérogènes provenant de sources externes, en l'absence de modèles préétablis, pour améliorer l'interopérabilité et la performance dans un contexte multi-modèles et multi-métiers.

L'étude débute par une analyse approfondie du contexte industriel, explorant l'évolution de l'Industrie 4.0 vers l'Industrie 5.0 et les concepts fondamentaux tels que l'Internet Industriel des Objets (IIoT) et l'Ingénierie des Systèmes Basée sur les Modèles (MBSE). Elle met en lumière les défis actuels dans la mise en œuvre des jumeaux numériques, notamment la fragmentation des modèles MBSE et la difficulté d'intégrer de manière cohérente les données IIoT.

La problématique centrale de cette recherche est identifiée comme suit : Comment enrichir un jumeau numérique par l'intégration d'ensembles de données provenant de sources externes, en l'absence de modèles préétablis, pour améliorer l'interopérabilité et la performance dans un environnement multi-modèle et multi-métiers ?

Pour répondre à cette question, la thèse propose une approche novatrice basée sur l'utilisation combinée des technologies du Web sémantique et de l'intelligence artificielle. Cette approche vise à créer un cadre d'interopérabilité robuste permettant l'intégration harmonieuse des données hétérogènes au sein des jumeaux numériques.

La méthodologie développée comprend trois axes majeurs :

Ontologification des données : Une approche hybride combinant techniques d'IA et expertise humaine pour convertir les données brutes en ontologies, avec une attention particulière portée aux spécificités des données IIoT.

Alignement et intégration ontologique : Une méthodologie pour la création d'une ontologie de référence par les experts métier, couplée à un processus semi-automatique d'alignement et d'intégration des ontologies générées par l'IA avec cette référence.

Exploration et exploitation des connaissances : L'exploitation d'interfaces intuitives et d'outils de visualisation adaptés permettant aux acteurs métier d'explorer et d'exploiter efficacement les ontologies intégrées dans le jumeau numérique.

La validation de cette approche est réalisée à travers un cas d'étude industriel dans le secteur aéronautique, démontrant son applicabilité et son efficacité dans un environnement réel. Les résultats montrent une amélioration significative de l'interopérabilité et de l'exploitation des données hétérogènes au sein des jumeaux numériques.

Les contributions principales de cette thèse incluent une méthodologie complète pour l'enrichissement des jumeaux numériques, un framework d'interopérabilité basé sur les technologies du Web sémantique et l'IA, des techniques avancées pour le traitement et l'intégration des données hétérogènes, et un prototype de système d'exploration des connaissances intégrées.

Cette recherche ouvre de nouvelles perspectives pour l'utilisation des jumeaux numériques dans l'industrie, en proposant des solutions concrètes aux défis d'interopérabilité et d'intégration de données hétérogènes. Elle contribue ainsi à l'avancement de l'Industrie 4.0 et prépare le terrain pour les futures innovations dans le domaine de l'ingénierie des systèmes et de l'Internet Industriel des Objets.

En conclusion, cette thèse apporte une contribution significative à la résolution des problèmes d'interopérabilité et d'intégration de données hétérogènes dans le contexte des jumeaux numériques, offrant ainsi aux industries un moyen plus efficace d'exploiter les avantages de l'Industrie 4.0 et de se préparer aux défis de l'Industrie 5.0.

Mots clés : architecture des systèmes d'information, intégration de données, interopérabilité sémantique, machine learning, jumeau numérique, PLM

Title :

Contribution to the modeling of semantic specification networks and the discovery of semantic links between "product" data throughout its life cycle

Abstract :

This thesis is situated within the context of Industry 4.0 and addresses the challenges associated with enriching digital twins in complex industrial environments. It focuses particularly on integrating heterogeneous data from external sources, in the absence of pre-established models, to enhance interoperability and performance in a multi-model and multi-disciplinary context.

The study begins with an in-depth analysis of the industrial context, exploring the evolution from Industry 4.0 to Industry 5.0 and fundamental concepts such as the Industrial Internet of Things (IIoT) and Model-Based Systems Engineering (MBSE). It highlights current challenges in implementing digital twins, notably the fragmentation of MBSE models and the difficulty of coherently integrating IIoT data.

The central research question is identified as follows: How can a digital twin be enriched through the integration of datasets from external sources, in the absence of pre-established models, to improve interoperability and performance in a multi-model and multi-disciplinary environment?

To address this question, the thesis proposes an innovative approach based on the combined use of Semantic Web technologies and artificial intelligence. This approach aims to create a robust interoperability framework allowing for the seamless integration of heterogeneous data within digital twins.

The developed methodology comprises three major pillars:

Data Ontologization: A hybrid approach combining AI techniques and human expertise to convert raw data into ontologies, with particular attention paid to the specificities of IIoT data.

Ontological Alignment and Integration: A methodology for the creation of a reference ontology by domain experts, coupled with a semi-automatic process for aligning and integrating AI-generated ontologies with this reference.

Knowledge Exploration and Exploitation: The utilization of intuitive interfaces and adapted visualization tools allowing domain actors to effectively explore and exploit the integrated ontologies within the digital twin.

The validation of this approach is carried out through an industrial case study in the aeronautical sector, demonstrating its applicability and effectiveness in a real-world environment. The results show a significant improvement in the interoperability and exploitation of heterogeneous data within digital twins.

The main contributions of this thesis include a comprehensive methodology for enriching digital twins, an interoperability framework based on Semantic Web technologies and AI, advanced techniques for processing and integrating heterogeneous data, and a prototype system for exploring integrated knowledge.

This research opens new perspectives for the use of digital twins in industry by proposing concrete solutions to the challenges of interoperability and integration of heterogeneous data. It thus contributes to the advancement of Industry 4.0 and paves the way for future innovations in the field of systems engineering and the Industrial Internet of Things.

In conclusion, this thesis makes a significant contribution to solving the problems of interoperability and integration of heterogeneous data in the context of digital twins, thereby offering industries a more effective means of exploiting the advantages of Industry 4.0 and preparing for the challenges of Industry 5.0.

Keywords : information systems architecture, data integration, semantic interoperability, machine learning, digital twin, PLM

Unité de recherche

ESTIA Recherche

RNSR N° 201420655V

N°97, allée Théodore Monod

ESTIA 2,Technopole Izarbel

64210 Bidart

Dédicace

À mon père,
qui aurait tant aimé voir l'aboutissement de cette thèse.

Remerciements

Cette thèse est l'aboutissement d'un travail qui n'aurait pu voir le jour sans le soutien et la contribution de nombreuses personnes que je tiens à remercier chaleureusement.

Je souhaite tout d'abord exprimer ma profonde gratitude envers mon directeur de thèse, le Professeur Christophe Merlo, pour la confiance qu'il m'a accordée en acceptant d'encadrer ce travail doctoral. Ses conseils avisés et son expertise dans le domaine du génie industriel ont été déterminants dans la réussite de ce projet.

Je remercie vivement ma co-encadrante, la MDC Zina Boussaada, pour son accompagnement constant, sa rigueur scientifique et ses précieuses contributions dans le domaine de l'intelligence artificielle. Ses encouragements et sa disponibilité ont grandement facilité l'avancement de mes travaux.

Ma reconnaissance va également à Monsieur Simon Rincé, mon superviseur industriel chez Airbus, pour son expertise technique, sa vision industrielle et son soutien sans faille tout au long de cette aventure. Son expérience et ses conseils ont été essentiels pour ancrer mes recherches dans une réalité industrielle concrète.

Je souhaite également exprimer ma sincère gratitude à Yvan Baudin, de l'ID Lab chez Airbus, qui m'a généreusement accueilli dès le début de ma thèse, à une période pourtant difficile liée à la COVID. Il m'a également guidé dans la découverte des bases du PLM.

Je tiens à remercier Airbus pour avoir rendu possible ce projet de recherche et m'avoir offert un environnement de travail stimulant. Ma gratitude va particulièrement à Matthijs Plokker pour m'avoir accueilli au sein de son équipe et avoir créé les conditions propices à la réussite de ce projet. Je suis particulièrement reconnaissant envers l'équipe DGYX pour leur accueil chaleureux et leur collaboration précieuse. Les échanges quotidiens et les discussions techniques avec mes collègues ont grandement enrichi ma réflexion et contribué à la qualité de ce travail.

Je tiens également à remercier l'équipe d'ESTIA Recherche pour leur accueil lors de mes passages au laboratoire. Bien que mes travaux se soient principalement déroulés en entreprise dans le cadre de cette thèse CIFRE, les échanges avec les chercheurs et doctorants du laboratoire ont toujours été enrichissants et constructifs.

Enfin, je souhaite exprimer ma gratitude envers ma famille et mes proches pour leur soutien indéfectible et leurs encouragements constants tout au long de ce parcours. Je remercie particulièrement mes amis qui ont su être présents dans les moments de doute comme dans les réussites, apportant leur soutien moral et leurs encouragements précieux. Merci à toutes les personnes qui, de près ou de loin, ont contribué à la réussite de cette aventure par leurs conseils, leur écoute ou simplement leur présence bienveillante.

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1 Contexte et problématique industrielle

« La quatrième révolution industrielle se caractérise par une fusion des technologies qui gomme les frontières entre les sphères physique, numérique et biologique. »

Klaus Schwab - Fondateur et Président exécutif, Forum économique mondial de Genève

1.1 Introduction

Dans le contexte de la quatrième révolution industrielle, plusieurs concepts clés ont émergé pour aider les entreprises à tirer pleinement parti des données collectées par les capteurs et des systèmes numériques. Parmi ces concepts, l'Ingénierie des Systèmes Basée sur des Modèles (MBSE, Model-Based Systems Engineering) permet aux ingénieurs de mieux comprendre les changements liés aux choix de conception et d'analyser ces derniers avant leur développement. Le Jumeau Numérique instancie ces modèles pour créer un avatar numérique d'un objet physique. L'Internet des Objets (IoT, Internet of Things), quant à lui, fournit des moyens de collecter des données à partir d'un grand nombre de capteurs et de dispositifs, permettant une surveillance et une analyse en temps réel pour une meilleure prise de décision. Enfin, l'interconnectivité entre les différents systèmes et données est également cruciale pour permettre aux entreprises d'optimiser les performances, de réduire les coûts et d'augmenter la productivité.

1.2 De l'Industrie 4.0 à l'Industrie 5.0

1.2.1 Définitions

La première révolution industrielle, celle de la vapeur, a débuté au XVIIIe siècle en Grande-Bretagne. Elle a été marquée par l'invention de machines à vapeur et l'utilisation de l'énergie mécanique pour remplacer la force humaine et animale dans la production. Cette révolution a également vu l'émergence de la production en série dans l'industrie textile, ainsi que la création de nouvelles méthodes de transport, tels que les chemins de fer.

La deuxième révolution industrielle, celle de l'électricité, a commencé à la fin du XIXe siècle et s'est poursuivie jusqu'au début du XXe siècle. Elle a été caractérisée par l'utilisation de l'électricité pour alimenter les machines et la production de masse. Cette révolution a également vu l'émergence de nouvelles industries, telles que l'automobile et l'aviation, ainsi que des progrès importants dans les communications, comme le téléphone et la radio.

La troisième révolution industrielle, celle de l'ordinateur, a débuté dans les années 1970 et s'est poursuivie jusqu'au début des années 2000. Elle a été marquée par l'émergence de l'informatique et des technologies de l'information. Cette révolution a vu l'utilisation de l'informatique pour automatiser la production et la création de nouveaux secteurs économiques, tels que les technologies de l'information et de la communication.

La quatrième révolution industrielle, telle que décrite par (Schwab, 2017), est celle des systèmes cyber-physiques. Elle est caractérisée par l'utilisation croissante de technologies numériques et de l'intelligence artificielle pour améliorer les processus de production et les opérations

industrielles. Cette révolution a vu l'essor de la puissance de calcul, de l'intelligence artificielle, de l'internet des objets, de l'impression 3D, des objets autonomes et de la robotique avancée.

Cette révolution a également vu l'émergence de nouveaux modèles d'affaires basés sur les données, où les entreprises peuvent utiliser les informations collectées par les capteurs et sur Internet pour prendre des décisions plus éclairées, améliorer la qualité de leurs produits et services, et fournir des expériences plus personnalisées aux clients. Comme le montre la Figure 1, les technologies de l'Industrie 4.0 offrent un large éventail de bénéfices, allant de l'augmentation de la productivité à la réduction des coûts et des délais de production.

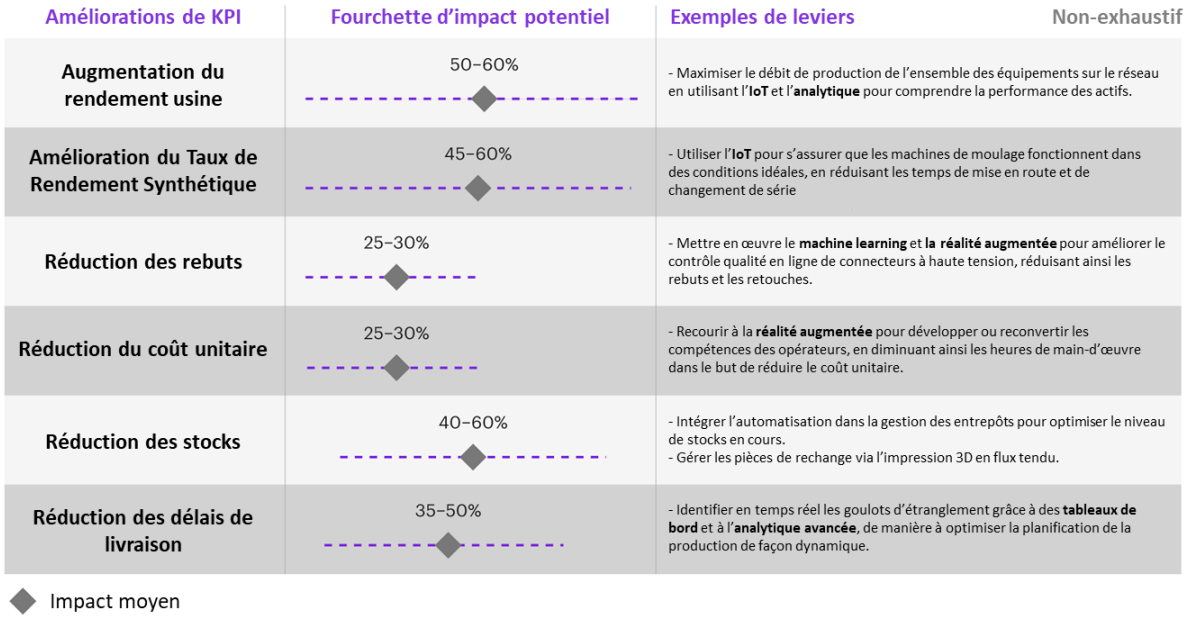


Figure 1 Bénéfices potentiels des technologies de l'Industrie 4.0 (Mehl & Anderson, 2021)

L'impact de la quatrième révolution industrielle peut être positif, en permettant une plus grande efficacité, des gains de productivité, une meilleure qualité de vie pour les individus et des solutions innovantes à des problèmes mondiaux tels que le changement climatique. Cependant, elle soulève également des défis importants en matière de sécurité, de confidentialité et d'éthique, en particulier en ce qui concerne l'utilisation des données personnelles et la prise de décisions autonomes par les machines. Il est donc crucial de développer des normes et des réglementations appropriées pour garantir que ces technologies sont utilisées de manière responsable et éthique.

L'objectif principal de la quatrième révolution industrielle est de créer des usines intelligentes et connectées, où les machines sont capables de communiquer entre elles et avec les travailleurs humains, de surveiller et d'optimiser les processus de production, et d'adapter leur fonctionnement en temps réel pour maximiser l'efficacité et la productivité.

Les données collectées peuvent inclure des informations sur les conditions environnementales, les performances des machines, la qualité des produits, les préférences des clients, les niveaux de stocks et bien plus encore. Les industriels disposent maintenant de nuées de capteurs, fournissant chacun des quantités massives de données. Ces nuées de capteurs permettent de collecter des données en temps réel sur pratiquement tous les aspects des opérations industrielles, de la production aux services après-vente en passant par la maintenance et la logistique.

Ces capteurs se trouvent à la fois partout du fait de leur nombre, et nulle part du fait de leur taille. Cette prolifération de capteurs a permis de collecter des données à une échelle sans précédent, fournissant des informations précieuses sur l'ensemble de l'écosystème industriel. Chaque outil, chaque pièce est un élément au sein d'un tout, et les données qu'un tel élément fournit font que le tout est plus grand que la somme de ses parties.

Les entreprises se servent des données collectées par l'entité physique afin d'établir un lien entre monde numérique et monde physique. Toutes les pièces évoquées précédemment sont à la fois identifiées et existent physiquement et numériquement.

Les réseaux de télécommunications, comme décrits par (Rao & Prasad, 2018, p. 145-159) jouent un rôle clé dans la connectivité de cette révolution. Ces nouvelles technologies nécessitent des réseaux haut débit et à faible latence pour fonctionner efficacement, ce qui pousse les opérateurs de télécommunications à investir massivement dans des infrastructures de réseau avancées.

La 5G, en particulier, offre des vitesses de transmission de données plus rapides, une plus grande capacité et une latence ultra-faible, ce qui est essentiel pour prendre en charge les applications critiques. Les réseaux de télécommunications sont également utilisés pour connecter les travailleurs, les machines et les capteurs à travers des réseaux locaux sans fil (WLAN, Wireless Local Area Network) et des réseaux de capteurs sans fil (WSN, Wireless Sensor Network), permettant ainsi de créer des systèmes de production et de logistique plus connectés et donc plus efficaces. En fin de compte, les réseaux de télécommunications jouent un rôle crucial dans la réalisation de l'Industrie 4.0 en fournissant la connectivité nécessaire pour permettre aux entreprises d'exploiter pleinement les avantages de la quatrième révolution industrielle. La Figure 2 illustre ainsi les trois catalyseurs principaux de l'Industrie 4.0 : l'explosion des données, l'évolution de la puissance de calcul et l'adoption massive d'Internet.

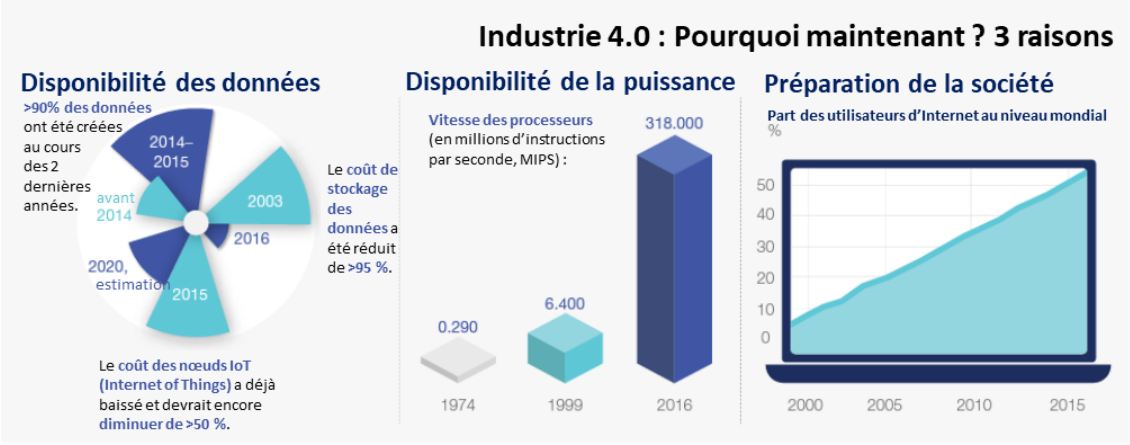


*Figure 2 Les catalyseurs de l'Industrie 4.0 (*Industry 4.0: Reinvigorating ASEAN manufacturing for the future*, s. d.)*

Ces quantités massives de données peuvent ensuite être analysées à l'aide de technologies d'analyse de données avancées pour découvrir des tendances, des modèles et des anomalies, ce qui peut aider les entreprises à prendre des décisions plus éclairées et à améliorer leur efficacité opérationnelle. Ces données sont utilisées pour améliorer la qualité des produits, optimiser les processus de production et de logistique, et fournir des expériences plus personnalisées aux clients.

Dans le prolongement de cette évolution, il convient de mentionner l'émergence de l'Industrie 5.0, qui représente une extension et un raffinement des concepts de l'Industrie 4.0. Cette nouvelle itération met l'accent sur une approche centrée sur l'humain, en reconnaissant l'importance cruciale de l'interaction entre l'homme et la machine. L'Industrie 5.0 se distingue par sa focalisation sur la flexibilité et la reconfigurabilité des systèmes de production, visant à optimiser la collaboration entre les travailleurs humains et les technologies avancées. Plutôt que d'introduire de nouvelles technologies révolutionnaires, elle cherche à exploiter de manière plus efficace et éthique les innovations existantes, en mettant l'accent sur la durabilité, la personnalisation de masse et l'adaptabilité des processus de production. (Nahavandi, 2019) souligne l'importance de l'intégration harmonieuse des compétences humaines avec les capacités technologiques pour créer des environnements de travail plus intelligents, plus réactifs et plus centrés sur l'humain.

1.2.2 L'Internet Industriel des Objets

Le concept d'Internet des Objets (Internet of Things, IoT) a été introduit par (Ashton, 2009) en 1999, bien que l'idée d'une informatique ubiquitaire ait été évoquée dès 1991 par (Weiser, 1991), qui imaginait des technologies "entrelacées dans le tissu de la vie quotidienne jusqu'à ce qu'elles en soient indifférenciables". Cependant, ce n'est qu'à partir de 2010 que le concept a véritablement gagné en popularité.

L'IoT désigne l'interconnexion d'objets du quotidien à Internet. Il utilise des capteurs et des technologies de communication pour connecter des objets physiques à Internet, créant ainsi un réseau d'objets capables de communiquer entre eux et avec des systèmes de traitement de données. Cette interconnexion permet de transmettre des informations et des commandes, rendant les processus plus intelligents.

(Atzori et al., 2010) considère l'IoT comme une évolution majeure d'Internet et de la connectivité, car il permet de relier le monde physique au monde numérique, ouvrant la voie à de nouveaux services et applications auparavant impossibles. Les données collectées par ces objets connectés peuvent être utilisées pour :

- Surveiller leur état en temps réel
- Collecter des données précises
- Prendre des décisions rapides et éclairées
- Optimiser leur utilisation
- Automatiser des tâches
- Fournir des informations utiles aux utilisateurs

Il existe de nombreuses applications de l'IoT dans la vie quotidienne, comme le soulignent (Risteska Stojkoska & Trivodaliev, 2017). En voici quelques exemples :

- Les montres connectées qui surveillent la santé et l'activité physique, offrant des informations en temps réel sur le rythme cardiaque, le nombre de pas effectués, la qualité du sommeil et même la saturation en oxygène du sang.

- Les thermostats connectés qui ajustent automatiquement la température de votre maison en fonction de votre présence et de vos préférences, permettant ainsi une gestion énergétique optimisée et un confort personnalisé.
- Les voitures connectées qui peuvent surveiller les performances, fournir des mises à jour en temps réel sur l'état du véhicule, et même ajuster la conduite pour économiser de l'énergie. Ces véhicules peuvent également communiquer avec l'infrastructure routière pour améliorer la sécurité et l'efficacité du trafic.
- Les appareils électroménagers connectés qui peuvent être contrôlés à distance, comme les lave-linges ou les réfrigérateurs intelligents. Un exemple est le réfrigérateur qui émet une alerte de porte ouverte sur le téléphone de son propriétaire, permettant ainsi d'éviter le gaspillage énergétique et la détérioration des aliments. Un autre exemple est l'aspirateur autonome qui adapte sa trajectoire en fonction de son analyse de son environnement, optimise la puissance d'aspiration en fonction de la surface et revient à sa base lorsque les batteries ont besoin d'être rechargées, offrant ainsi une solution de nettoyage automatisée et efficace.
- Les systèmes de sécurité domestique intelligents qui peuvent surveiller l'activité et alerter les propriétaires en cas d'intrusion, intégrant souvent des caméras, des détecteurs de mouvement et des serrures connectées pour une protection complète et personnalisable.

L'Internet Industriel des Objets (Industrial Internet of Things, IIoT) représente l'intégration des technologies IoT dans le contexte industriel. Son objectif est de concevoir des chaînes de valeur entièrement numérisées, autocontrôlées et décentralisées (C. K. M. Lee & Zhang, 2016; Müller & Voigt, 2018).

(Gamache et al., 2017) présentent l'IIoT comme un pilier essentiel de l'industrie 4.0, crucial pour les systèmes cyber-physiques. Il constitue une infrastructure intelligente des technologies de l'information et de la communication, permettant une communication et une coopération en temps réel entre les objets (machines et appareils) et établissant une connexion entre le monde physique et virtuel dans un environnement dynamique (Botta et al., 2015) (Khodkari & Maghrebi, 2016) (Y. Liu et al., 2015).

L'IIoT transforme les données brutes fournies par les objets en informations exploitables, pouvant conduire à des actions concrètes (Kohler Dorothée, 2016). Comme le soulignent (Vermesan, 2013), cette transformation des données en informations actionnables est un aspect clé de l'IIoT, permettant aux entreprises d'optimiser leurs processus et de prendre des décisions plus éclairées.

(Sisinni et al., 2018) mettent en évidence certaines différences entre l'IoT et l'IIoT, comme illustré dans le Tableau 1 ci-dessous :

Tableau 1 Comparaison entre l'IoT grand public et l'IoT industriel

	IoT grand public	**IoT industriel**
Impact	Révolution	Évolution
Modèle de service	Centré sur l'humain	Orienté machine

Statut actuel	Nouveaux appareils et normes	Appareils et normes existants
Connectivité	Ad-hoc (infrastructure non tolérée ; les nœuds peuvent être mobiles)	Modes structurés (nœuds fixes ; gestion de réseau centralisée)
Criticité	Non critique (excluant les applications médicales)	Critique pour la mission (temps, fiabilité, sécurité, confidentialité)
Volume de données	Moyen à élevé	Élevé à très élevé

L'IoT permet la création d'un réseau capable de connecter des objets entre eux, de les surveiller, de les gérer et de les contrôler en échangeant des données et des informations (*The Internet of Things*, 2012)

Ces données sont collectées à l'aide de systèmes embarqués, puis analysées par des logiciels. Le système agit ensuite sur les composants physiques et virtuels connectés (Vermesan, 2013). L'interaction peut être physique (déclenchement d'un mouvement ou d'une tâche) ou virtuelle (via des applications mobiles ou un système d'information central) (LaWell, 2015) (Mattern & Floerkemeier, 2010)

Pour être efficaces dans un environnement IoT, les objets doivent répondre à cinq critères essentiels :

1. Posséder une identité propre (codes-barres, puces RFID, numéros de série)
2. Être capables de communiquer (Wi-Fi, Bluetooth, etc.)
3. Disposer de capteurs (thermomètre, GPS, accéléromètre, etc.)
4. Pouvoir être contrôlés à distance
5. Avoir des capacités d'auto-apprentissage (facultatif mais souhaitable)

L'IoT nécessite des politiques de sécurité et de fiabilité strictes, ainsi que des normes pour faciliter l'interconnexion et l'interopérabilité entre les machines et les logiciels (Botta et al., 2016). Il sert de support pour la connectivité des différents sous-systèmes qui composent un système cyber-physique.

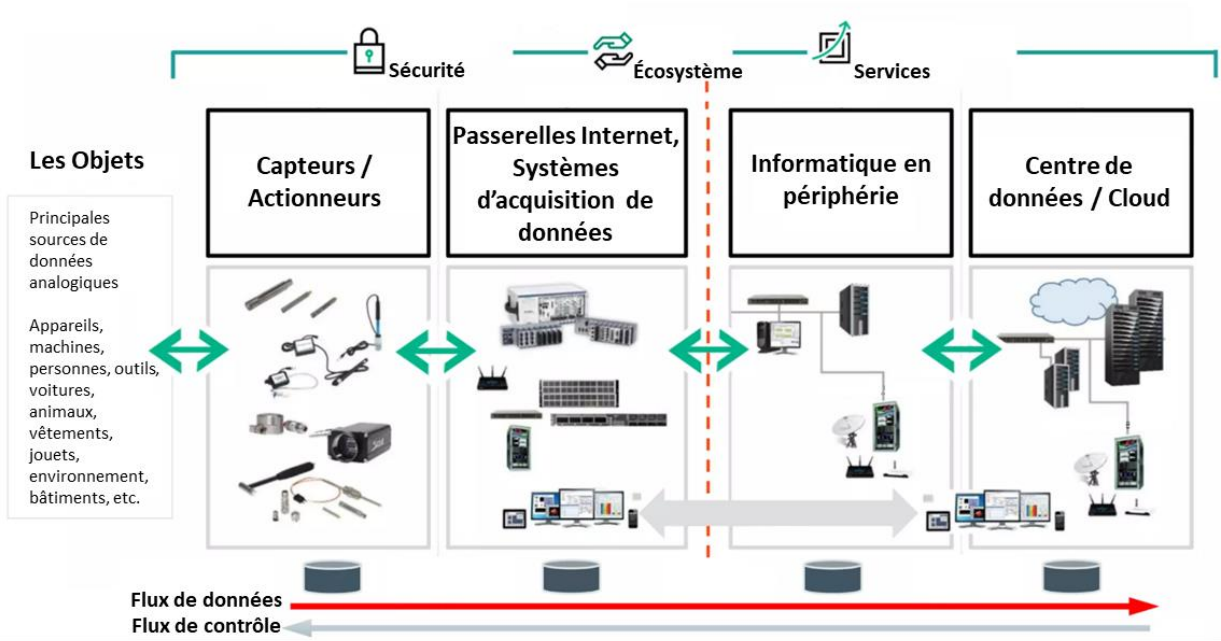


Figure 3 Architecture en quatre étages des systèmes IoT (Hewlett Packard Enterprise, 2016)

(Hewlett Packard Enterprise, 2016) propose une architecture structurée en quatre sous-parties pour décrire les systèmes IoT, comme illustré dans la Figure 3:

1. Capteurs et actionneurs : Ces dispositifs connectés surveillent et influencent les processus réels. Les capteurs recueillent des données sur l'état des processus ou les conditions environnementales. Le traitement des données est de préférence effectué à proximité du système pour minimiser les latences.
2. Systèmes d'acquisition de données : Ils collectent, agrègent et formatent les données envoyées par les capteurs avant de les transmettre via des passerelles Internet (WAN, réseau cellulaire, Wi-Fi). Dans le domaine industriel, où les quantités de données produites sont massives, un filtrage et une compression des données sont souvent nécessaires avant la transmission.
3. Traitement en périphérie (« Edge Computing ») : Les données subissent un traitement supplémentaire avant d'atteindre le centre de données partagé ou le cloud. Des appareils d'« Edge Computing » effectuent des analyses avancées et du pré-traitement pour fournir une capacité de rétroaction rapide. Ce traitement est généralement effectué sur le site qui a produit la donnée, au plus proche du capteur.
4. Centres de données de l'entreprise : Ils effectuent des traitements approfondis à l'aide d'applications conçues et gérées par des professionnels de la data. C'est à ce niveau que les données provenant de plusieurs sites sont agrégées pour obtenir une vision globale du système IoT. Ces centres de données analysent et identifient des tendances générales, des motifs récurrents ou des anomalies pour aider à la supervision des opérations.

Du suivi de fabrication à la fabrication autonome, la mise en place massive de nouvelles capacités de connectivité permet l'implémentation de nouvelles fonctionnalités pour les processus, produits et services (Danjou et al., 2017). (Porter & Heppelmann, 2014) regroupent ces nouvelles capacités en quatre classes, allant du plus simple au plus ambitieux :

- Surveillance
- Contrôle
- Optimisation
- Autonomie

Ces capacités sont interdépendantes : la surveillance constitue la base du contrôle, qui est lui-même essentiel pour permettre l'optimisation. Enfin, la surveillance, le contrôle et l'optimisation sont nécessaires pour atteindre l'autonomie. Par exemple, l'ajustement de l'orientation des pales de l'éolienne pour réguler la puissance générée, illustrant ainsi la capacité d'optimisation.

Il est important de noter que cette classification en fonction des capacités ne correspond pas à un niveau de maturité, et il n'existe pas de classe idéale. Chaque entreprise doit définir les capacités souhaitées de ses processus, produits ou services pour maximiser la valeur apportée à ses clients et définir son positionnement stratégique.

Cette tendance à la hausse est corroborée par les données présentées dans la Figure 4, qui illustre l'évolution du nombre total de connexions de dispositifs, incluant à la fois les objets IoT et non-IoT, de 2010 à 2025. Selon ces projections, le nombre total de connexions devrait atteindre 41,2 milliards en 2025, dont 30,9 milliards seront des dispositifs IoT. Cette croissance significative, avec un taux de croissance annuel composé de 13% entre 2019 et 2025, souligne l'importance croissante de l'IoT dans le paysage technologique mondial et les défis potentiels liés à cette expansion rapide.

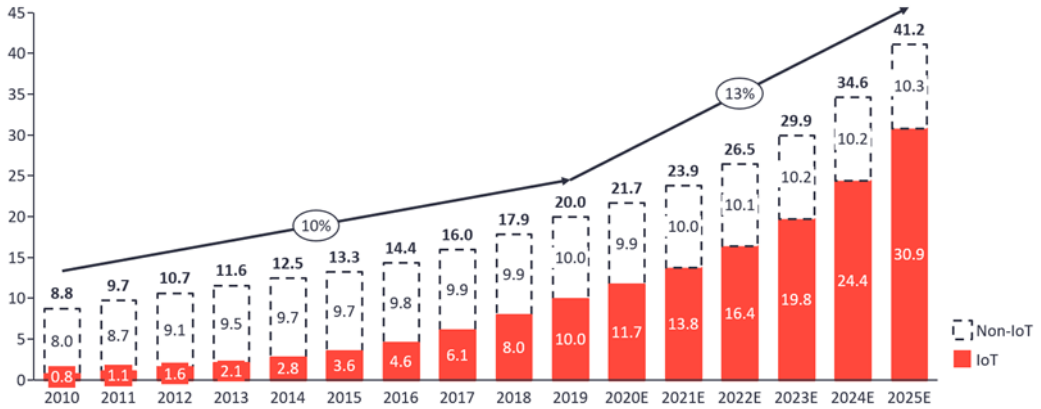


*Figure 4 Évolution du nombre total de connexions de dispositifs (IoT et non-IoT) de 2010 à 2025 (*State of the IoT*, 2020)*

Cette croissance rapide soulève des questions importantes concernant la gestion, la sécurité et l'interopérabilité de ces objets connectés. Il devient de plus en plus complexe de faire communiquer tous ces objets de manière cohérente et sécurisée.

1.2.3 Synthèse

L'industrie 4.0, caractérisée par l'intégration des technologies numériques et de l'intelligence artificielle dans les processus industriels, marque une nouvelle ère dans l'évolution des systèmes de production. Au cœur de cette révolution se trouve l'Internet des Objets Industriels (IIoT), qui permet une interconnexion sans précédent des machines, des capteurs et des systèmes de gestion.

Cette interconnexion offre de nombreux avantages, notamment une meilleure efficacité opérationnelle, une optimisation des processus et une capacité accrue à prendre des décisions basées sur des données en temps réel. L'IIoT permet de collecter et d'analyser des quantités massives de données provenant de multiples sources tout au long du cycle de vie des produits et des systèmes industriels.

Cependant, cette vision holistique du cycle de vie, couplée à la remontée massive de données IoT, soulève plusieurs défis importants :

- Gestion de l'hétérogénéité des données : Les données proviennent de sources diverses (capteurs, machines, systèmes de gestion) et sous des formats variés, ce qui complique leur intégration et leur analyse cohérente.
- Interopérabilité des systèmes : La multiplicité des protocoles de communication et des standards utilisés par les différents dispositifs IoT peut entraver la communication fluide entre les systèmes.
- Sécurité et confidentialité des données : L'interconnexion accrue augmente les risques de cyberattaques et soulève des questions sur la protection des données sensibles tout au long du cycle de vie du produit.
- Scalabilité des infrastructures : L'augmentation exponentielle du nombre d'objets connectés met à l'épreuve les capacités de stockage et de traitement des données.
- Intégration des données dans les processus décisionnels : Transformer efficacement les données brutes en informations exploitables pour améliorer les processus tout au long du cycle de vie du produit reste un défi.
- Maintenance et mise à jour des systèmes : Assurer la pérennité et l'évolutivité des systèmes IoT sur l'ensemble du cycle de vie des produits nécessite des stratégies de maintenance et de mise à jour complexes.
- Compétences et formation : L'exploitation efficace de ces nouvelles technologies requiert des compétences spécifiques qui ne sont pas toujours disponibles dans les entreprises traditionnelles.

Ces défis soulignent la nécessité de développer des approches intégrées et des solutions innovantes pour exploiter pleinement le potentiel de l'IIoT dans une perspective de cycle de vie complet des produits et systèmes industriels. La résolution de ces problématiques sera cruciale pour réaliser pleinement la promesse de l'industrie 4.0 et créer des systèmes de production véritablement intelligents et adaptatifs.

1.3 Le Jumeau Numérique

Le Jumeau Numérique (« Digital Twin ») correspond à l'avatar numérique d'un objet physique, qui l'accompagne non seulement au cours de son cycle de conception et de fabrication, mais également tout au long de son cycle de vie. En se basant sur l'expertise métier liée à l'objet, ainsi que sur les informations recueillies par des capteurs, il devient possible de vérifier les modèles définis pendant le développement du produit par rapport à des mesures réelles, de superviser les performances, et également de réactualiser ces modèles ainsi que les actions à réaliser en fonction de ces données. Cela permet, par exemple, de prolonger son maintien en conditions opérationnelles ou de planifier au mieux les opérations de maintenance. Le jumeau numérique peut être considéré comme une extension de l'approche MBSE (Model-Based Systems Engineering).

1.3.1 L'approche Model-Based Systems Engineering

Les entreprises industrielles développent des systèmes industriels de plus en plus complexes. Si pendant longtemps, des documents textuels, sujets à interprétations, étaient utilisés pour la réalisation de tels produits (comme des spécifications, des rapports d'étude, des comptes-rendus ou des rapports d'expérimentations), la complexité croissante des systèmes et le besoin d'innovation ont poussé à l'adoption de l'ingénierie des systèmes basée sur des modèles, aussi nommée « Model-Based Systems Engineering » (MBSE). Le MBSE est une approche de l'ingénierie utilisant la modélisation comme base afin de faciliter les phases de spécification des besoins, de conception, d'analyse et de vérification, ceci pour une exploitation tout au long du cycle de vie du produit.

Le MBSE est utilisé dans de nombreux domaines tels que l'automobile (Nouacer et al., 2016), les villes intelligentes (Hause et al., 2018), l'aérospatiale (Weiss et al., 2019) ou l'industrie ferroviaire (B. Chen et al., 2018).

Le concept de Model-Based Systems Engineering (MBSE) a été inventé par (Wymore, 1993) avec pour objectif de donner aux étudiants les bases théoriques fondamentales sur lesquelles s'appuient les systèmes d'ingénierie, d'expliciter les théories mathématiques à la base des développements de modèles et d'introduire les considérations à prendre en compte lors de l'application de la théorie à la conception de systèmes réels.

Néanmoins, ce terme est resté peu utilisé jusqu'à sa reprise par l'International Council on Systems Engineering dans sa (Systems Engineering Vision 2020, 2007) qui lui a donné une première définition formelle. Le MBSE est alors défini comme 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 conceptuelle et se poursuivant tout au long du développement et des phases ultérieures du cycle de vie. » Les modèles sont alors placés au centre du processus de développement, remplaçant les documents jusque-là utilisés dans l'ingénierie.

Cette approche, mettant donc l'accent sur le modèle, voit l'essor de systèmes spécialement prévus à cet effet. Il est possible de citer par exemple, ECAD et MCAD, SysML et UML. Elle permet aux ingénieurs de mieux comprendre les changements liés aux choix de conception et d'analyser ces derniers avant leur développement.

De plus, le MBSE bénéficie grandement de l'ajout de données réelles. Ces données permettent d'avoir une compréhension accrue d'un système, et leur intégration est facilitée dès la conception.

Le Tableau 2 présente une analyse comparative détaillée des avantages du MBSE par rapport à l'approche traditionnelle basée sur les documents.

Tableau 2 Avantages du MBSE par rapport à l'approche basée sur les documents (Hart, 2015)

	Basé sur les documents	**Basé sur le MBSE**
Information	Principalement du texte Diagrammes ad hoc Faiblement couplés, répétés dans plusieurs documents	Constructions visuelles et textuelles Définies une fois et réutilisées Partagées entre les domaines Notation cohérente dans les diagrammes Relations définies
Vues de l'information	Par document	Fournit des points de vue Filtres par domaine, espace de problème, etc.
Mesure de l'impact des changements	S'étend sur plusieurs documents Souvent, les exigences textuelles sont isolées de la structure et du comportement	Les relations définissent les chemins de traçabilité Partie naturelle du processus de modélisation Automatisé par programmation
Mesure de l'intégrité - Complétude, qualité et précision	Par inspection manuelle	Automatisé par programmation Animation des spécifications

Dans l'exemple concret de l'utilisation du MBSE pour la conception d'un avion, les différents systèmes qui composent l'avion (par exemple, les systèmes électriques, hydrauliques, mécaniques, etc.) seraient modélisés à l'aide d'un langage de modélisation tel que SysML ou UML. Ces modèles seraient utilisés pour spécifier les exigences, concevoir, analyser et vérifier chaque système individuellement, mais aussi pour intégrer l'ensemble des systèmes dans l'avion complet. En utilisant le MBSE, les ingénieurs pourraient ainsi modéliser l'ensemble du système d'avion et évaluer les impacts potentiels de changements de conception avant même de réaliser des prototypes physiques.

Dans le monde aéronautique, un modèle en V (Figure 5) est souvent utilisé. Ce processus en V permet, dans les systèmes de grande complexité, de garantir que les exigences de sécurité et de sûreté sont prises en compte tout au long du développement, du déploiement et de la maintenance des systèmes.

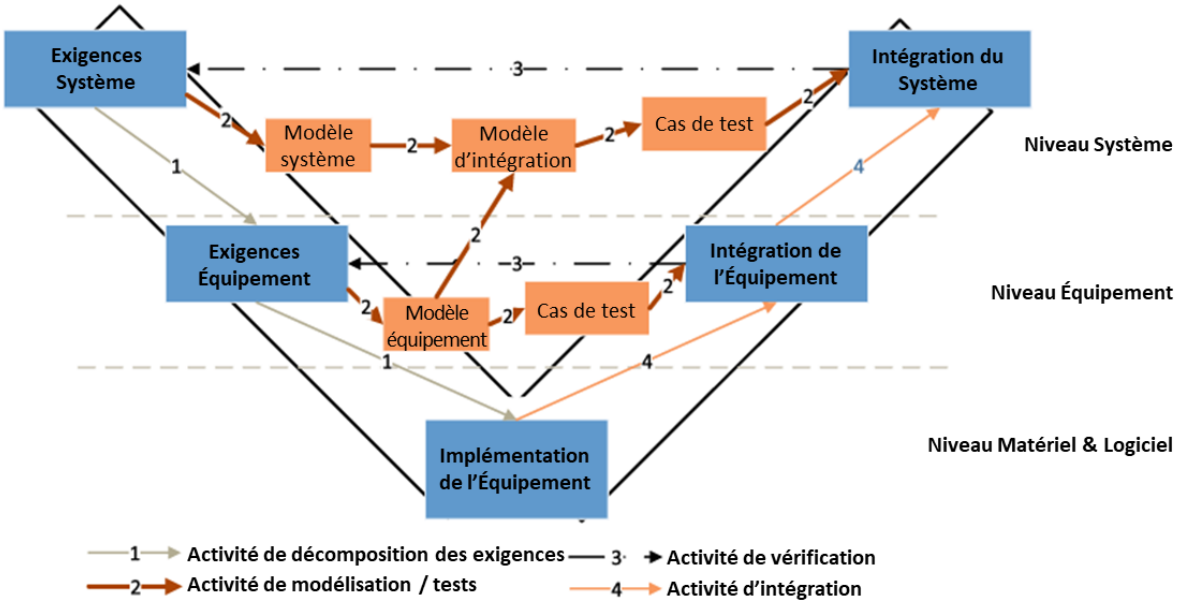


Figure 5 Développement de système pour l'aéronautique basé sur les modèles et les tests (Xing et al., 2021)

De multiples cadres méthodologiques existent pour l'application du MBSE. Un exemple de ces cadres est celui défini par le groupe CESAMES, dont un exemple d'application pour l'aéronautique est montré en Figure 6 et Figure 7.

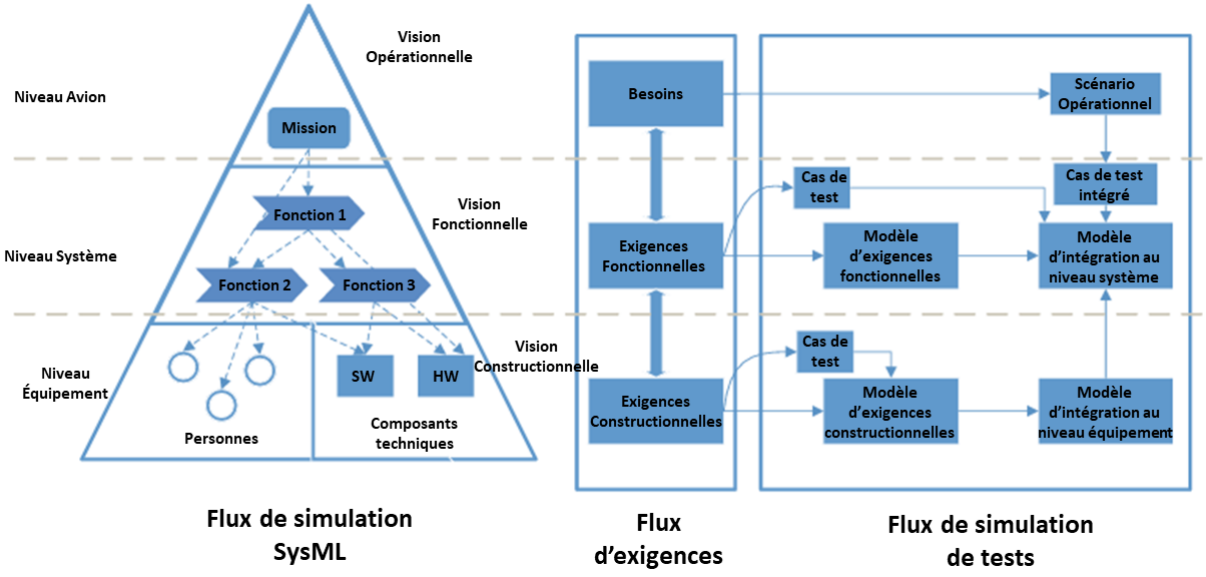


Figure 6 Cadre méthodologie CESAM pour le MBSE (Xing et al., 2021)

Ce cadre méthodologique utilise la modélisation SysML pour analyser et modéliser des systèmes intégrés complexes selon trois visions architecturales génériques : la vision opérationnelle, la vision fonctionnelle et la vision constructive. La vision opérationnelle se concentre sur l'interface avec l'environnement du système et permet de comprendre les besoins des différentes parties prenantes. La vision fonctionnelle analyse le comportement abstrait du système sans considérer les composants utilisés, tandis que la vision constructive se concentre sur les composants physiques du système.

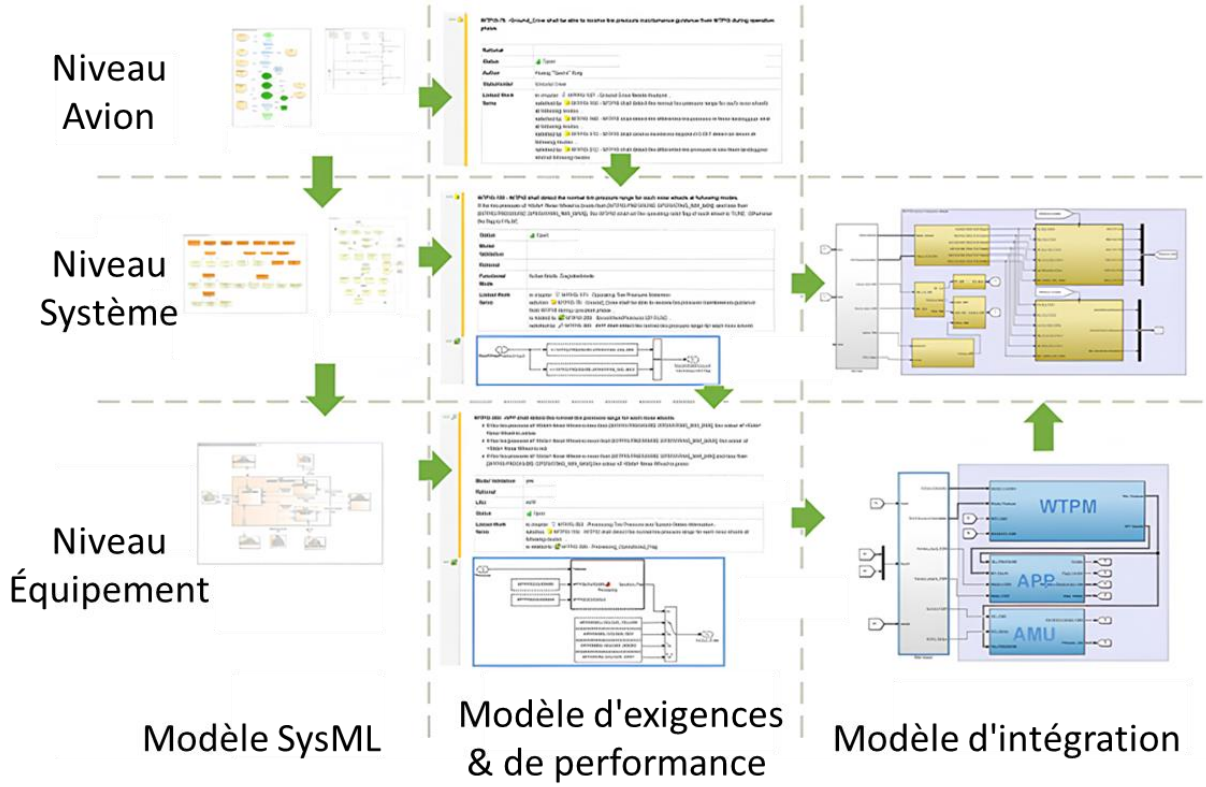


Figure 7 Implémentation de ce cadre méthodologique pour un système de pression des pneus (Xing et al., 2021)

Le MBSE permet d'aider à la simulation d'éléments complexes tels que les trains d'atterrissage. Cette simulation est une tâche complexe en raison du grand nombre de paramètres qui entrent en compte, comme montré dans la Figure 8.

Paramètres de configuration de l'aéronef

Aircraft Attitude	Aircraft Weight (kg)			Aircraft C.G. Position (m)		Aircraft Lift to Weight Ratio		Tricycle Landing Gear Layout(m)	
level landing	MRW	MTOW	MLW	Fwd C.G.	Aft C.G.	landing	drop test	Wheel Base	Wheel Track
	W =71000	W =70500	W =61500	A = 5	A = 5.4	L/W = 1	L/W = 0	D = 6	T = 5

Paramètres de configuration du train d'atterrissage

LG Design Assumptions and Constraints Set				LG Designing Parameters					
				LG Static Load (lb)					
Load Factor	Safety Factor	MG Strut Quantity	Tire Quantity per Strut	NG Maximum Static Load Nose Tire Maximum Static Load			MG Maximum Static Load per Strut Main Tire Maximum Static Load		
n = 3	1.07	N = 2	2	$F_{S,N}$ = 26088 $F_{Sti,N}$ = 13957	Load Condition Fwd C.G.	W = MRW	$F_{S,M}$ = 70438 $F_{Sti,M}$ = 37684	Load Condition Aft C.G.	W = MRW

Paramètres de configuration du pneu et de la jambe d'absorption oléo-pneumatique

Tire and Strut Design Assumptions and Constraints Set				
Tire Polytropic Index	Strut Polytropic Index	Discharge Coefficient	Hydraulic Oil Density (kg/m^3)	Initial Charging Pressure (psi)
r = 1	k = 1.1	C_d = 0.9	ρ_h = 880	P_1 = 375

Tire Designing Parameters							
Tire Rating Load (lb)		Tire Nominal Diameter (inch)		Wheel Nominal Diameter (inch)		Stiffness Coefficient (N)	
nose tire	main tire	nose tire	main tire	nose tire	main tire	nose tire	main tire
F_{rating} = 14340	F_{rating} = 40600	d_n = 30	d_n = 43.5	d_w = 15	d_w = 20	K_n = 7.649*10^5	K_n = 2.004*10^6

Oleo-Pneumatic Shock Absorber Strut Designing Parameters							
Piston Area (m^2)		Chamber Length (m)		Orifice Area (m^2)		Strut Stroke (m)	
nose gear	main gear	nose gear	main gear	nose gear	main gear	nose gear	main gear
A_h = 1.121*10^{-2}	A_h = 3.028*10^{-2}	l = 0.4574	l = 0.4323	A_o = 9.882*10^{-5}	A_o = 2.128*10^{-4}	S = 0.3869	S = 0.3644

Figure 8 Exemple de paramètres pour la configuration des trains d'atterrissage d'un avion (L. Li et al., 2019)

En effet, les trains d'atterrissage doivent être capables de supporter le poids de l'avion lors de l'atterrissage et de fournir une bonne adhérence sur la piste. Les facteurs tels que la taille et la forme des pneus, les propriétés des matériaux, la pression des pneus, la vitesse de l'avion et les conditions météorologiques sont autant de variables qui doivent être prises en compte dans la simulation. Les simulateurs doivent également tenir compte de la variabilité de ces facteurs et de la manière dont ils affectent les performances du train d'atterrissage. En raison de leur importance critique pour la sécurité des vols, la simulation précise des trains d'atterrissage est essentielle pour garantir la fiabilité et la sécurité des avions.

La présence d'un grand nombre de modèles permet également d'obtenir des modèles à plus grande échelle. En faisant varier un paramètre, il est ainsi possible d'observer son impact à travers les chaînes de modèles successifs. Ces nouveaux modèles permettent ainsi d'étendre l'approche MBSE à des systèmes toujours plus grands, de faire évoluer les modèles automatiquement et de simplifier la conception et la validation d'hypothèses de conception, ainsi que l'étude d'alternatives lors de l'analyse de solutions.

Les modèles sont liés entre eux de façon implicite par nature. Par exemple, un modèle de poids sera lié à un modèle de coût : en cas de diminution du poids de l'avion, les coûts de l'avion en seront diminués. Mais ces conclusions sont induites par des relations entre les modèles. Dans l'exemple, la diminution du poids entraîne mécaniquement une baisse de la quantité de carburant nécessaire, ce qui diminue donc le prix d'un vol. Une difficulté est d'être capable de trouver ces interdépendances, de les suggérer, de les mettre à jour et de les maintenir lors des évolutions. Cela permettra de regarder le système à travers d'autres prismes : si je fige un matériau, quels seront par exemple les impacts sur le prix et la complexité d'assemblage ?

Pour améliorer la modélisation de systèmes complexes, la collecte de données réelles est devenue essentielle. Cette nécessité est permise par l'Internet des objets industriels (IIoT), qui utilise des capteurs pour récupérer des données en temps réel à partir de machines et d'équipements. Cela permet ainsi aux entreprises de prendre des décisions plus éclairées et de réduire les risques liés à la conception, à la production, l'efficacité et la sécurité.

1.3.2 Définition du jumeau numérique

Le concept de jumeau numérique (ou Digital Twin) trouve son origine dans les travaux de Michael Grieves. Dans son ouvrage (Grieves, 2016), il explique avoir introduit cette idée lors d'un cours sur le Product Lifecycle Management (PLM) à l'Université du Michigan. Le terme exact de « Digital Twin » est apparu dans (Grieves, 2011), et Grieves l'attribue à l'un de ses collègues, John Vickers de la NASA. À l'époque où ce concept fut introduit, il décrivait la représentation virtuelle des produits comme « nouvelle et immature », ajoutant que les informations collectées sur le produit physique pendant sa fabrication étaient « limitées, collectées manuellement et principalement sur papier ».

Le concept de jumeau numérique, illustré Figure 9, est constitué de trois parties principales : les produits physiques dans l'« Espace Réel », les produits virtuels dans l'« Espace Virtuel », et les flux de données et d'informations qui relient les deux produits ensemble.

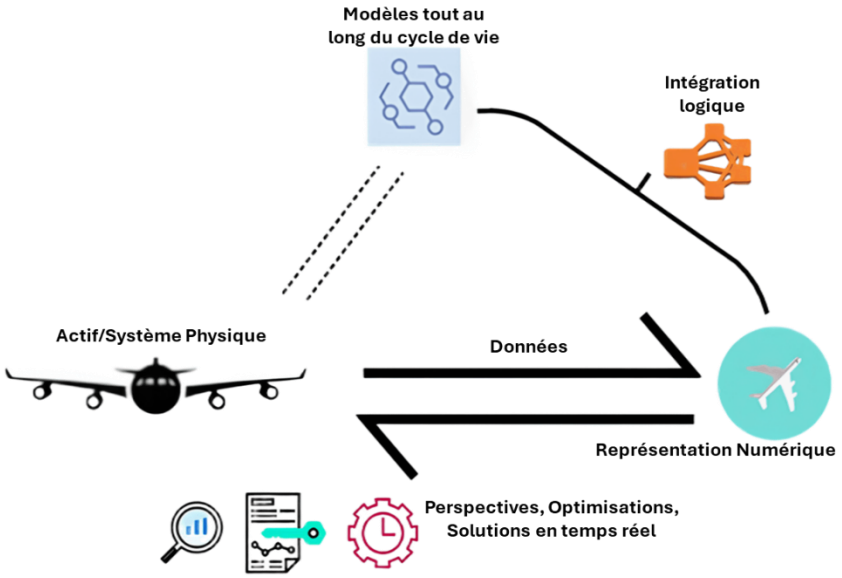


Figure 9 Concept du Digital Twin de (Grieves, 2015)

Le jumeau numérique est donc l'avatar virtuel d'un système physique réel spécifique. Comme Grieves le décrit, un jumeau numérique représente totalement un produit physique, « de l'échelle micro atomique jusqu'à un niveau macro géométrique ». Il identifie également deux grands types de jumeaux numériques :

- Le jumeau numérique prototype, qui contient toutes les informations nécessaires pour la production d'une version physique.
- Le jumeau numérique d'instance, qui décrit un produit spécifique.

Le jumeau numérique d'instance reste lié à son produit physique tout au long de son cycle de vie grâce au lien que les deux systèmes possèdent et entretiennent. Les données de la vie effective du produit viennent enrichir les capacités de son jumeau numérique, ce qui permet à son tour d'affiner les prédictions qu'il sera possible de réaliser sur le produit physique.

Ainsi, à la différence d'une représentation statique, figée dans le temps, telle que pourrait être un modèle, le jumeau numérique dispose aussi d'un aspect temps réel. (Madni et al., 2019) présentent ainsi le jumeau numérique comme le prolongement de l'approche MBSE.

(Jones et al., 2020) présentent certaines caractéristiques des jumeaux numériques dans la recherche. Ils identifient 12 caractéristiques fréquemment présentes ainsi que 7 caractéristiques manquantes qui pourraient être exploitées dans le futur. Ces caractéristiques sont listées dans le Tableau 2.

Tableau 3 12 caractéristiques communes des jumeaux numériques et 7 manquantes (Jones et al., 2020)

	Thème	**Description**
1	Entité Physique	Un artefact du 'monde réel', par exemple un véhicule, un composant, un produit, un système, un modèle.
2	Entité Virtuelle	Une représentation générée par ordinateur de l'artefact physique, par exemple un véhicule, un composant, un produit, un système, un modèle.
3	Environnement Physique	L'environnement mesurable du 'monde réel' dans lequel l'entité physique existe.

4	Environnement Virtuel	Un nombre quelconque de 'mondes virtuels' ou de simulations qui reproduisent l'état de l'environnement physique et sont conçus pour des cas d'utilisation spécifiques, par exemple la surveillance de la santé, l'optimisation du calendrier de production.
5	Fidélité	Le nombre de paramètres transférés entre les entités physiques et virtuelles, leur précision et leur niveau d'abstraction. Les exemples trouvés dans la littérature incluent : entièrement complet, ultra-réaliste, haute-fidélité, données provenant de sources multiples, du niveau micro-atomique au niveau macro-géométrique.
6	État	La valeur actuelle de tous les paramètres de l'entité/environnement physique ou virtuel.
7	Paramètres	Les types de données, d'informations et de processus transférés entre les entités, par exemple la température, les scores de production, les processus.
8	Connexion Physique-Virtuel	La connexion de l'environnement physique à l'environnement virtuel. Comprend les étapes de métrologie physique et de réalisation virtuelle.
9	Connexion Virtuel-Physique	La connexion de l'environnement virtuel à l'environnement physique. Comprend les étapes de métrologie virtuelle et de réalisation physique.
10	Jumelage et Taux de Jumelage	L'acte de synchronisation entre les deux entités et le taux auquel la synchronisation se produit.
11	Processus Physiques	Les objectifs physiques et le processus dans lequel l'entité physique s'engage, par ex. une ligne de production.
12	Processus Virtuels	Les techniques de calcul employées dans le monde virtuel, par exemple l'optimisation, la prédiction, la simulation, l'analyse, la multi-physique intégrée, multi-échelle, simulation probabiliste.
13	Avantages Perçus	Les avantages envisagés réalisés en concrétisant le Jumeau Numérique, par exemple l'amélioration de la conception, du comportement, de la structure, de la fabricabilité, de la conformité, etc.
14	Jumeau Numérique à travers le Cycle de Vie du Produit	Le cycle de vie du Jumeau Numérique (cycle de vie complet, profil numérique évolutif, données historiques)
15	Cas d'Utilisation	Les applications du Jumeau Numérique, par exemple la réduction des coûts, l'amélioration du service, l'aide à la prise de décision.
16	Implémentations Techniques	La technologie utilisée pour réaliser le Jumeau Numérique, par exemple l'Internet des Objets.
17	Niveaux de Fidélité	Le nombre de paramètres, leur précision et le niveau d'abstraction qui sont transférés entre le jumeau virtuel et physique / l'environnement.
18	Propriété des Données	La propriété légale des données stockées dans le Jumeau Numérique.
19	Intégration entre Entités Virtuelles	Les méthodes requises pour permettre la communication entre différentes entités virtuelles.

En 2011, (Tuegel et al., 2011) ont proposé une nouvelle vision pour prédire la durée de vie d'un avion. Lors de la vente d'un avion, son jumeau numérique serait fourni avec. Le jumeau numérique représenterait alors cet avion précisément, et de la façon la plus poussée possible. Y seraient inclus ainsi les anomalies de fabrication, les détails de la microstructure du matériau, jusqu'aux conditions environnementales de fabrication de l'avion. Le jumeau numérique ainsi créé pourrait être utilisé pour simuler le comportement réel de l'avion. Théoriquement, il serait possible de faire voler virtuellement l'avion pendant des heures pour prédire les futurs problèmes qui surviendront sur l'appareil, et effectuer de la maintenance préventive pour éviter les possibles accidents.

Bien entendu, cette vision d'une modélisation aussi précise permettant de prédire de façon déterministe le futur d'un avion est pour l'instant purement théorique, mais elle a le mérite d'exposer trois problèmes clés à résoudre pour permettre l'accomplissement d'une telle idée :

- Le premier évoqué est le problème de modélisation des phénomènes physiques. Plus l'industriel veut pouvoir modéliser de façon réaliste la physique, plus il sera requis d'avoir un modèle physique précis. Cette précision nécessitera à la fois des données plus précises ainsi qu'une puissance de calcul croissante. De plus, la « tyrannie des échelles » posera des problèmes. En effet, les interactions qui se produisent à l'échelle microscopique ne sont pas les mêmes que celles qui se produisent à l'échelle macroscopique, ce qui empêche d’assurer une continuité numérique directe entre les différents modèles.
- Le deuxième problème est la différence entre les capacités permises par le matériel, et les performances effectives du logiciel. Il ne suffira pas d'accroître la puissance de calcul des ordinateurs pour résoudre tous les problèmes de performance, il faudra également disposer d'algorithmes plus performants pour en tirer profit.
- Enfin, le dernier problème est l'inertie des processus de conception dans une telle entreprise. Beaucoup de processus sont éprouvés, anciens, qui souffrent d'une certaine lenteur pour être mis à jour. De plus, une industrie comme l'aéronautique doit porter une attention particulière à la sécurité et à la vérification des produits afin de pouvoir les commercialiser.

(Wagg et al., 2020) proposent une classification des jumeaux numériques en fonction de leurs capacités (Figure 10). Le niveau le plus basique est la supervision de processus, ou la surveillance de l'état de l'usine. De l'ajout de capacités opérationnelles découle la deuxième catégorie, appelée opérationnelle. La catégorie suivante est obtenue en ajoutant au jumeau numérique la capacité de simuler le jumeau physique grâce à des modèles. Nous sommes là proches du modèle présenté par (Tuegel et al., 2011), c'est-à-dire que le jumeau numérique permettra à l'utilisateur d'anticiper les actions à prendre.

Ensuite viennent deux types de jumeaux numériques qui sont pour l'instant davantage dans le domaine de l'anticipation. La catégorie suivante gagne un degré d'intelligence, avec la capacité d'apprendre des données afin d'affiner ses modèles de prédiction. Enfin, la toute dernière catégorie serait autonome, et donnerait au jumeau numérique la capacité de prendre des décisions afin de réduire au minimum les interventions humaines.

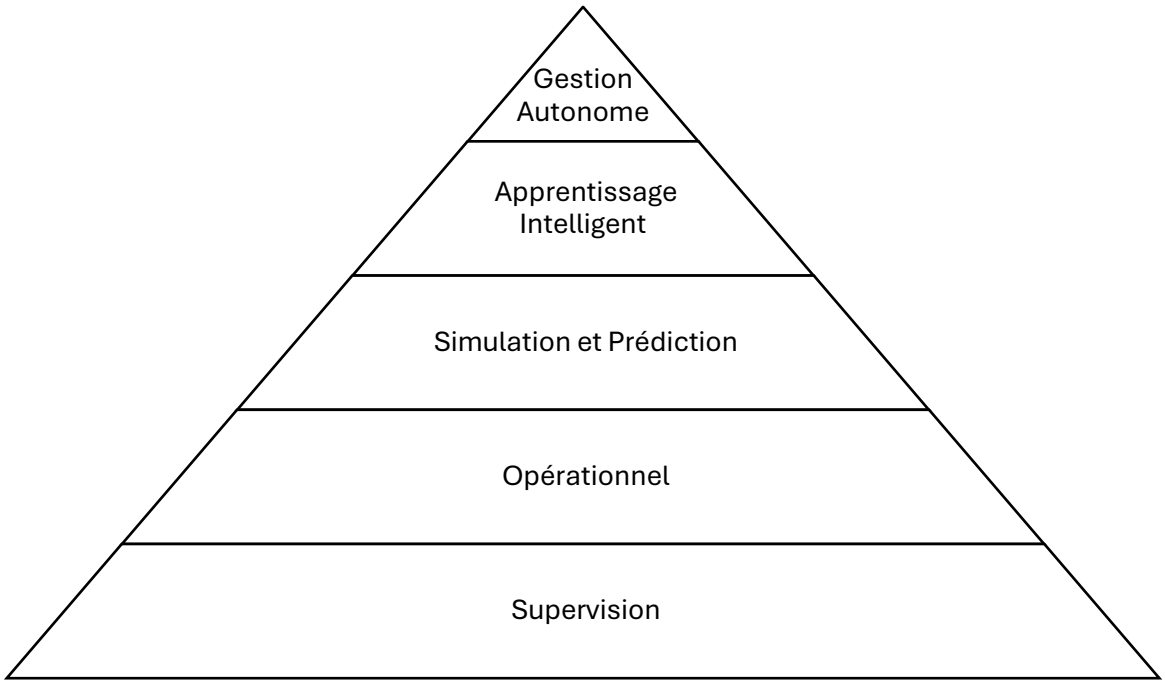


Figure 10 classification des jumeaux numériques par (Wagg et al., 2020)

Le jumeau numérique bénéficie de façon importante des avancées en informatique. Ainsi, l'utilisation des ontologies et des graphes de connaissances (« knowledge graphs ») permet de construire des jumeaux numériques (Banerjee et al., 2017), les techniques d'apprentissage automatique (« machine learning ») permettent de créer des modèles basés sur la physique (Ritto & Rochinha, 2020) et la 5G permet d'intégrer encore davantage jumeau numérique et IoT (Cheng et al., 2018).

Aujourd'hui, le jumeau numérique a été adopté et est utilisé dans de nombreux pans de l'industrie. Des exemples d'applications et d'utilisations peuvent être recensés par exemple dans l'industrie automobile (Damjanovic-Behrendt, 2018), dans l'industrie pétrolière (Sharma et al., 2017) et même dans le management de fermes (Verdouw & Kruize, 2017). Plus proche de notre sujet, se trouvent évidemment des exemples dans l'aviation, comme (Tuegel et al., 2011) cité précédemment, ou (Singh et al., 2021).

Le jumeau numérique est en grande partie possible grâce à l'apport des big data, et cet apport a eu besoin du support, apporté par l'Industrie 4.0, qu'est l'IIoT.

1.3.3 Synthèse

Le jumeau numérique (JN) représente une évolution majeure dans la conception et le suivi des systèmes complexes, en particulier dans l'industrie aéronautique. Il s'appuie sur les principes du Model-Based Systems Engineering (MBSE) pour créer une représentation virtuelle complète et dynamique d'un objet physique tout au long de son cycle de vie.

Le MBSE, introduit par (Wymore, 1993) et formalisé par l'International Council on Systems Engineering (INCOSE), place les modèles au cœur du processus de conception, remplaçant l'approche traditionnelle basée sur les documents. Cette méthode permet une meilleure compréhension des systèmes complexes, facilitant l'analyse des impacts des choix de conception avant même la réalisation de prototypes physiques.

Le jumeau numérique va au-delà du MBSE en intégrant des données en temps réel provenant de capteurs (Internet Industriel des Objets, IIoT) pour maintenir une représentation fidèle et dynamique du système physique. Il permet non seulement de simuler le comportement du système, mais aussi d'anticiper les problèmes potentiels et d'optimiser la maintenance prédictive.

Cependant, la mise en œuvre du jumeau numérique soulève des défis importants liés à sa définition même. Il nécessite l'intégration cohérente de multiples modèles de simulation, devant assurer une représentation fidèle du système physique et simuler son fonctionnement à travers les différentes phases de son cycle de vie.

D'un point de vue industriel, la difficulté majeure réside dans la cohérence nécessaire entre ces différents modèles. En effet, certains modèles existent déjà mais sont partiellement implémentés dans divers outils logiciels, tandis que d'autres doivent être reconstruits. Assurer la cohérence et l'interopérabilité de ces modèles, tout en les maintenant à jour avec les données du système physique, représente un défi technique et organisationnel considérable pour les industries adoptant cette approche.

Cette problématique de cohérence et d'intégration des modèles dans un environnement industriel complexe constitue donc un axe de recherche et de développement crucial pour la réalisation effective des jumeaux numériques dans l'industrie aéronautique et au-delà.

1.4 Mise en œuvre du jumeau numérique

1.4.1 Contexte industriel

Le jumeau numérique est une technologie clé de l'industrie 4.0. Le jumeau numérique est un outil qui permet de regrouper un ensemble de modèles, servant à différents stades du cycle de vie du produit. Il permet de créer une copie numérique d'un objet physique, d'un processus ou d'un système, en utilisant des données collectées par des capteurs et d'autres sources.

Les jumeaux numériques peuvent être utilisés pour simuler et optimiser des processus, pour prévoir des pannes ou des défaillances, ou pour surveiller et améliorer les performances. Ils jouent également un rôle crucial dans la maintenance prédictive, en anticipant les besoins de maintenance avant qu'une panne ne se produise, ce qui peut aider à réduire les temps d'arrêt et les coûts de maintenance.

Le jumeau numérique n'est pas une entité statique, mais est amené à évoluer à mesure que sont créées, inventées ou découvertes de nouvelles sources d'informations et de données pour les modèles. Le jumeau numérique est alors à la fois fournisseur et consommateur de modèles et de données. En reliant ces deux éléments, il est possible de parvenir à affiner toujours plus les hypothèses : les données obtenues sur le terrain sont confirmées par les modèles théoriques, et les modèles sont renforcés par l'apport de nouvelles données expérimentales.

En fin de compte, l'utilisation de ces technologies peut aider les entreprises à améliorer leur efficacité opérationnelle, à réduire les coûts et à améliorer la qualité des produits et services qu'elles fournissent. Néanmoins, ces évolutions sont complexes : leur mise en application est rendue difficile par la multiplication des parties prenantes et par la segmentation et le cloisonnement des activités industrielles.

1.4.2 Les outils MBSE

La mise en pratique du MBSE soulève plusieurs défis, notamment en ce qui concerne la gestion des modèles et la communication entre les différents outils des systèmes d'information (SI) utilisés dans l'industrie. Le jumeau numérique, qui représente une réplique virtuelle d'un objet ou système réel, peut jouer un rôle essentiel dans la réalisation des avantages potentiels du MBSE. Cependant, l'intégration du jumeau numérique dans la pratique du MBSE soulève également certaines problématiques.

Le jumeau numérique repose sur des modèles de systèmes pour représenter l'état actuel et futur d'un objet ou d'un système réel. Dans le milieu industriel, les entreprises ont souvent recours à plusieurs outils pour répondre aux besoins spécifiques de leurs métiers et usages. Cette diversité d'outils peut conduire à un cloisonnement des modèles, rendant difficile la réunion et l'exploitation globale de ces modèles pour les jumeaux numériques. Cette fragmentation entraîne une perte d'efficacité dans la conception et la gestion des systèmes industriels.

La grande disparité entre les outils informatiques constitue un frein majeur à l'adoption du MBSE. Le manque de communication entre les logiciels propriétaires peut entraîner des difficultés pour importer, exporter ou partager des modèles et des données entre différentes applications. Les formats de données, les protocoles de mise à jour et les systèmes de stockage varient souvent d'un outil à l'autre, rendant difficile l'échange et la synchronisation des informations entre les modèles. Cette situation peut nuire à l'efficacité globale du processus de conception, en limitant la capacité des équipes à collaborer et à exploiter pleinement les avantages du MBSE.

L'utilisation fréquente de logiciels propriétaires en MBSE peut poser des problèmes importants en termes de communication entre les outils. Ces logiciels ont tendance à imposer leurs propres formats de données et normes, ce qui rend difficile l'échange et la synchronisation des informations entre les différents outils utilisés au sein d'une même organisation ou entre partenaires. Ces logiciels peuvent également engendrer des coûts supplémentaires pour les licences et l'assistance technique, ainsi qu'une dépendance vis-à-vis du fournisseur du logiciel. Cette dépendance peut limiter la flexibilité et l'évolutivité des solutions MBSE, en restreignant les choix de technologies et d'outils disponibles pour les entreprises.

La présence de modèles en double mais non cohérents est une préoccupation majeure lors de l'implémentation du MBSE. Ce problème peut survenir lorsque des parties de modèles sont copiées et modifiées indépendamment dans différents outils ou par différentes équipes, sans mise à jour appropriée des autres instances. Cette situation conduit à des incohérences entre les modèles, ce qui peut entraîner des erreurs dans la conception et la validation des systèmes industriels.

Les jumeaux numériques nécessitent une mise à jour et une synchronisation constante des modèles pour refléter avec précision l'état réel des objets ou des systèmes. Lorsqu'un modèle est modifié dans un outil spécifique, il est essentiel de mettre à jour les autres modèles en conséquence. Établir des mécanismes de contrôle de version et de synchronisation entre les outils et les équipes, afin de garantir la cohérence et la traçabilité des modèles tout au long du processus de conception est un début de solution. Toutefois, cette tâche s'avère complexe lorsque les modèles sont présents dans différents outils (Figure 11) et que la communication

entre ces outils est limitée. Cette situation augmente les risques d'incohérences et d'erreurs dans les modèles, compromettant ainsi la qualité des systèmes industriels.

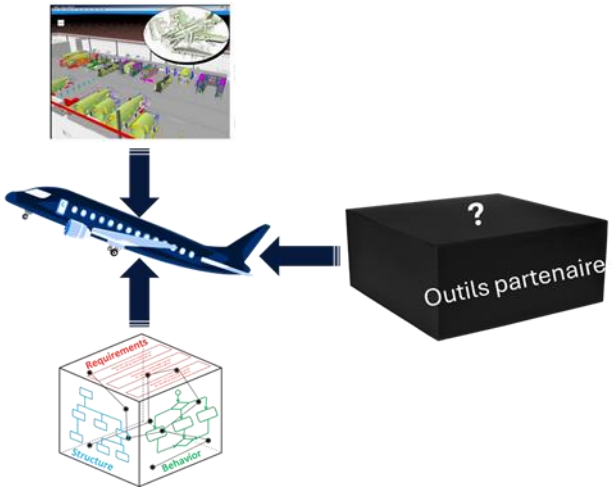


Figure 11 Illustration de la dispersion des modèles dans différents outils MBSE

Il est également essentiel de préserver l'intégrité et la complétude des modèles de systèmes tout au long du processus de conception. Toutefois, certaines parties du modèle initial peuvent être perdues lorsqu'elles ne sont pas correctement conservées ou intégrées dans les outils métiers spécifiques. Cette perte d'information peut nuire à la qualité et à l'efficacité des systèmes développés, ainsi qu'à la traçabilité des exigences et des décisions de conception.

Enfin, le besoin de collaborer avec des partenaires externes ou d'autres entreprises est de plus en plus courant dans l'industrie, ce qui nécessite l'intégration de modèles MBSE provenant de différentes sources. Cette intégration pose des défis supplémentaires en matière de compatibilité, d'échange et de synchronisation des modèles. Lorsque des modèles MBSE sont partagés entre différentes entreprises et partenaires, la confidentialité des données et des informations devient un enjeu crucial. Il est essentiel de mettre en place des mécanismes pour contrôler et limiter l'accès aux modèles, afin de protéger les informations sensibles et de préserver la propriété intellectuelle des parties impliquées.

En abordant ces problématiques, les entreprises peuvent améliorer l'intégration du jumeau numérique dans le MBSE et ainsi maximiser les avantages offerts par cette approche.

1.4.3 L'intégration des données IIoT

L'intégration des systèmes Internet Industriel des Objets (IIoT) est cruciale pour l'implémentation réussie des jumeaux numériques dans le cadre de l'industrie 4.0. Les jumeaux numériques représentent des modèles virtuels précis et à jour des objets physiques ou des systèmes industriels, et leur efficacité dépend en grande partie de la capacité à intégrer et à exploiter les données provenant de divers capteurs et dispositifs IIoT. Toutefois, plusieurs défis se posent en matière d'interopérabilité des systèmes IIoT pour les jumeaux numériques.

Les dispositifs IIoT utilisés dans l'industrie sont souvent hétérogènes, utilisant des protocoles de communication, des formats de données et des architectures logicielles différents. Cette hétérogénéité rend difficile la collecte, l'échange et l'analyse des données provenant de divers dispositifs pour créer et maintenir un jumeau numérique précis et cohérent.

Un aspect clé de l'interopérabilité des systèmes IIoT pour les jumeaux numériques est la capacité à intégrer et à traiter les données en temps réel. Les jumeaux numériques doivent être constamment mis à jour pour refléter l'état actuel des objets physiques qu'ils représentent. Cependant, la collecte, l'échange et l'analyse des données en temps réel peuvent poser des défis techniques et organisationnels, notamment en termes de latence, de bande passante et de synchronisation entre les dispositifs IIoT et les systèmes de jumeaux numériques.

La complexité inhérente des objets physiques peut limiter la capacité à interconnecter les connaissances métier avec les données issues des capteurs. Les entreprises doivent donc trouver des moyens efficaces de modéliser et de représenter ces objets physiques pour créer des modèles précis et complets qui peuvent être exploités par les différents systèmes.

Contrairement au MBSE, les données IIoT n'ont pas été conçues pour être cohérentes par définition. Cette non-cohérence peut entraver l'échange d'informations entre les systèmes et compromettre la qualité des modèles et des analyses. La problématique de modéliser une donnée sans disposer du modèle sur lequel est bâti cette donnée est un réel défi pour la création du jumeau numérique.

La gestion massive de données collectées par les capteurs pose des défis importants en termes de stockage, de traitement et de protection des données. Il est crucial pour les entreprises de développer des systèmes et des processus efficaces pour gérer ces données de manière responsable et efficiente. Cela peut impliquer l'utilisation de technologies pour collecter et analyser les données, ainsi que des processus pour faciliter la communication et l'intégration entre différents systèmes.

Concrètement, les données IIoT sont généralement structurées sous forme de tables, souvent désignées sous le nom générique de « source IIoT ». Une source IIoT correspond typiquement à une table dans laquelle les données sont remontées à partir de capteurs. Ces tables peuvent contenir des informations telles que :

- Un identifiant unique pour chaque enregistrement.
- Un horodatage indiquant le moment de la collecte, du traitement ou de la réception de la donnée.
- L'identifiant du capteur ou du groupe de capteurs.
- La valeur mesurée (par exemple, température, pression, vibration, etc.).
- Des métadonnées supplémentaires (localisation du capteur, type de mesure, etc.).

Il est important de noter qu'une source IIoT peut être consolidée à partir de plusieurs ensembles de capteurs. Par exemple, une table "Température_Usine" pourrait regrouper les données de tous les capteurs de température répartis dans différentes zones d'une usine. Cette consolidation peut se faire au niveau de la collecte des données ou lors de leur traitement.

Les mesures qui devraient être identiques mais qui sont en pratique différentes sont aussi un frein à l'intégration de l'IIoT. Ces différences peuvent être dues à divers paramètres, comme des différences de capteurs utilisés, des différences physiques sur la mesure (une température pouvant par exemple varier selon l'endroit où elle a été prise), la prise en compte des marges des capteurs, ainsi que le transport et formatage de la donnée par le système informatique.

En surmontant ces défis, les entreprises peuvent améliorer l'interopérabilité des systèmes IIoT pour les jumeaux numériques, facilitant ainsi la création et la gestion de jumeaux numériques précis et efficaces dans le cadre de l'industrie 4.0. Cela nécessite une approche holistique qui prend en compte non seulement les aspects techniques, mais aussi les processus organisationnels et les compétences nécessaires pour gérer efficacement ces systèmes complexes.

1.4.4 Synthèse

La mise en œuvre du jumeau numérique dans le contexte de l'industrie 4.0 soulève des défis majeurs, particulièrement en ce qui concerne l'intégration du MBSE (Model-Based Systems Engineering) et des données de l'Internet Industriel des Objets (IIoT). Cette synthèse met en lumière deux problématiques centrales :

1. L'intégration du MBSE dans le jumeau numérique :

   Le MBSE, bien que prometteur pour la conception et la gestion de systèmes complexes, se heurte à des obstacles liés à la diversité des outils utilisés dans l'industrie. Cette diversité entraîne un cloisonnement des modèles, des problèmes de compatibilité entre logiciels propriétaires, et des difficultés de synchronisation et de mise à jour des modèles. Ces défis compromettent la cohérence et l'intégrité des jumeaux numériques, limitant ainsi leur efficacité.

2. L'intégration des données IIoT dans le jumeau numérique :

   L'incorporation des données IIoT dans les jumeaux numériques est essentielle pour refléter l'état réel des systèmes physiques. Cependant, l'hétérogénéité des dispositifs IIoT, la gestion de volumes massifs de données en temps réel, et la nécessité d'assurer la cohérence et l'unicité des données posent des défis considérables. La complexité inhérente des objets physiques et la non-cohérence intrinsèque des données IIoT compliquent davantage cette intégration.

Face à ces enjeux, la problématique industrielle centrale qui émerge est la suivante : Comment établir une corrélation efficace entre l'ensemble des données (qu'elles proviennent des systèmes d'information existants ou de l'IIoT) et l'ensemble des modèles utilisés dans le MBSE et les jumeaux numériques ?

Pour répondre à cette problématique, le besoin d'interopérabilité apparaît comme crucial. L'interopérabilité doit être considérée à plusieurs niveaux :

- entre les différents outils MBSE et les systèmes d'information de l'entreprise ;
- entre les dispositifs IIoT et les systèmes de traitement de données ;
- entre les données collectées et les modèles du jumeau numérique.

En conclusion, relever le défi de l'interopérabilité est essentiel pour permettre une intégration harmonieuse des données et des modèles, facilitant ainsi la création et la gestion de jumeaux numériques précis et efficaces. Cette approche permettra aux entreprises de tirer pleinement parti des avantages offerts par l'industrie 4.0, en améliorant l'efficacité opérationnelle, en réduisant les coûts et en optimisant la qualité des produits et services.

1.5 Problématique industrielle

1.5.1 Interopérabilité : définitions, périmètre

L'interopérabilité est définie par l'Organisation internationale de normalisation (ISO) 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" (International Organization for Standardization, 2015)

De manière plus générale, elle est décrite comme la capacité de deux systèmes (ou plus) à échanger de l'information et à atteindre réciproquement leurs fonctionnalités (Bourey et al., 2007). Dans le cadre de la conception de nombreux systèmes technologiques complexes, la diversité des technologies utilisées et les nombreux systèmes d'information permettant de stocker les données générées rendent cette notion d'interopérabilité incontournable pour parvenir à un système fonctionnel et efficace.

L'interopérabilité facilite la communication et la collaboration entre les différents outils, modèles et systèmes, permettant ainsi de surmonter les obstacles liés à la fragmentation des modèles, à l'hétérogénéité des dispositifs IoT, et à la gestion des données. En améliorant l'interopérabilité, les entreprises peuvent optimiser leurs processus de conception et de gestion, réduire les erreurs et les incohérences, et maximiser les bénéfices de l'industrie 4.0 et du MBSE.

Dans le contexte du jumeau numérique et de l'industrie 4.0, l'interopérabilité doit être prise en compte à deux niveaux principaux : lors de la mise en œuvre du jumeau numérique et lors de l'intégration de données IoT.

1.5.2 Un besoin d'interopérabilité lors de la mise en œuvre du jumeau numérique

Il est crucial d'assurer une communication efficace entre les différents outils des systèmes d'information (SI) utilisés dans l'industrie. Cela implique l'utilisation de normes ouvertes, de formats de données communs et de mécanismes d'échange d'informations pour faciliter l'intégration des modèles et des données entre les différents outils.

Le modèle MBSE initial peut être fragmenté lorsqu'il est réparti dans différents outils métiers et chez des partenaires. Cette fragmentation peut entraîner des incohérences et des pertes d'information. L'interopérabilité est cruciale pour restaurer la cohérence et intégrer les contributions de tous les outils et partenaires dans un modèle MBSE complet et cohérent.

La gestion des modèles et la synchronisation entre les différentes parties prenantes et les équipes impliquées dans le processus de conception sont essentielles pour assurer la cohérence et la traçabilité des modèles. Des mécanismes de contrôle de version et de synchronisation doivent être mis en place pour garantir que les modèles sont constamment mis à jour et cohérents entre les différents outils et équipes.

L'intégration des modèles MBSE provenant de différentes sources, notamment d'autres entreprises ou partenaires, pose des défis supplémentaires en termes de compatibilité, d'échange et de synchronisation des modèles. Des mécanismes de contrôle d'accès et de protection des données doivent être mis en place pour préserver la confidentialité des informations et la propriété intellectuelle des parties impliquées.

1.5.3 Un besoin d'interopérabilité lors de l'intégration de données IoT

L'interopérabilité des données IoT est un aspect essentiel pour le développement et le fonctionnement efficaces des jumeaux numériques. Elle permet d'assurer une intégration harmonieuse des informations provenant de diverses sources dans un modèle cohérent et exploitable.

La cohérence des données est un élément clé de cette interopérabilité. Elle implique que les informations provenant de divers capteurs et dispositifs IoT soient compatibles et harmonisées, permettant ainsi une analyse et une exploitation sans faille. Pour assurer cette cohérence, il est nécessaire d'établir des règles et des procédures pour :

- La synchronisation des données en temps réel
- L'alignement des unités de mesure
- La résolution des conflits entre différentes sources de données

L'unicité de la donnée est un autre aspect crucial de l'interopérabilité. Elle se réfère à la capacité de consolider et de gérer les informations provenant de différentes sources IoT de manière cohérente, en évitant les doublons ou les incohérences. Pour garantir cette unicité, plusieurs processus sont nécessaires :

- La validation des données pour s'assurer de leur exactitude et de leur complétude
- La normalisation des données pour les convertir en un format commun
- La gestion continue des données pour maintenir leur qualité et leur pertinence

L'interopérabilité des données IoT facilite également la mise en place de mécanismes d'échange d'informations efficaces. Ces mécanismes doivent permettre la collecte, l'analyse et l'exploitation des données en temps réel ou en différé, et peuvent inclure :

- Des plateformes d'intégration de données
- Des API standardisées
- Des services cloud facilitant l'échange d'informations

En garantissant l'interopérabilité des données IoT, les jumeaux numériques peuvent fournir des informations plus fiables et précises sur les objets et les systèmes qu'ils représentent. Cela permet une meilleure prise de décision basée sur les données et améliore l'efficacité globale des processus et des opérations liés aux jumeaux numériques.

1.5.4 Synthèse

L'interopérabilité est un enjeu crucial pour l'implémentation réussie du jumeau numérique et l'adoption de l'industrie 4.0. En développant des solutions favorisant l'interopérabilité à différents niveaux, les entreprises peuvent optimiser l'intégration du jumeau numérique dans le MBSE et maximiser les avantages offerts par cette approche. Cela peut conduire à une amélioration significative de l'efficacité opérationnelle, une réduction des coûts et une augmentation de la qualité des produits et services.

Les défis liés à l'interopérabilité des jumeaux numériques, l'intégration des modèles MBSE et des données IoT soulignent l'importance de développer des solutions innovantes. En répondant à ces problématiques, les entreprises pourront tirer pleinement parti des avantages offerts par les jumeaux numériques, le MBSE et l'Internet des objets, renforçant ainsi leur compétitivité et leur performance sur le marché mondial.

La synthèse des points précédents met en lumière plusieurs questions clés :

- **Comment concevoir et mettre en œuvre des jumeaux numériques pour répondre efficacement aux défis d'interopérabilité entre divers systèmes d'information, technologies et domaines métier ?**
- **Quels mécanismes, protocoles et normes adopter pour faciliter l'intégration, la synchronisation et la gestion des modèles MBSE dans les jumeaux numériques, tout en assurant la cohérence, la traçabilité et la confidentialité des informations ?**
- **Comment assurer une intégration transparente et efficace des données provenant de divers capteurs et dispositifs IoT dans les jumeaux numériques, en tenant compte des défis liés à la cohérence, l'unicité, la synchronisation et la gestion des données ?**

Dans le cadre de nos travaux, nous nous focalisons donc sur la problématique industrielle suivante :

Comment établir la corrélation entre l'ensemble des données (qu'elles soient issues des systèmes d'information existants ou de l'IoT) et l'ensemble des modèles utilisés dans le MBSE et les jumeaux numériques, afin de permettre aux utilisateurs métiers d'exploiter efficacement ces données dans leurs activités ?

Cette problématique vise à adresser les défis d'interopérabilité tout en facilitant l'exploitation concrète des données par les utilisateurs finaux dans le contexte des jumeaux numériques.

Dans le chapitre suivant, nous allons explorer cette problématique industrielle afin de construire notre problématique scientifique. Le Chapitre 3 sera constitué d'un état de l'art plus approfondi visant à identifier les bases sur lesquelles nous proposerons, dans le Chapitre 4, nos contributions. Le Chapitre 5 montrera comment la méthodologie et les outils que nous avons développés peuvent être implémentés via un cas d'application représentatif. Enfin, dans la Conclusion générale, nous ferons une synthèse de nos travaux et proposerons des pistes pour le futur.

Chapitre 2 Étude de l'interopérabilité pour la mise à jour du jumeau numérique

2.1 Introduction

L'analyse des enjeux présentés dans le chapitre précédent met en lumière des défis majeurs d'interopérabilité au sein des systèmes de jumeaux numériques, ainsi que dans l'interaction entre les modèles d'Ingénierie des Systèmes Basée sur les Modèles (MBSE) et les données de l'Internet des Objets (IoT). Cette constatation souligne l'importance cruciale de comprendre et d'améliorer l'interopérabilité pour optimiser l'utilisation de ces technologies dans le contexte de l'industrie 4.0.

Les problématiques industrielles clés identifiées peuvent être résumées comme suit :

1. **Jumeaux Numériques** : comment assurer leur interopérabilité avec divers systèmes d'information, technologies et domaines métier ?
2. **Modèles MBSE** : quels mécanismes, protocoles et normes adopter pour faciliter l'intégration, la synchronisation et la gestion des modèles MBSE dans les jumeaux numériques ?
3. **Données IoT** : comment intégrer de manière transparente et efficace les données provenant de divers capteurs et dispositifs IoT dans les jumeaux numériques, tout en garantissant leur cohérence et leur unicité ?

Pour relever ces défis et apporter des réponses pertinentes à ces questions, il est essentiel d'examiner en profondeur la notion d'interopérabilité. Cela implique d'évaluer l'état de l'art actuel, d'identifier les lacunes existantes et d'explorer les pistes d'amélioration potentielles. Dans un contexte industriel en constante évolution, la question de l'interopérabilité revêt une importance capitale pour assurer l'efficacité et la pérennité des systèmes.

Ce chapitre s'articulera autour de plusieurs axes :

- Une définition précise et multidimensionnelle de l'interopérabilité, en explorant ses différentes facettes et niveaux d'application.
- Une analyse des défis et obstacles actuels liés à l'interopérabilité dans le contexte des jumeaux numériques, du MBSE et de l'IoT.
- Un examen des approches existantes pour améliorer l'interopérabilité, en mettant l'accent sur les technologies émergentes et les meilleures pratiques.
- Une réflexion sur les implications de l'interopérabilité pour la gestion des données et des connaissances dans un environnement industriel complexe.

En abordant ces aspects, nous visons à établir les éléments fondamentaux pour développer des solutions innovantes qui répondront aux défis d'interopérabilité identifiés, tout en ouvrant la voie à une exploitation plus efficace et intégrée des jumeaux numériques dans l'industrie 4.0.

2.2 Interopérabilité

L'interopérabilité est un concept fondamental dans le domaine des systèmes d'information et des technologies de communication. Elle représente la capacité des systèmes à collaborer efficacement, échanger des informations et utiliser ces informations de manière cohérente. Plusieurs définitions complémentaires permettent d'appréhender la richesse de ce concept.

(Konstantas et al., 2006) définissent l'interopérabilité comme « l'aptitude de systèmes à pouvoir travailler ensemble sans effort particulier pour les utilisateurs de ces systèmes ». Cette définition met l'accent sur la transparence et la facilité d'utilisation pour l'utilisateur final.

Le (Department of Defense Dictionary of Military and Associated Terms, 2017) des États-Unis propose une définition axée sur les systèmes de communication : "la condition atteinte parmi les systèmes de communications-électroniques ou les équipements de communications-électroniques lorsque les informations ou les services peuvent être échangés directement et de manière satisfaisante entre eux et/ou leurs utilisateurs". Cette définition souligne l'importance de la qualité et de la directivité des échanges.

L'(« IEEE Standard Computer Dictionary », 1991) offre une définition plus générale : « la capacité de deux ou plusieurs systèmes ou composants à échanger des informations et à utiliser les informations qui ont été échangées ». Cette définition met en évidence non seulement l'échange d'informations, mais aussi leur utilisation effective.

L'(International Organization for Standardization, 2015) propose une définition qui souligne l'aspect utilisateur : « la capacité de communiquer, d'exécuter des programmes ou de transférer des données parmi diverses unités fonctionnelles d'une manière qui nécessite que l'utilisateur ait peu ou pas de connaissance des caractéristiques uniques de ces unités ». Cette définition met l'accent sur la transparence du processus pour l'utilisateur.

Pour illustrer concrètement le concept d'interopérabilité, (Wegner, 1996) propose l'exemple des prises électriques. Cet exemple met en lumière la nécessité d'une compatibilité à la fois statique (forme de la prise) et dynamique (tension et fréquence), ainsi que l'utilisation d'adaptateurs pour assurer l'interopérabilité en l'absence de correspondance directe.

(Becker, 2012), quant à lui, illustre l'interopérabilité dans le contexte du cloud computing. Il souligne les défis pratiques et les considérations de marché, tels que la réticence des leaders du marché à promouvoir l'interopérabilité qui pourrait avantager leurs concurrents, ainsi que la nécessité d'une standardisation technique, légale et normative.

Ces définitions et exemples convergent vers une compréhension commune de l'interopérabilité comme une capacité abstraite et de haut niveau, essentielle à l'efficacité et à l'efficience des systèmes modernes. Elle permet une collaboration et une communication sans entrave entre divers systèmes, rendant le concept applicable et pertinent dans une multitude de scénarios.

Pour approfondir la compréhension de l'interopérabilité, il est crucial d'examiner les différentes approches et les méthodes pour les traiter.

2.2.1 Définitions

L'interopérabilité est un concept multidimensionnel qui se manifeste à travers différents niveaux, chacun jouant un rôle crucial dans la facilitation de la communication et de la collaboration entre systèmes hétérogènes. La compréhension de l'interopérabilité s'étend sur plusieurs niveaux, notamment technique, sémantique, syntaxique, applicatif, fonctionnel, et organisationnel. Chacun de ces niveaux est essentiel pour assurer une interaction fluide et efficace entre différents systèmes, particulièrement dans le contexte des jumeaux numériques, des modèles d'Ingénierie des Systèmes Basée sur les Modèles (MBSE), et des données de l'Internet des Objets (IoT).

Le (Référentiel Général d'Interopérabilité V1, 2009) (RGI) distingue trois catégories principales d'interopérabilité : technique, sémantique et syntaxique.

- L'interopérabilité technique aborde les défis liés à la connexion entre différents systèmes. Elle englobe la définition des interfaces, les formats de données, les protocoles de communication et les télécommunications. Cette forme d'interopérabilité décrit la capacité de différentes technologies à communiquer et à échanger des données en se basant sur des normes d'interface clairement définies et largement adoptées.
- L'interopérabilité syntaxique concerne la manière dont les données sont codées et formatées, y compris la nature, le type et le format des messages échangés. La syntaxe est le processus de traduction de la signification en symboles. La relation entre la sémantique et la syntaxe est similaire à celle entre le fond et la forme. Elle mène à la notion de système ouvert, permettant la coexistence d'éléments hétérogènes.
- L'interopérabilité sémantique garantit que la signification précise des informations échangées est compréhensible pour toutes les applications concernées, même si elles n'ont pas été initialement conçues pour cet objectif spécifique. Les conflits sémantiques peuvent survenir lorsque différents systèmes interprètent l'information de manière différente, selon les organisations. Pour atteindre l'interopérabilité sémantique, les parties impliquées doivent se référer à un modèle de référence commun pour l'échange d'informations.

Au-delà de ces trois catégories, il y a d'autres niveaux à considérer pour une interopérabilité complète Comme le soulignent (D. Chen et al., 2008), ces niveaux supplémentaires comprennent :

- Niveaux Application et Fonctionnel : ils se rapportent à la manière dont les applications interagissent et communiquent entre elles, et à la manière dont elles prennent en charge les processus métier spécifiques.
- Niveau Organisationnel : ce niveau se concentre sur l'adaptation des processus métiers et des pratiques organisationnelles pour permettre une collaboration et une interaction efficaces entre différentes entités ou départements.

Les défis de l'interopérabilité ne sont pas uniquement techniques. Ils impliquent également la gestion des incompatibilités conceptuelles et organisationnelles.

(Wache et al., 2001), cités par (Villeneuve et al., 2019) décrivent trois approches permettant d'éviter les problèmes d'incompatibilité entre systèmes, qui sont la principale cause de non-interopérabilité :

- Les approches intégrées sont basées sur la définition d'un format partagé pour tous les modèles et correspondent à l'implémentation d'interfaces dédiées entre un système et le format partagé.
- Les approches unifiées s'appuient sur la caractérisation d'un métamodèle partagé pour la cartographie des concepts basée sur un point de vue sémantique. Un moyen de définir un métamodèle est d'utiliser des ontologies de concepts.
- Les approches fédérées sont généralement plus complexes à implémenter et permettent aux systèmes de s'adapter dynamiquement pour recevoir les données d'autres systèmes. Une ontologie est utilisée pour structurer la connaissance du domaine de travail puis pour gérer une cartographie dynamique entre les concepts des différents systèmes.

Plusieurs travaux existent dans la littérature avec des domaines de connaissances dédiés comme l'agriculture, notamment Agrovoc décrit par (Soergel et al., 2004), ou AgOnt présenté par (Hu et al., 2011). Ces travaux mettent en évidence une problématique globale récurrente lors de l'introduction d'objets connectés et de savoir humain dans des applications à base de connaissances, à savoir l'interopérabilité des différents systèmes en jeu.

Dans le cadre de l'approche intégrée, il faut développer autant d'interfaces de communication que de systèmes, et veiller aux synchronisations éventuelles entre certains de ces systèmes. La production d'ontologies, notamment dans le cadre de l'approche unifiée, auprès de chacun des systèmes principaux permet de faciliter ce travail en procédant à l'alignement de ces ontologies puis en intégrant les résultats de ces alignements par configuration plutôt que par développement lors de l'ajout de nouveaux systèmes. L'approche fédérée, plus complexe, est intéressante lorsque de nombreux acteurs interviennent dans le cadre de collaborations étroites pour le développement, la production et le maintien en conditions opérationnelles d'un environnement complexe d'objets connectés et de systèmes (avion par exemple).

Une seconde problématique qu'il est nécessaire de traiter réside dans la prise en compte des différents aspects de l'interopérabilité (données, service, processus et métier) ainsi que les différents niveaux de problèmes d'incompatibilité (conceptuel, technologique et organisationnel). La problématique de l'interopérabilité est réputée comme ne pouvant pas être résolue uniquement par l'angle technique vu précédemment (c'est-à-dire par la gestion des problèmes technologiques et conceptuels).

En nous basant sur une approche d'ingénierie dirigée par les modèles, « définie comme une approche pour le développement de systèmes qui augmente la puissance du modèle en le déplaçant d'un rôle descriptif à un rôle prescriptif » (Miller & Mukerji, 2003), certains problèmes d'interopérabilité sont étudiés en s'appuyant sur des Architectures Dirigées par les Modèles (Model Driven Architecture - MDA).

(Merlo et al., 2014) proposent la modélisation multi-niveaux suivante pour formaliser ce problème de l'interopérabilité, identifier des solutions pour chaque niveau, et les intégrer au travers d'une démarche structurée :

- Un niveau « métier » afin de caractériser les différentes parties prenantes (humaines et technologiques) du système étudié, leurs rôles, leurs relations, leurs processus métier et les décisions qu'elles prennent.
- Un niveau « système d'information » dédié à l'identification des données, informations et connaissances et aux flux d'informations entre les parties prenantes (en incluant les outils numériques) afin de caractériser les collaborations et les échanges entre les parties prenantes. C'est la description fonctionnelle de l'ensemble de la plateforme numérique.
- Un niveau « technologique » qui permet de prendre en considération les contraintes techniques pour spécifier et implémenter la plateforme numérique intégrant le système d'aide à la décision.

Au niveau "métier", c'est l'ensemble des acteurs métiers participant à la production de connaissances qui est concerné. La façon dont ces acteurs travaillent ensemble dans le cadre de processus collaboratifs doit permettre d'anticiper sur les futurs services métiers à leur proposer une fois les systèmes que nous étudions opérationnels. Il s'agit d'une étape nécessaire s'intégrant dans une démarche d'accompagnement de ces acteurs.

D'autre part, au niveau « système d'information », ce sont l'ensemble des fournisseurs d'objets connectés et de systèmes informatiques qui doivent être considérés afin de prendre en compte les supports adéquats de communication pour concevoir une architecture fonctionnelle. Il est important de noter que ces fournisseurs adoptent souvent leurs propres standards, ce qui souligne la nécessité d'une approche adaptative pour assurer l'interopérabilité entre différents systèmes et standards.

Par conséquent, comme le souligne (Baïna, 2006), les solutions techniques doivent être définies en accord avec les modèles organisationnels à partir des niveaux « métier » et « système d'information ». Cette approche garantit une cohérence entre les aspects techniques et organisationnels du système.

Pour aborder efficacement l'interopérabilité, il est crucial de reconnaître que cette problématique s'étend bien au-delà des aspects techniques. Les approches intégrées, unifiées, et fédérées, telles que décrites, mettent en lumière la nécessité d'une harmonisation à plusieurs niveaux - du partage de formats et de métamodèles à l'adaptation dynamique des systèmes pour une meilleure collaboration. Ces approches, en particulier lorsqu'elles impliquent des domaines aussi vastes que l'agriculture ou des systèmes complexes comme les objets connectés, soulignent l'importance de considérer l'interopérabilité sous plusieurs angles : technique, conceptuel et organisationnel.

Cependant, cette perspective multi-dimensionnelle de l'interopérabilité nous amène à un autre aspect fondamental. Alors que l'interopérabilité syntaxique traite de la structuration et de la transmission des données, l'interopérabilité sémantique se concentre sur leur signification et leur compréhension mutuelle. Cette dernière joue un rôle pivot dans l'échange de données entre systèmes hétérogènes, car elle assure non seulement la cohérence dans la structure des données mais aussi dans la transmission de leur signification réelle. Comme le souligne (Sheth, 1999), l'interopérabilité sémantique devient essentielle pour surmonter les défis de communication et d'intégration, par exemple dans des environnements complexes où coexistent les jumeaux numériques, les modèles MBSE et les technologies IoT.

Ainsi, en passant de la nécessité d'une approche harmonisée pour la gestion de l'interopérabilité à tous les niveaux, nous nous orientons vers l'importance cruciale de l'interopérabilité sémantique. Celle-ci se révèle indispensable pour assurer une compréhension mutuelle et efficace des données, une intégration cohérente des différents systèmes et technologies, et une collaboration entre les différents acteurs métiers et systèmes impliqués. Comme le soulignent (Benson & Grieve, 2016) ainsi que (Yang et al., 2017), l'interopérabilité sémantique concerne non seulement la structure des données (syntaxe), mais aussi la transmission simultanée de leur signification (sémantique) via l'ajout de métadonnées et le recours à un vocabulaire contrôlé et partagé.

L'interopérabilité syntaxique est donc une condition préalable à l'interopérabilité sémantique et concerne les mécanismes de structuration et de transmission des données. Les normes actuelles sont utilisées pour structurer les données de manière que leur traitement soit interprétable à partir de la structure. Cependant, les systèmes doivent également être en mesure de traduire avec précision les informations entre différentes syntaxes.

L'interopérabilité sémantique est particulièrement pertinente pour répondre aux problématiques soulevées précédemment, car elle se concentre sur la compréhension mutuelle des informations échangées entre différents systèmes et technologies. Elle permet notamment d'adresser les problèmes suivants :

- Compréhension mutuelle des données : il est crucial que les différents systèmes et acteurs impliqués comprennent la signification précise des données échangées. L'interopérabilité sémantique garantit que les informations sont interprétées de manière cohérente et correcte par toutes les parties, même si elles proviennent de domaines métier différents ou utilisent des technologies distinctes.
- Réduction des conflits sémantiques : lorsque différents systèmes utilisent des interprétations différentes des informations, des conflits sémantiques peuvent survenir, rendant difficile la communication et la collaboration entre eux. L'interopérabilité sémantique permet de surmonter ces conflits en établissant un modèle de référence commun pour l'échange d'informations, assurant ainsi une compréhension unifiée des données.
- Facilitation de l'intégration des modèles MBSE : les modèles MBSE sont souvent élaborés à partir de différentes normes et langages de modélisation, tels que SysML (Object Management Group, 2019), Modelica (Modelica Association, 2023) et XMI (Object Management Group, 2015). Comme le souligne (Shani, 2017), l'interopérabilité sémantique facilite l'intégration de ces modèles dans les jumeaux numériques en garantissant une compréhension commune des informations et des concepts modélisés, indépendamment des normes utilisées.
- Intégration transparente des données IoT : les données provenant de divers capteurs et dispositifs IoT peuvent présenter des défis en termes de cohérence, d'unicité et de synchronisation. Comme le soulignent (Gyrard et al., 2015) et (Venceslau et al., 2019), l'interopérabilité sémantique permet de garantir une intégration transparente et efficace de ces données dans les jumeaux numériques en assurant que leur signification et leur contexte sont bien compris par tous les systèmes concernés.

L'interopérabilité sémantique, au cœur de notre problématique liée à la collaboration entre les jumeaux numériques, le MBSE et l'IoT, assure une compréhension mutuelle et cohérente des données échangées entre les différents systèmes et acteurs impliqués, et revêt une importance particulière pour la communication et l'échange d'informations entre différentes applications. En garantissant que les informations sont interprétées de manière cohérente et précise malgré les différences entre les systèmes et les modèles de données, elle souligne la nécessité de recourir à des méthodes et des techniques facilitant la compréhension et la gestion des connaissances.

2.2.2 Constats

En lien avec les définitions précédemment établies, il est évident que pour aborder l'interopérabilité sous toutes ses formes, une compréhension approfondie du « modèle » au sens large est fondamentale. Ce modèle représente l'architecture sous-jacente permettant la gestion et l'interprétation des données dans un contexte spécifique.

Rappelons le problème industriel posé : un jumeau numérique est essentiellement constitué de modèles connus. Ces modèles facilitent la gestion de données structurées, dont le sens et l'utilité sont clairement définis par un contexte métier spécifique. De plus, ces modèles permettent d'établir des liens conceptuels entre différents ensembles de données, qu'ils appartiennent au même domaine métier ou à des domaines différents, voire à d'autres modèles.

Pour qu'une donnée soit intégrée efficacement dans un jumeau numérique, elle doit préalablement :

a. **Être structurée** : la donnée doit avoir une forme organisée et constante dans le temps qui permette son traitement et son analyse (syntaxe).

b. **Avoir un sens** : la donnée doit être compréhensible et significative dans le contexte donné (sémantique).

c. **Avoir des relations établies** : les liens conceptuels entre les données doivent être clairement définis pour faciliter leur intégration et leur utilisation.

Ces caractéristiques sont essentielles pour qu'un acteur métier trouve une utilité pratique à ces données. En d'autres termes, pour l'utilisateur, une donnée doit pouvoir se transformer en information qui, à son tour, enrichit ses connaissances, comme le souligne (Ralph L. Ackoff, 1989) dans sa hiérarchie de la connaissance.

En abordant la question des données non structurées, dont le sens n'est pas immédiatement apparent et sans relations préétablies, il devient impératif de rechercher des méthodes pour reconstruire une structure, redéfinir un sens et créer des liens. Cela implique une phase de "préparation" des données. Nous devons donc explorer comment passer de la donnée brute à l'information structurée et significative.

Cette transformation implique une série de processus et de techniques, qui seront détaillés dans les sections suivantes de cette thèse, en s'appuyant sur les concepts et les cadres méthodologiques discutés précédemment. Ces processus incluent l'analyse de données, la modélisation sémantique, et le développement de systèmes de liaison et d'intégration, formant ainsi le pont entre la donnée brute et l'information exploitable dans un jumeau numérique.

2.2.3 Synthèse

L'interopérabilité est un concept multidimensionnel essentiel dans le contexte des jumeaux numériques, du MBSE et de l'IoT. Comme le soulignent (Rezaei et al., 2014), elle se décline en plusieurs niveaux, notamment technique, syntaxique et sémantique, chacun jouant un rôle crucial dans la facilitation de la communication et de la collaboration entre systèmes hétérogènes.

Trois approches principales sont identifiées pour résoudre les problèmes d'incompatibilité : intégrée, unifiée et fédérée. Ces approches soulignent l'importance d'une harmonisation à plusieurs niveaux, allant du partage de formats à l'adaptation dynamique des systèmes.

L'interopérabilité sémantique émerge comme un aspect fondamental, assurant non seulement la cohérence dans la structure des données mais aussi dans la transmission de leur signification réelle. Selon (Sheth, 1999), elle est particulièrement pertinente pour répondre aux défis de compréhension mutuelle des données, de réduction des conflits sémantiques, d'intégration des modèles MBSE et d'intégration transparente des données IoT.

Pour qu'une donnée soit efficacement intégrée dans un jumeau numérique, elle doit être structurée, avoir un sens clair et des relations établies. Cette transformation de données brutes en informations structurées et significatives implique une phase de "préparation" des données, nécessitant des processus d'analyse, de modélisation sémantique et de développement de systèmes d'intégration.

En somme, comme le soulignent (Gyrard et al., 2015), l'interopérabilité, en particulier dans sa dimension sémantique, est cruciale pour assurer une compréhension mutuelle et efficace des données, une intégration cohérente des différents systèmes et technologies, dont l'IoT, et une collaboration effective entre les différents acteurs métiers et systèmes impliqués dans l'écosystème des jumeaux numériques.

Dans la section suivante, nous allons étudier comment intégrer différentes sources de données, qu'elles soient issues de systèmes informatiques ou de sources IoT, dans le but de donner ou redonner du sens, c'est-à-dire d'être capable de fournir, à partir des données brutes issues des sources, une information qui puisse être utile à un acteur métier.

2.3 De la donnée à l'information

2.3.1 Traitement manuel des données

La transformation des données brutes en informations exploitables constitue une étape cruciale dans le développement et l'exploitation des jumeaux numériques. Le traitement manuel des données, bien que traditionnel, demeure une approche pertinente dans certains contextes, notamment lorsqu'il s'agit d'interpréter des données complexes ou ambiguës nécessitant une expertise humaine approfondie.

Dans le cadre de l'utilisation des jumeaux numériques et de l'intégration de données IoT, le traitement manuel se manifeste sous diverses formes, chacune jouant un rôle spécifique dans le processus de transformation des données. Le nettoyage et la préparation des données impliquent un examen minutieux par des experts pour identifier et corriger les erreurs, les incohérences ou les valeurs aberrantes, et combler les manques de données. Cette étape est

particulièrement critique dans des domaines tels que l'aéronautique, où la précision des données peut avoir des implications significatives sur la sécurité et les performances, comme le soulignent (Tao, Cheng, et al., 2018).

La catégorisation et l'étiquetage manuels des données facilitent leur utilisation ultérieure en les classant dans des catégories prédéfinies. Cette approche est couramment utilisée dans les environnements industriels complexes, où la contextualisation des données est essentielle à leur interprétation correcte. L'interprétation contextuelle, quant à elle, nécessite l'intervention d'experts du domaine capables de comprendre les nuances et les implications des données dans leur contexte spécifique d'application. (Grieves & Vickers, 2017) mettent en évidence l'importance de cette approche dans leurs travaux.

La création manuelle de relations entre les données est un processus crucial pour établir une vue holistique du système représenté. Cette étape permet d'identifier des corrélations et des causalités qui pourraient ne pas être évidentes dans les données brutes. Enfin, la validation et la vérification manuelles des données intégrées sont essentielles pour garantir la fidélité de la représentation numérique par rapport au système physique réel dans un jumeau numérique.

En ce qui concerne l'interopérabilité, plusieurs approches manuelles sont employées. L'utilisation de formats spécifiques implique la conversion manuelle des données dans des formats propriétaires adaptés. (Pratt, 2001) souligne que l'adoption de formats standards normalisés, tels que STEP (Standard for the Exchange of Product model data) pour l'industrie manufacturière, facilite l'échange de données entre différents systèmes et organisations.

Les approches de couplage manuel, telles que l'interfaçage manuel et l'alignement manuel, jouent un rôle important dans la connexion des sources de données IoT au reste de l'infrastructure. Comme le notent (Atzori et al., 2010), ces méthodes sont particulièrement utiles pour l'intégration de systèmes hérités (« legacy systems »), obsolètes et parfois encore utilisés activement, ou la gestion de projets impliquant plusieurs fournisseurs avec des terminologies divergentes.

Malgré son utilité dans certains contextes, le traitement manuel des données présente des limitations significatives. La consommation importante de temps et de ressources, le risque accru d'erreurs humaines, le manque d'évolutivité face à des volumes croissants de données, et les potentielles incohérences dues à des interprétations variables sont autant de défis à surmonter. De plus, les retards inhérents au traitement manuel peuvent s'avérer problématiques pour les applications nécessitant des mises à jour en temps réel.

Néanmoins, le traitement manuel conserve sa pertinence dans certains scénarios spécifiques. (Kritzinger et al., 2018) mettent en évidence son utilité particulière pour l'initialisation de systèmes automatisés, la gestion de cas complexes ou exceptionnels nécessitant une expertise humaine approfondie, et la validation finale des données intégrées dans le jumeau numérique.

Ainsi, bien que le traitement manuel des données joue un rôle important dans certains aspects du développement et de la maintenance des jumeaux numériques, ses limitations en termes d'efficacité et d'évolutivité soulignent la nécessité d'explorer des approches plus automatisées et évolutives. (Qi & Tao, 2018) soulignent que cette transition vers des méthodes automatisées soulève des questions scientifiques importantes concernant l'optimisation des processus de traitement des données, l'amélioration de la précision et de la fiabilité des systèmes automatisés,

et le développement de techniques d'apprentissage machine capables de gérer efficacement la complexité et la diversité des données IoT dans le contexte des jumeaux numériques.

2.3.2 Traitement automatisé des données

Le traitement automatisé des données joue un rôle crucial dans la transformation des données brutes en informations exploitables pour les jumeaux numériques. Cette section explore les principales approches de traitement automatisé des données pertinentes pour notre problématique : l'exploration de données (data mining), l'exploration de texte (text mining) et les technologies de données massives (Big Data). Comme le soulignent (M. Chen et al., 2014), nous commencerons par examiner la complexité inhérente aux données industrielles, puis nous aborderons le concept de Big Data et ses caractéristiques essentielles.

2.3.2.1 De la complexité des données industrielles

Dans le contexte de l'Industrie 4.0, l'accès en temps réel à toutes les informations pertinentes est devenu essentiel. Cette connectivité permet de naviguer efficacement entre les différents éléments d'un site industriel, qu'il s'agisse d'objets physiques, de logiciels ou de systèmes cyber-physiques. (L. D. Xu et al., 2018) soulignent l'importance de cette interconnexion en considérant les besoins métiers suivants, qui mettent en évidence la nécessité d'un référentiel d'informations unifié pour le site industriel :

1. Accès à des informations contextualisées : par exemple, pour le bâtiment A, quels sont les plans validés pour les robots de plus de 50K € ayant été audités au cours des 3 derniers mois
2. Accès à l'historique et aux cas d'usage : Pouvoir retracer l'historique d'un élément et ses applications dans l'usine, en partant d'un système d'information géographique ou d'un modèle numérique de modélisation des informations de la construction.
3. Capitalisation des informations : Être capable de capitaliser les informations reçues des prestataires lors d'un appel d'offres, non seulement pour la préparation de l'offre en cours, mais aussi pour de futurs appels.
4. Reconfiguration rapide : Être en mesure de reconfigurer un atelier en fonction de l'état du produit fabriqué dans les plus brefs délais, en identifiant tous les impacts potentiels avant d'apporter des modifications aux éléments de l'atelier ou à d'autres services.
5. Accès direct aux informations : Pouvoir accéder directement au manuel d'utilisation et aux informations sur l'équipement en faisant clignoter son QR code directement sur l'équipement dans l'atelier.

Ces exemples soulignent l'importance cruciale d'une utilisation efficace de toutes les données générées via l'Internet des Objets (IoT) par les systèmes cyber-physiques présents sur un site industriel. C'est un enjeu majeur pour l'Industrie 4.0, qui nécessite des approches avancées de traitement et d'analyse des données.

2.3.2.2 Données massives et le concept des 5V

Le terme « données massives », traduction française de « Big Data », désigne des ensembles de données volumineux et complexes dont l'ampleur et la diversité surpassent les capacités des

techniques conventionnelles de gestion et de traitement des données. Selon (Hilbert, 2016), cette définition fait largement consensus au sein de la communauté scientifique.

(Gandomi & Haider, 2015) présentent le concept des 5V comme une méthode couramment utilisée pour décrire les caractéristiques clés du Big Data. Ces 5V sont :

1. Volume : il s'agit de la quantité massive de données générées et stockées. Le Big Data implique la gestion et l'analyse d'ensembles de données qui dépassent souvent les capacités des systèmes de gestion de bases de données traditionnels, se mesurant en téraoctets, pétaoctets, voire en unités plus importantes.
2. Vitesse : ce terme fait référence à la rapidité avec laquelle les données sont générées, collectées et traitées. Dans le contexte du Big Data, les données sont souvent produites en temps réel ou quasi-réel, nécessitant des solutions capables de traiter et d'analyser rapidement ces flux de données.
3. Variété : la variété concerne les différents types et formats de données inclus dans le Big Data. Ces données peuvent être structurées (comme dans les bases de données relationnelles), semi-structurées (comme les fichiers XML), ou non structurées (comme les textes, images, vidéos ou interactions sur les réseaux sociaux).
4. Véracité : ce terme se réfère à la fiabilité et à la qualité des données. Dans le contexte du Big Data, il est crucial de s'assurer de l'exactitude et de la pertinence des données analysées, malgré leur volume et leur diversité.
5. Valeur : la valeur représente l'utilité et la pertinence des informations extraites des données. L'objectif ultime du Big Data est de transformer ces vastes ensembles de données en insights précieux qui peuvent être utilisés pour améliorer les processus, la prise de décision et les résultats commerciaux.

2.3.2.3 Exploitation du Big Data industriel

L'exploitation efficace du Big Data dans un contexte industriel offre de nombreux avantages potentiels. (Cavanillas et al., 2016) mettent en évidence plusieurs de ces avantages, notamment :

1. D'effectuer des analyses prédictives et comportementales
2. D'anticiper la demande du marché
3. De réduire les coûts commerciaux et de gestion des données
4. De faciliter une prise de décision rapide et éclairée
5. De fournir de nouveaux produits et services en examinant le modèle économique

Pour exploiter efficacement le Big Data d'un site industriel, la première étape cruciale n'est pas simplement de collecter des données, mais de collecter les bonnes données. Cela implique de :

1. Structurer les données de manière appropriée en considérant les concepts importants
2. Corréler les données d'un point de vue business
3. Maintenir les données à jour et les faire évoluer au fil du temps

L'objectif est d'établir des corrélations entre les différentes sources de données d'une entreprise. Par exemple, relier une documentation utilisée pour la maintenance avec des logiciels de gestion

des fournisseurs et des modèles numériques de produits commercialisés peut permettre de définir un service à forte valeur ajoutée pour l'entreprise. Cela pourrait se traduire par l'identification d'améliorations potentielles pour le produit en repérant les pannes récurrentes, tout en évitant les pannes liées aux défauts de maintenance préventive.

Il est important de noter que si les données véhiculent intrinsèquement des informations, c'est leur analyse qui rend le système véritablement intelligent. (J. Lee et al., 2015) soulignent que cette analyse approfondie permet de transformer les données brutes en insights actionnables, contribuant ainsi à l'optimisation des processus industriels et à l'amélioration de la prise de décision.

2.3.2.4 Méthodes classiques d'analyse de données

Les méthodes classiques d'analyse de données constituent le fondement de nombreuses techniques modernes d'exploration et d'interprétation des données. Ces approches, bien qu'antérieures à l'ère du Big Data, restent pertinentes et sont souvent intégrées dans des frameworks plus avancés. Comme le soulignent (Hastie et al., 2009), cette section se concentre sur deux catégories principales de méthodes classiques : les méthodes probabilistes et l'inférence statistique.

2.3.2.4.1 Méthodes probabilistes

Les méthodes probabilistes s'appuient sur la théorie des probabilités pour modéliser l'incertitude inhérente aux données et aux processus étudiés. Ces approches permettent d'estimer la probabilité de différents événements, de quantifier l'incertitude associée à ces événements et d'effectuer des prédictions robustes.

(Christopher M. Bishop, 2006) identifie plusieurs modèles probabilistes couramment utilisés, parmi lesquels :

1. Les modèles de Markov : utilisés pour modéliser des systèmes qui évoluent dans le temps, où l'état futur ne dépend que de l'état présent.
2. Les chaînes de Markov cachées : une extension des modèles de Markov où l'état du système n'est pas directement observable.
3. Les modèles graphiques probabilistes : des représentations graphiques de distributions de probabilité, permettant de visualiser et de raisonner sur des relations complexes entre variables.
4. Les réseaux bayésiens : un type spécifique de modèle graphique probabiliste qui permet de représenter des relations de causalité entre variables.
5. Les processus gaussiens : des modèles non paramétriques flexibles, particulièrement utiles pour la régression et la classification.

L'utilisation de ces méthodes probabilistes dans l'analyse de données a conduit à des avancées significatives dans divers domaines, tels que la finance, la médecine, la biologie et l'ingénierie. Elles permettent d'obtenir des informations précieuses sur les tendances, les modèles et les relations cachées au sein des données, facilitant ainsi la compréhension et la résolution de problèmes complexes.

2.3.2.4.2 Inférence statistique

L'inférence statistique est une approche qui vise à tirer des conclusions à partir d'un échantillon de données en utilisant des méthodes mathématiques et statistiques. Elle se divise principalement en deux branches : l'inférence fréquentiste et l'inférence bayésienne.

Selon (Wasserman, 2004), l'inférence fréquentiste repose sur la notion de fréquence relative des événements observés. Elle utilise des techniques telles que :

1. L'estimation ponctuelle : estimation d'un paramètre de population à partir d'un échantillon.
2. Les intervalles de confiance : plage de valeurs probables pour un paramètre de population.
3. Les tests d'hypothèses : évaluation de la validité d'une hypothèse sur un paramètre de population.
4. La régression : modélisation de la relation entre variables.

(Gelman et al., 2015) expliquent que l'inférence bayésienne, quant à elle, utilise la théorie des probabilités pour combiner les connaissances préalables (sous forme de distribution a priori) et les données observées (la vraisemblance) afin de mettre à jour les estimations des paramètres du modèle (la distribution a posteriori). Cette approche permet une intégration naturelle des informations antérieures et une quantification explicite de l'incertitude dans les estimations.

Les méthodes classiques d'analyse de données, telles que les approches probabilistes et l'inférence statistique, fournissent un cadre solide pour extraire des informations et tirer des conclusions à partir de données. Elles offrent plusieurs avantages :

1. Robustesse : ces méthodes ont été largement testées et validées sur une variété de problèmes.
2. Interprétabilité : les résultats de ces méthodes sont souvent plus faciles à interpréter que ceux de certaines techniques d'apprentissage automatique plus complexes.
3. Efficacité computationnelle : pour certains types de problèmes, ces méthodes peuvent être plus efficaces que des approches plus complexes.

Cependant, comme le soulignent (Efron & Hastie, 2016) ces méthodes présentent aussi des limitations, notamment lorsqu'il s'agit de traiter des données à très grande échelle ou hautement dimensionnelles. C'est pourquoi elles sont souvent combinées ou complétées par des techniques plus avancées d'apprentissage automatique et d'intelligence artificielle dans le contexte du Big Data et de l'Industrie 4.0.

2.3.2.5 Exploration de données (Data mining)

L'exploration de données, ou data mining, est un processus d'extraction de connaissances à partir de grandes quantités de données structurées. (Han et al., 2012) soulignent que cette approche est particulièrement pertinente dans le contexte des jumeaux numériques, où elle peut être utilisée pour découvrir des modèles, des tendances et des relations cachées dans les données collectées par les capteurs de l'Internet des Objets (IoT) et d'autres sources.

Dans l'industrie aéronautique, par exemple, (Gholami & Hafezalkotob, 2017) montrent que l'exploration de données peut être appliquée aux données de vol pour identifier des schémas de

comportement anormaux des composants d'un avion. Ces informations peuvent ensuite être intégrées au jumeau numérique pour améliorer les prévisions de maintenance et optimiser les performances de l'appareil.

Parmi les techniques couramment utilisées en exploration de données, se trouvent :

1. L'analyse de regroupement (clustering) : cette technique permet de regrouper des données similaires, ce qui peut être particulièrement utile pour identifier des modes de fonctionnement similaires dans les systèmes industriels complexes, comme l'explique (Jain, 2010).
2. Les règles d'association : utilisées pour découvrir des relations intéressantes entre variables dans de grands ensembles de données.
3. La détection d'anomalies : particulièrement utile pour identifier des comportements inhabituels ou des défaillances potentielles dans les systèmes industriels.
4. L'analyse de séries temporelles : essentielle pour comprendre l'évolution des systèmes au fil du temps et faire des prévisions.

2.3.2.6 Exploration de texte (Text Mining)

L'exploration de texte, ou text mining, est une extension de l'exploration de données appliquée aux données textuelles non structurées. (Aggarwal & Zhai, 2012) expliquent que cette technique est particulièrement pertinente pour extraire des informations à partir de documents techniques, de rapports de maintenance ou de retours d'expérience, qui sont souvent riches en informations mais difficiles à intégrer directement dans les modèles numériques.

Dans le contexte des jumeaux numériques, l'exploration de texte peut être utilisée pour :

1. Analyser les rapports de maintenance et en extraire automatiquement des informations sur les défaillances récurrentes, les procédures de réparation efficaces ou les conditions environnementales associées à certains problèmes, comme le démontrent (Arif-Uz-Zaman et al., 2017).
2. Extraire automatiquement des informations à partir de la documentation technique, permettant ainsi de maintenir à jour les modèles du jumeau numérique en fonction des dernières spécifications ou normes, sans nécessiter une intervention manuelle constante.
3. Analyser les commentaires des utilisateurs ou les rapports d'incidents pour identifier les problèmes émergents ou les opportunités d'amélioration.
4. Extraire des informations pertinentes à partir de brevets ou de publications scientifiques pour alimenter les processus d'innovation.

2.3.2.7 Application des technologies Big Data aux jumeaux numériques

Ayant précédemment défini le concept de Big Data et ses caractéristiques (les 5V), nous allons maintenant examiner comment ces technologies s'appliquent spécifiquement aux jumeaux numériques et quels avantages elles apportent dans ce contexte.

Selon (Tao, Qi, Wang, & Nee, 2019), l'intégration des technologies Big Data dans les jumeaux numériques permet de :

1. Traiter et analyser en temps réel les vastes quantités de données générées par les capteurs IoT et autres sources, offrant ainsi une représentation plus fidèle et actualisée du système physique.
2. Fusionner efficacement des données provenant de sources multiples et hétérogènes, y compris des données structurées (mesures de capteurs) et non structurées (rapports de maintenance), pour créer un modèle plus complet et précis du système physique.
3. Réaliser des analyses prédictives avancées, améliorant ainsi la maintenance prédictive et l'optimisation des processus industriels.
4. Mieux comprendre les interactions complexes entre les différents composants d'un système industriel grâce à l'analyse de grandes quantités de données historiques et en temps réel.
5. Faciliter la prise de décision rapide et informée grâce à la capacité d'analyser de grandes quantités de données en temps réel.

(Qi & Tao, 2018) soulignent que ces capacités offrent des avantages significatifs pour les jumeaux numériques, notamment une meilleure précision des modèles, une capacité accrue à détecter des anomalies ou des tendances subtiles, et une plus grande flexibilité pour s'adapter aux changements dans l'environnement physique.

2.3.2.8 Défis et limitations

Malgré les avantages considérables du traitement automatisé des données, plusieurs défis persistent dans le contexte des jumeaux numériques. Ces défis soulèvent des questions scientifiques importantes et ouvrent la voie à de nouvelles approches, notamment l'utilisation de l'intelligence artificielle (IA).

1. Qualité et fiabilité des données : (Saha & Srivastava, 2014) soulignent que la qualité et la fiabilité des données restent des préoccupations majeures, particulièrement lors de l'intégration de données provenant de sources multiples et hétérogènes. Les erreurs ou incohérences dans les données peuvent conduire à des modèles inexacts ou à des décisions erronées.
2. Complexité croissante des modèles : (Kritzinger et al., 2018) mettent en évidence que la gestion de la complexité croissante des modèles de jumeaux numériques à mesure que davantage de données sont intégrées pose un défi majeur. Il devient crucial de développer des méthodes pour maintenir la cohérence et la traçabilité des informations au sein du jumeau numérique, tout en permettant une mise à jour continue basée sur les nouvelles données collectées.
3. Interprétation des résultats : L'interprétation des résultats issus de ces techniques automatisées nécessite souvent une expertise humaine pour être pleinement exploitée dans le contexte spécifique de l'application. Cela soulève des questions sur la manière de combiner efficacement l'intelligence artificielle et l'expertise humaine.
4. Gestion du volume et de la vélocité des données : Le traitement en temps réel de grandes quantités de données pose des défis en termes d'infrastructure et d'algorithmes capables de gérer ce flux continu d'informations.

5. Sécurité et confidentialité des données : (Tao, Qi, Wang, & Nee, 2019) soulignent qu'avec l'augmentation de la collecte et du partage des données, les questions de sécurité et de confidentialité deviennent cruciales, en particulier dans les environnements industriels sensibles.

Face à ces défis, (J. Lee et al., 2018) proposent que l'intégration de techniques d'intelligence artificielle (IA) et d'apprentissage automatique émerge comme une solution prometteuse. Selon ces auteurs, l'IA peut potentiellement :

- Améliorer la qualité des données en détectant automatiquement les anomalies et les incohérences.
- Gérer la complexité croissante des modèles en identifiant automatiquement les relations pertinentes entre les variables.
- Faciliter l'interprétation des résultats en fournissant des explications intelligibles des décisions prises par les modèles.
- Optimiser le traitement en temps réel des données massives grâce à des algorithmes plus efficaces.
- Renforcer la sécurité des données en détectant les comportements suspects et les tentatives d'intrusion.

Il est important de noter que l'IA ne remplace pas les méthodes traditionnelles, mais les complète et les étend. En réalité, de nombreuses techniques d'IA s'appuient sur des fondements statistiques et probabilistes. L'intégration de l'IA avec les approches classiques permet de tirer partie des forces de chaque méthode, offrant ainsi une boîte à outils plus complète pour relever les défis de l'analyse de données dans l'industrie moderne.

Dans les sections suivantes, nous explorerons comment les techniques d'IA peuvent être appliquées pour enrichir l'analyse des données industrielles, en mettant l'accent sur leur capacité à traiter des données complexes et à grande échelle, tout en reconnaissant l'importance continue des méthodes traditionnelles dans ce processus.

2.3.3 Apports de l'intelligence artificielle

L'intelligence artificielle (IA) émerge comme une solution prometteuse pour relever les défis complexes liés à l'intégration et à l'interprétation des données dans les jumeaux numériques. Son potentiel transformateur réside dans sa capacité à apprendre à partir des données, à s'adapter à des environnements changeants et à traiter des informations complexes et hétérogènes. Dans le contexte de notre problématique industrielle, l'IA offre des perspectives nouvelles pour améliorer l'interopérabilité et la pertinence des jumeaux numériques.

L'un des apports majeurs de l'IA dans ce domaine est sa capacité à gérer et à interpréter de grandes quantités de données hétérogènes. Les techniques d'apprentissage automatique (machine learning, ML) et d'apprentissage profond (deep learning, DL) permettent d'identifier des motifs et des relations complexes dans les données qui pourraient échapper aux méthodes traditionnelles d'analyse. Comme le soulignent (J. Lee et al., 2019), dans l'industrie aéronautique, ces techniques peuvent être utilisées pour analyser simultanément des données de capteurs, des

rapports de maintenance et des informations de vol, offrant ainsi une vue plus complète et nuancée de l'état d'un aéronef.

L'IA contribue également à améliorer la précision et la fiabilité des prédictions dans les jumeaux numériques. (Carvalho et al., 2019) mettent en évidence que les modèles de ML peuvent être entraînés sur des données historiques pour prédire les comportements futurs des systèmes physiques avec une grande précision. Cette capacité est particulièrement précieuse dans des domaines tels que la maintenance prédictive, où l'anticipation précise des défaillances peut conduire à des économies significatives et à une amélioration de la sécurité.

Un autre aspect crucial est la capacité de l'IA à s'adapter en temps réel aux changements dans l'environnement opérationnel. Selon les travaux de (Mnih et al., 2015), les algorithmes d'apprentissage par renforcement peuvent continuellement ajuster les modèles du jumeau numérique en fonction des nouvelles données entrantes, assurant ainsi que le jumeau reste une représentation fidèle et à jour du système physique.

Dans le contexte de l'interopérabilité, l'IA offre des solutions innovantes pour surmonter les barrières sémantiques entre différents systèmes et sources de données. (Gyrard et al., 2018) suggèrent que les techniques de traitement du langage naturel (Natural Language Processing, NLP) peuvent être utilisées pour interpréter et aligner automatiquement les terminologies et les ontologies utilisées dans différents domaines ou par différents fournisseurs, facilitant ainsi l'intégration cohérente des données.

L'IA permet également d'aborder le défi de la qualité et de la fiabilité des données. (Chandola et al., 2009) expliquent que les techniques de ML peuvent être employées pour détecter automatiquement les anomalies et les incohérences dans les données, améliorant ainsi la qualité globale des informations intégrées dans le système informatique. Cette capacité est particulièrement importante dans les environnements industriels où la précision des données est critique pour la prise de décision.

Cependant, l'intégration de l'IA dans un système informatique soulève également de nouvelles questions scientifiques. L'une des problématiques centrales, comme le soulignent (Gunning & Aha, 2019), est la nécessité de développer des modèles d'IA qui soient non seulement précis, mais aussi interprétables et explicables. Dans de nombreux contextes industriels, en particulier dans des domaines sensibles comme l'aéronautique ou la santé, il est crucial de comprendre comment et pourquoi un modèle d'IA arrive à une conclusion particulière.

Une autre question scientifique importante concerne l'adaptation des techniques d'IA aux contraintes spécifiques des environnements industriels. (J. Wang et al., 2018) mettent en avant la nécessité de développer des modèles capables de fonctionner avec des ressources de calcul limitées, de s'adapter à des environnements dynamiques et de gérer efficacement les incertitudes inhérentes aux données du monde réel.

La gestion de la confidentialité et de la sécurité des données dans les systèmes d'IA pour les jumeaux numériques constitue également un défi majeur. Comme le soulignent (Tao, Qi, et al., 2018), il est nécessaire de développer des approches qui permettent d'exploiter pleinement le potentiel des données tout en respectant les réglementations en matière de protection des données et en préservant les avantages concurrentiels des entreprises.

Enfin, l'intégration de l'expertise humaine dans les systèmes d'IA pour les jumeaux numériques reste un défi important. (Kamar, 2016) souligne l'importance de développer des approches qui permettent une collaboration efficace entre les experts humains et les systèmes d'IA, combinant ainsi l'intuition et l'expérience humaines avec la puissance de calcul et d'analyse des algorithmes.

En conclusion, l'IA offre des perspectives prometteuses pour améliorer l'interopérabilité et la pertinence des jumeaux numériques dans le contexte industriel. Cependant, son application efficace soulève de nombreuses questions scientifiques qui nécessitent une recherche approfondie. Ces défis incluent le développement de modèles d'IA interprétables et adaptés aux contraintes industrielles, la gestion de la confidentialité et de la sécurité des données, et l'intégration efficace de l'expertise humaine. Relever ces défis scientifiques est essentiel pour exploiter pleinement le potentiel de l'IA dans l'amélioration des jumeaux numériques et, par extension, dans l'optimisation des processus industriels dans le cadre de l'industrie 4.0.

2.3.4 Synthèse

La transformation des données brutes en informations exploitables est une étape cruciale dans le développement et l'exploitation des jumeaux numériques. Cette section a exploré les différentes approches de traitement des données, allant du traitement manuel aux techniques automatisées et à l'intelligence artificielle (IA).

Le traitement manuel, bien que précis pour certaines tâches complexes nécessitant une expertise humaine approfondie, présente des limitations significatives en termes d'efficacité et d'évolutivité face aux volumes croissants de données générées dans l'industrie 4.0. Comme le soulignent (Tao, Qi, Wang, & Nee, 2019), ces limitations mettent en évidence la nécessité d'adopter des approches plus avancées et automatisées pour le traitement des données à grande échelle.

Les approches automatisées, telles que le « data mining », le « text mining » et les technologies « Big Data », offrent des solutions plus évolutives pour traiter de grandes quantités de données hétérogènes. D'après (M. Chen et al., 2014), ces techniques permettent d'extraire des connaissances précieuses à partir de vastes ensembles de données structurées et non structurées, facilitant ainsi la prise de décision basée sur les données dans les environnements industriels complexes. Cependant, elles soulèvent des défis importants en termes de qualité des données, d'interprétation des résultats et de gestion de la complexité croissante des modèles.

L'intelligence artificielle, notamment à travers le « machine learning » et le « deep learning », émerge comme une solution prometteuse pour surmonter ces limitations. (J. Lee et al., 2019) soulignent qu'elle offre des capacités avancées pour l'analyse prédictive, l'adaptation en temps réel et l'interprétation de données complexes. L'IA permet non seulement d'améliorer la précision et la fiabilité des prédictions dans les jumeaux numériques, mais aussi de gérer efficacement l'interopérabilité entre différents systèmes et sources de données.

Néanmoins, l'intégration de l'IA dans les jumeaux numériques soulève de nouvelles questions scientifiques. (Gunning & Aha, 2019) mettent en avant des défis notamment en termes d'interprétabilité des modèles, d'adaptation aux contraintes industrielles spécifiques, de gestion de la confidentialité et de la sécurité des données, et d'intégration efficace de l'expertise humaine. Ces défis nécessitent une recherche approfondie pour exploiter pleinement le potentiel de l'IA dans l'optimisation des processus industriels.

Ces défis mettent en lumière la nécessité d'une approche plus sophistiquée pour gérer non seulement les données et les informations, mais aussi les connaissances qu'elles véhiculent. C'est dans ce contexte que l'ingénierie des connaissances devient particulièrement pertinente pour notre problématique. Comme l'expliquent (Studer et al., 1998), cette discipline offre des méthodes et des outils pour formaliser, représenter et raisonner sur les connaissances complexes inhérentes aux systèmes industriels, facilitant ainsi l'intégration cohérente des différentes sources d'information dans les jumeaux numériques.

La transition des données brutes aux informations exploitables, puis aux connaissances structurées, représente un continuum essentiel dans le développement de jumeaux numériques efficaces et interopérables. Cette progression nous amène à explorer plus en profondeur le domaine de l'ingénierie des connaissances et son rôle crucial dans la mise en œuvre de l'interopérabilité sémantique, sujet qui sera abordé dans la section suivante.

2.4 De l'information aux connaissances : mise en œuvre de l'interopérabilité sémantique

2.4.1 Introduction à l'ingénierie des connaissances

L'ingénierie des connaissances est un domaine interdisciplinaire qui se situe à l'intersection de l'intelligence artificielle, de l'informatique et de la gestion des connaissances. Comme le soulignent (Studer et al., 1998), elle vise à représenter, structurer, manipuler et exploiter les connaissances humaines dans des systèmes informatiques, avec pour objectif de créer des applications intelligentes et des systèmes d'aide à la décision performants.

Plusieurs principes clés sous-tendent le domaine de l'ingénierie des connaissances :

- Représentation des connaissances : i s'agit de déterminer comment représenter et organiser les connaissances de manière formelle et compréhensible par les machines. (Davis et al., 1993) notent que les méthodes de représentation des connaissances incluent les ontologies, les réseaux sémantiques, les logiques de description et les systèmes de règles.
- Acquisition des connaissances : cela implique la collecte, l'organisation et la structuration des connaissances à partir de diverses sources, telles que les experts humains, les documents, les bases de données et le Web. Selon (Schreiber et al., 1999), les techniques d'acquisition des connaissances comprennent les entretiens avec des experts, l'analyse de textes, la fouille de données et l'apprentissage automatique.
- Raisonnement et inférence : une fois les connaissances représentées et structurées, les systèmes d'ingénierie des connaissances doivent être capables de raisonner sur ces connaissances pour résoudre des problèmes, prendre des décisions et répondre à des questions. (Russell et al., 2010) expliquent que les mécanismes de raisonnement peuvent être basés sur des règles, des modèles probabilistes, des algorithmes de recherche, ou des techniques d'apprentissage automatique plus avancées.
- Interaction homme-machine : les systèmes d'ingénierie des connaissances doivent être capables de communiquer et d'interagir avec les utilisateurs humains de manière naturelle et efficace. (Card et al., 1999) soulignent que cela implique le développement

d'interfaces utilisateur intuitives, de systèmes de dialogue avancés et de techniques de visualisation des connaissances innovantes.

- Maintenance et évolution des connaissances : les connaissances évoluent constamment, il est donc crucial de maintenir et de mettre à jour les systèmes d'ingénierie des connaissances. (Flouris et al., 2008) précisent que cela peut inclure la vérification et la validation des connaissances, la détection et la résolution des conflits, et l'intégration de nouvelles connaissances de manière cohérente et efficace.

L'ingénierie des connaissances fournit ainsi des méthodes et des outils essentiels pour représenter, structurer et manipuler les connaissances humaines, facilitant leur intégration dans les systèmes informatiques complexes tels que les jumeaux numériques. Cependant, comme le font remarquer (Liao et al., 2017), pour permettre une collaboration et un partage de ces connaissances à grande échelle entre différentes applications et systèmes, notamment dans le contexte de l'industrie 4.0, il est crucial d'adopter des technologies et des normes communes.

Cette nécessité d'interopérabilité sémantique nous amène à explorer plus en détail les technologies du Web sémantique, qui offrent un cadre standardisé pour la représentation et l'échange de connaissances à l'échelle du Web, et qui peuvent être adaptées aux besoins spécifiques de l'industrie 4.0 et des jumeaux numériques.

2.4.2 Le Web sémantique

Le Web sémantique, tel que conceptualisé par (Berners-Lee et al., 2001), représente un changement de paradigme dans la création de contenu sur le Web. Les normes du World Wide Web Consortium (W3C) ont été développées pour faciliter l'échange, la fédération et l'inférence de données sémantiques à l'échelle du Web.

Dans le paradigme du World Wide Web traditionnel, une page web est conçue principalement pour être lisible par un être humain. Les utilisateurs échangent des documents hétérogènes (texte, images, HTML), mais c'est à eux d'interpréter ces documents en données compréhensibles. Ces documents sont identifiés par une URL (Uniform Resource Locator), qui spécifie leur emplacement sur le Web.

En revanche, dans le paradigme du Web sémantique, une page web est conçue pour être à la fois lisible par un humain et interprétable par une machine. Les utilisateurs échangent des données sémantiquement structurées dont la signification est prédéfinie. Le Web sémantique repose sur l'utilisation d'URI (Uniform Resource Identifier), chacun identifiant une ressource unique. Les liens entre les ressources sont également définis.

L'exemple de la réservation de vacances illustre bien la différence entre ces deux approches. Dans le Web traditionnel, l'utilisateur doit effectuer manuellement une série d'actions pour planifier ses vacances, tandis que dans le Web sémantique, un agent informatique pourrait automatiser ce processus en interprétant les données sémantiquement structurées.

Comme le soulignent (Gerber et al., 2006), le Web sémantique favorise l'interopérabilité entre les sites internet et le partage de connaissances. Les ressources et les concepts communs sont conçus pour être facilement échangeables.

L'architecture du Web sémantique, telle que présentée dans la Figure 12, illustre les différentes couches technologiques qui le composent. Cette structure en couches permet une construction

progressive et modulaire du Web sémantique, chaque niveau s'appuyant sur les fonctionnalités des niveaux inférieurs.

L'interopérabilité sémantique, qui permet de transmettre le sens des données avec les données elles-mêmes sous forme d'un "paquet d'informations" auto-descriptif, repose sur un vocabulaire partagé et des liens associés à une ontologie. Cette base offre la capacité d'interprétation, d'inférence et de logique par les machines.

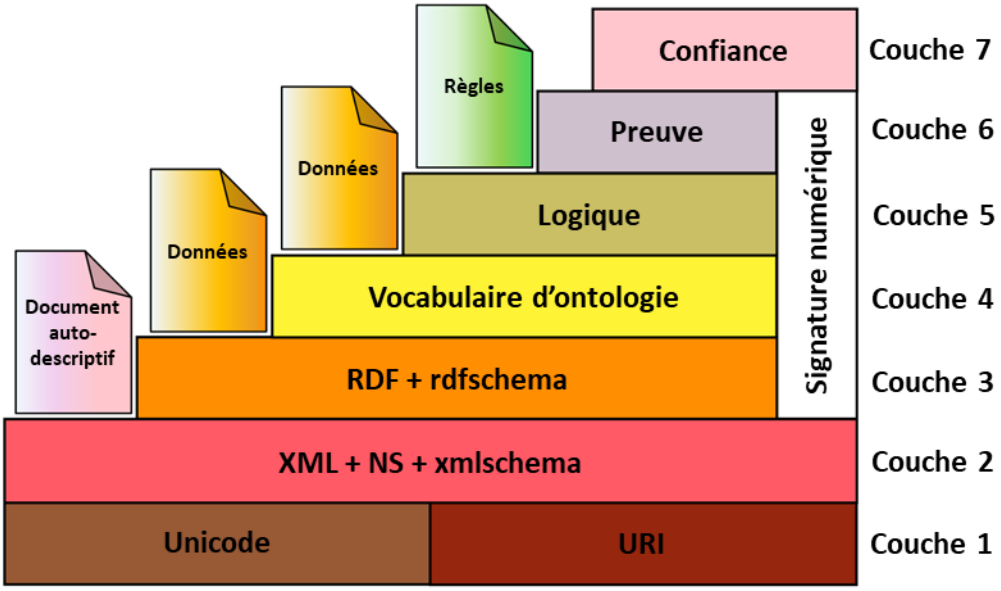


Figure 12 Architecture du web sémantique (Berners-Lee et al., 2001)

Pour généraliser cette interopérabilité, il est nécessaire de disposer de mécanismes capables de décrire et de structurer les connaissances de manière formelle et compréhensible par les machines. Les ontologies, au cœur du Web sémantique, peuvent aider à résoudre les problèmes d'ambiguïtés et de synonymes en fournissant un ensemble de concepts dont les significations et les relations sont stables et acceptées par les utilisateurs.

En adoptant les normes et les technologies du Web sémantique, l'interopérabilité sémantique peut être renforcée, permettant ainsi une meilleure intégration et un partage efficace des connaissances entre les différents acteurs et systèmes.

2.4.3 Les ontologies comme solution pour l'interopérabilité

Le concept d'ontologie, initialement issu de la philosophie où il signifie "l'étude de ce qui est", a été adapté et étendu dans le domaine de l'informatique. Dans ce contexte, une ontologie représente une formalisation structurée des connaissances d'un domaine spécifique.

(Gruber, 1993) propose une définition largement citée qui décrit les ontologies comme "une spécification explicite et formelle d'une conceptualisation partagée". Cette définition met l'accent sur les aspects clés des ontologies : leur nature explicite, formelle, et partagée.

Pour approfondir cette définition, (Antoniou et al., 2012) expliquent qu'une ontologie peut être représentée comme une base de connaissances contenant un ensemble d'entités avec des définitions précises, qu'il s'agisse de types définis ou de concepts (comme "aéronef" et "avion"), formant ainsi la terminologie d'un domaine.

La structure d'une ontologie comprend plusieurs composants essentiels :

1. Classes (ou Concepts) : ce sont les noms des concepts représentés par l'ontologie, comme "avion" ou "aéronef". Elles correspondent à des ensembles d'entités et sont analogues à des prédicats unaires en logique.
2. Propriétés (ou Rôles) : ce sont les caractéristiques ou relations entre les entités. Par exemple, "décollerDe" et "atterrirA" sont des propriétés qui mettent en lien deux entités. En logique, elles correspondent à des prédicats binaires.
3. Instances (ou Individus) : ce sont des entités spécifiques qui instancient les classes. Par exemple, "AéroportDeToulouse" serait une instance de la classe "Aéroport".
4. Axiomes : ce sont des assertions sur les propriétés des classes et des instances. Ils définissent les règles et les contraintes au sein de l'ontologie.

Selon (Baader, 2003), une base de connaissances ontologique se compose généralement de deux éléments principaux :

1. TBox (Terminological Box) : elle contient les définitions des concepts et des propriétés, formant ainsi la structure conceptuelle du domaine.
2. ABox (Assertional Box) : elle contient les assertions sur les instances individuelles en relation avec les concepts et propriétés définis dans la TBox.

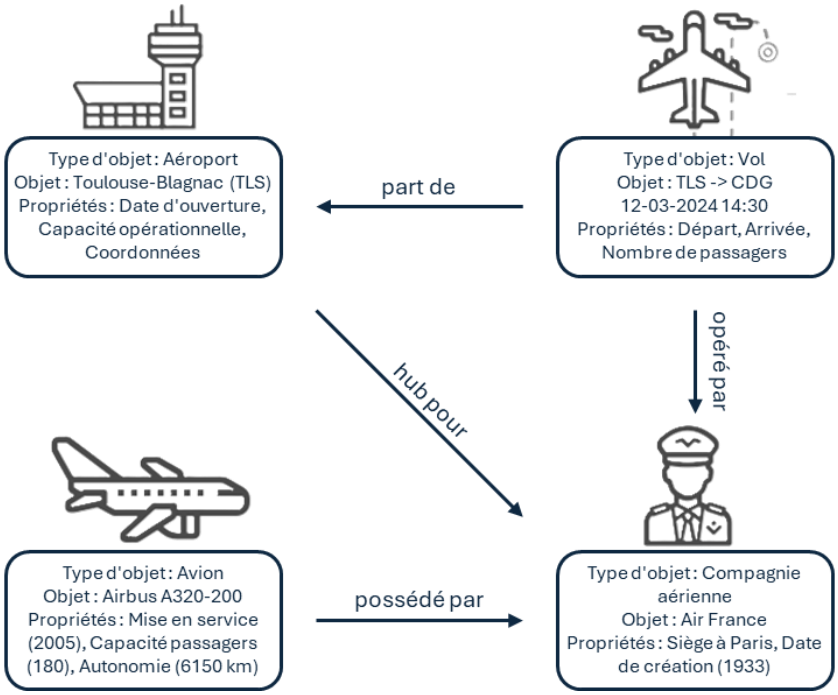


*Figure 13 Une ontologie simple de 4 types d'objets (*Ontology building*, 2024)*

La Figure 13 illustre un exemple d'ontologie simplifiée pour l'industrie aérienne. Cette représentation schématique met en évidence quatre types d'objets principaux - Aéroport, Vol, Compagnie aérienne et Avion - ainsi que leurs propriétés et relations essentielles. Cette ontologie offre un aperçu structuré des éléments clés et des interactions au sein des ensembles de données du secteur aérien, fournissant ainsi un cadre conceptuel pour l'analyse et la modélisation des opérations aériennes.

Dans le contexte des jumeaux numériques et de l'industrie 4.0, (Tao, Zhang, & Nee, 2019) soulignent que les ontologies peuvent jouer un rôle crucial en fournissant une base sémantique commune pour l'intégration des données provenant de diverses sources, telles que les capteurs IoT, les modèles MBSE, et les systèmes d'information existants. Ils affirment qu'elles peuvent aider à résoudre les problèmes d'hétérogénéité sémantique et faciliter l'interopérabilité entre les différents composants d'un système de jumeau numérique.

2.4.4 Synthèse

L'ingénierie des connaissances et les technologies du Web sémantique, en particulier les ontologies, émergent comme des solutions prometteuses pour relever les défis d'interopérabilité dans le contexte des jumeaux numériques et de l'industrie 4.0. Cette synthèse récapitule les points clés abordés et souligne leur importance pour notre problématique.

Comme le soulignent (Staab & Studer, 2009), les ontologies représentent un paradigme puissant pour la modélisation et la représentation formelle des connaissances d'un domaine. Elles offrent plusieurs avantages cruciaux pour l'interopérabilité sémantique :

1. Standardisation du vocabulaire : les ontologies fournissent un cadre commun et normalisé pour décrire les concepts et les relations d'un domaine, réduisant ainsi les ambiguïtés et facilitant la communication entre différents systèmes et acteurs.
2. Représentation formelle et machine-readable : la nature formelle des ontologies permet leur traitement automatique, facilitant l'échange et l'intégration des données entre systèmes hétérogènes.
3. Capacités de raisonnement : les ontologies permettent l'inférence de nouvelles connaissances à partir des connaissances existantes, offrant des possibilités d'analyse et de prise de décision avancées.
4. Réutilisabilité et extensibilité : les ontologies bien conçues peuvent être réutilisées et étendues dans différents contextes, favorisant la cohérence et l'interopérabilité à grande échelle.
5. Alignement et mapping : les techniques d'alignement d'ontologies permettent de relier des concepts similaires entre différentes ontologies, facilitant l'intégration de données provenant de sources hétérogènes.

Dans le contexte spécifique des jumeaux numériques et de l'industrie 4.0, (Tao, Zhang, Liu, et al., 2019) mettent en évidence que les ontologies peuvent jouer un rôle crucial en :

- Fournissant une base sémantique commune pour l'intégration des données provenant de diverses sources (capteurs IoT, modèles MBSE, systèmes d'information existants).
- Facilitant la compréhension mutuelle entre différents domaines d'expertise impliqués dans la conception et l'exploitation des jumeaux numériques.
- Permettant une représentation cohérente des processus métier et des flux de données au sein des systèmes industriels complexes.

Cependant, (Simperl & Luczak-Rösch, 2014) soulignent que la mise en œuvre efficace des ontologies dans un contexte industriel présente plusieurs défis :

1. Complexité de la modélisation : la représentation de domaines de connaissances vastes et dynamiques peut s'avérer complexe et nécessiter une expertise approfondie.
2. Consensus et standardisation : l'établissement d'un consensus entre les experts du domaine pour la définition des concepts et des relations est crucial mais parfois difficile à obtenir.
3. Évolution et maintenance : les ontologies doivent être régulièrement mises à jour pour refléter les changements dans le domaine modélisé, ce qui peut être un processus coûteux en temps et en ressources.
4. Intégration à grande échelle : l'intégration et l'alignement d'ontologies provenant de différentes sources ou domaines restent des défis techniques et organisationnels importants.

En conclusion, les ontologies offrent une solution prometteuse pour relever les défis de l'interopérabilité sémantique dans le contexte des jumeaux numériques et de l'industrie 4.0. Elles fournissent un cadre formel pour la représentation et le partage des connaissances, facilitant ainsi l'intégration et l'exploitation des données provenant de sources hétérogènes. Cependant, leur mise en œuvre efficace nécessite une approche méthodique et collaborative, ainsi qu'une attention particulière aux défis de modélisation, de maintenance et d'intégration à grande échelle.

La prochaine étape consiste à explorer comment ces concepts peuvent être concrètement mis en œuvre dans un contexte industriel, en tenant compte des contraintes et des spécificités propres à l'environnement des jumeaux numériques et de l'industrie 4.0.

2.5 Synthèse globale : Mise en œuvre dans un contexte industriel

L'utilisation des ontologies et des technologies du web sémantique dans l'industrie émerge comme une approche prometteuse pour relever les défis liés à l'interopérabilité, la communication et la gestion des connaissances, particulièrement dans le contexte des jumeaux numériques et de l'industrie 4.0. Cette synthèse explore les applications concrètes de ces technologies et leur potentiel pour améliorer les processus industriels.

2.5.1 Ontologies comme support pour les modèles industriels

Les ontologies se révèlent être un outil puissant pour servir de référentiel conceptuel dans le cadre d'approches d'Ingénierie des Systèmes Basée sur les Modèles (MBSE). Comme l'ont démontré (van Ruijven, 2015) et (Kharlamov et al., 2016), les ontologies s'avèrent particulièrement utiles comme support pour des modèles industriels. Néanmoins, il convient de noter que, selon (Karray et al., 2021), "les recherches actuelles concluent clairement que les ontologies existantes souffrent d'un manque d'interopérabilité", ce qui met en évidence un défi majeur à relever dans ce domaine.

2.5.2 Avantages et applications des ontologies dans l'industrie

Les ontologies offrent plusieurs avantages cruciaux pour l'industrie, notamment :

1. Une représentation structurée et formalisée des concepts et relations au sein d'un domaine spécifique.
2. La facilitation de la collaboration entre divers acteurs en favorisant l'échange d'informations.
3. L'assurance de la cohérence des données à travers différents systèmes et processus.
4. La modélisation et représentation des connaissances et des processus métier, facilitant la création de jumeaux numériques précis et complets.
5. Le support pour la simulation, l'optimisation des processus, la prévision des pannes et l'amélioration des performances.

Il est important de noter que l'utilisation du web sémantique et des ontologies dans l'industrie n'est pas récente. En effet, (Léger et al., 2005) ont rapporté des cas d'usage du web sémantique dans un contexte industriel dès 2003. De plus, (Maier et al., 2003) ont décrit l'usage d'ontologies dans l'industrie automobile comme une forme d'abstraction d'un système physique pour faciliter le changement ou l'ajout de sources de données.

2.5.3 Diversité et évolution des ontologies industrielles

De nombreux domaines industriels ont développé leurs propres ontologies, couvrant des aspects variés tels que les standards industriels sur les données produits (Fraga et al., 2018), le management de la chaîne logistique (Ameri et al., 2020), l'Industrie 4.0 (Kumar et al., 2019, p. 0), et les bancs de test (Pereira et al., 2020). Cette diversité souligne l'importance croissante des ontologies dans divers secteurs industriels, mais pose également des défis en termes d'interopérabilité et d'intégration.

Dans ce contexte, (Shani, 2017) a exploré l'utilisation des ontologies pour prévenir l'obsolescence des modèles MBSE. Les ontologies sont proposées comme un support à l'interopérabilité entre les outils de modélisation, favorisant la réutilisation des modèles et luttant contre leur vieillissement. L'article suggère que les ontologies pourraient servir de langage universel pour représenter les données de modélisation, permettant aux modèles de résister au temps grâce à la réutilisation entre différents outils et langages.

2.5.4 Cycle de vie du jumeau numérique

Le cycle de vie du jumeau numérique est un aspect crucial mais souvent négligé dans la littérature. Il comprend plusieurs phases importantes :

1. Conception : le jumeau numérique peut être créé dès la phase de conception du produit ou du système physique, offrant des opportunités d'optimisation et de simulation précoces.
2. Implémentation : cette phase implique la mise en place de l'infrastructure nécessaire pour synchroniser le jumeau numérique avec son homologue physique.
3. Exploitation : c'est la phase la plus étudiée, où le jumeau numérique est utilisé pour surveiller, analyser et optimiser le système physique en temps réel.

4. Mise à jour : un processus continu qui assure que le jumeau numérique reste une représentation fidèle du système physique tout au long de son cycle de vie.

L'intégration de l'IoT dès les phases de prototypage et de conception permet de collecter des données précieuses pour affiner le jumeau numérique. Comme le soulignent (Tao, Zhang, Liu, et al., 2019), cette correspondance entre les phases de vie du jumeau numérique et celles du système physique est essentielle pour maximiser la valeur ajoutée du jumeau numérique à chaque étape, de la conception initiale à la fin de vie du produit ou du système.

2.5.5 Développement et mise en œuvre de jumeaux numériques basés sur les ontologies

Dans leurs travaux, (Meyer et al., 2020) ont examiné l'importance des ontologies pour les jumeaux numériques dans le génie industriel et l'intelligence artificielle. Ils proposent une approche intégrant les concepts de base des ontologies pour améliorer l'interopérabilité entre les différents systèmes d'ingénierie, sous-systèmes, simulations et solutions basées sur l'IA. Les auteurs suggèrent une architecture multicouche et multiniveau pour les jumeaux numériques des processus de fabrication, s'appuyant sur des normes telles que (*ISA95*, s. d.) et (*IEC 61499*, s. d.).

2.5.6 Intégration de l'intelligence artificielle

L'intégration de l'intelligence artificielle (IA) avec les ontologies et les jumeaux numériques ouvre de nouvelles perspectives pour l'industrie 4.0. Selon (Panetto et al., 2019), l'IA peut améliorer significativement l'exploitation des ontologies et l'efficacité des jumeaux numériques de plusieurs manières :

1. Apprentissage automatique pour l'enrichissement des ontologies : les techniques d'apprentissage automatique peuvent être utilisées pour enrichir et mettre à jour automatiquement les ontologies à partir de nouvelles données, permettant ainsi une adaptation continue aux changements dans l'environnement industriel.
2. Raisonnement basé sur les ontologies : l'IA peut exploiter les connaissances formalisées dans les ontologies pour effectuer des raisonnements complexes, prendre des décisions plus éclairées et optimiser les processus industriels.
3. Traitement du langage naturel : les techniques de NLP (Natural Language Processing) peuvent faciliter l'interaction entre les utilisateurs et les systèmes basés sur les ontologies, rendant l'accès aux connaissances plus intuitif et efficace.
4. Analyse prédictive avancée : en combinant les données des capteurs IoT, les modèles ontologiques et les algorithmes d'IA, il est possible de développer des systèmes de maintenance prédictive plus précis et des optimisations de processus plus efficaces.
5. Automatisation cognitive : l'IA peut automatiser des tâches cognitives complexes en s'appuyant sur les connaissances formalisées dans les ontologies, permettant une prise de décision plus rapide et plus précise dans les environnements industriels.

L'intégration de l'IA avec les ontologies et les jumeaux numériques représente un domaine de recherche prometteur qui pourrait révolutionner la manière dont les industries gèrent leurs connaissances, optimisent leurs processus et prennent des décisions.

2.6 Problématique scientifique

L'étude approfondie des ontologies et de leur rôle dans l'implémentation des jumeaux numériques dans un environnement multi-modèle et multi-métiers a mis en évidence leur potentiel crucial pour assurer l'interopérabilité sémantique entre divers systèmes. Les ontologies facilitent la communication et la compréhension des données et des concepts entre les domaines métier, les technologies et les modèles intégrés au sein du jumeau numérique. Cependant, plusieurs défis émergent de cette approche :

1. La conception et l'alignement de vastes modèles d'ontologies s'avèrent complexes et chronophages, nécessitant des outils automatisés pour faciliter le traitement et l'alignement des ontologies.
2. L'Internet des Objets Industriel (IIoT) génère d'importantes quantités de données, posant des défis d'intégration et d'évolutivité qui requièrent l'utilisation de l'intelligence artificielle (IA) pour une gestion efficace.
3. La combinaison des ontologies avec l'IA offre des perspectives prometteuses pour exploiter pleinement les avantages de l'IIoT et créer des solutions évolutives et performantes.

Dans ce contexte, notre recherche se concentre sur la question centrale suivante :

Comment enrichir un jumeau numérique par l'intégration d'ensembles de données provenant de sources externes, en l'absence de modèles préétablis, pour améliorer l'interopérabilité et la performance dans un environnement multi-modèle et multi-métiers ?

Cette problématique repose sur l'hypothèse que l'absence de modèles préétablis représente un défi unique, nécessitant des approches novatrices pour la gestion et l'intégration des données dans les jumeaux numériques. Pour aborder cette question complexe, nous la déclinons en quatre questions de recherche principales, chacune accompagnée de deux sous-questions :

1. **De quelle manière convertir des données brutes, en particulier non structurées, en informations utiles ?**
    a. Comment traiter efficacement les données non structurées en général ?
    b. Quelles sont les particularités du traitement des données IIoT ?
2. **Comment structurer les informations pour leur intégration dans un jumeau numérique ?**
    a. Quel support utiliser dans la structuration des informations issues de données non structurées ?
    b. En quoi la structuration des données IIoT est-elle spécifique ?
3. **Quels sont les enjeux pour l'intégration d'informations structurées avec des informations non-structurées dans un jumeau numérique ?**
    a. Comment surmonter les défis d'intégration des données non structurées ?
    b. Quelle est la spécificité de l'intégration des données IIoT ?

4. **Comment rendre les informations intégrées accessibles pour les acteurs métier, afin de générer de nouvelles connaissances ?**
    a. Quels supports permettent de faciliter l'accès aux informations pour les acteurs métier ?
    b. Quelles technologies ou approches favorisent l'exploitation et l'exploration des données intégrées ?

Pour répondre à ces questions, nous allons explorer dans le chapitre suivant un état de l'art approfondi, structuré en quatre sections principales correspondant à chacune de nos questions de recherche. Nous examinerons :

1. Les techniques d'intelligence artificielle pour l'analyse et la conversion des données brutes en informations utiles, avec un focus particulier sur le traitement des données non structurées et IIoT.
2. Les méthodologies de construction et d'enrichissement d'ontologies, en mettant l'accent sur les approches adaptées au contexte industriel et aux données IIoT.
3. Les méthodes d'alignement et d'intégration d'ontologies, ainsi que les stratégies pour gérer l'hétérogénéité des données dans un environnement multi-modèle et multi-métiers.
4. Les techniques d'exploration et d'exploitation des ontologies, incluant les méthodes de requêtage sémantique, les outils de visualisation, et les approches d'analyse avancée basées sur les ontologies.

Cette exploration nous permettra d'identifier les techniques et outils les plus pertinents sur lesquels nous construirons notre contribution, visant à développer une approche innovante pour l'enrichissement des jumeaux numériques dans un contexte industriel complexe.

Chapitre 3 État de l'art

3.1 Introduction

Dans le chapitre précédent, nous avons exploré différents travaux liés à la problématique industrielle, ce qui nous a permis de caractériser notre problématique scientifique. L'objectif de ce chapitre est d'étudier en profondeur les approches existantes qui pourraient nous permettre de répondre à la question de recherche principale :

Comment enrichir un jumeau numérique par l'intégration d'ensembles de données provenant de sources externes, en l'absence de modèles préétablis, pour améliorer l'interopérabilité et la performance dans un environnement multi-modèle et multi-métiers ?

Pour aborder cette question complexe, nous structurerons notre état de l'art selon un processus en quatre étapes, reprenant les 4 questions de recherche, et reflétant la transformation progressive des données brutes en informations exploitables par un expert métier, lui permettant de construire de nouvelles connaissances :

Question de recherche 1 : de quelle manière convertir des données brutes, en particulier non structurées, en informations utiles ?

a. Comment traiter efficacement les données non structurées en général ?

b. Quelles sont les particularités du traitement des données IIoT ?

Dans la section 3.2, nous explorerons les techniques d'analyse de données et les outils de l'intelligence artificielle qui permettent d'extraire des informations à partir de données industrielles. Nous examinerons comment les algorithmes d'apprentissage automatique et d'apprentissage profond peuvent être utilisés pour traiter de grands volumes de données non structurées et en extraire des informations pertinentes. Nous nous intéresserons particulièrement aux techniques adaptées aux spécificités des données industrielles et de l'Internet des Objets Industriel (IIoT).

Question de recherche 2 : comment structurer les informations pour leur intégration dans un jumeau numérique ?

a. Quel support utiliser dans la structuration des informations issues de données non structurées ?

b. En quoi la structuration des données IIoT est-elle spécifique ?

Dans la section 3.3, nous étudierons comment les informations extraites peuvent être structurées de manière sémantique pour faciliter leur intégration dans un jumeau numérique. Nous nous concentrerons sur les méthodologies de construction d'ontologies et les approches pour l'enrichissement et l'adaptation d'ontologies existantes. Nous examinerons des approches de construction d'ontologies, s'appuyant sur des ontologies existantes, ainsi que les approches pour une découverte complète d'ontologies, en mettant l'accent sur les méthodes adaptées au contexte industriel.

Question de recherche 3 : quels sont les enjeux pour l'intégration d'informations structurées avec des informations non-structurées dans un jumeau numérique ?

a. Comment surmonter les défis d'intégration des données non structurées ?

b. Quelle est la spécificité de l'intégration des données IIoT ?

Dans la section 3.4, nous aborderons les défis de l'alignement d'ontologies et les techniques d'intégration de données hétérogènes dans un cadre ontologique unifié. Nous nous pencherons sur les méthodes d'alignement automatique et semi-automatique d'ontologies, ainsi que sur les stratégies pour gérer l'hétérogénéité des données dans un environnement multi-modèle et multi-métiers, en vue d'assurer l'interopérabilité sémantique nécessaire à l'intégration efficace des informations dans un jumeau numérique.

Question de recherche 4 : comment rendre les informations intégrées accessibles pour les acteurs métier, afin de générer de nouvelles connaissances ?

a. Quels supports permettent de faciliter l'accès aux informations pour les acteurs métier ?

b. Quelles technologies ou approches favorisent l'exploitation et l'exploration des données intégrées ?

La section 3.5 examinera les méthodes et outils permettant aux acteurs métier d'accéder et d'exploiter efficacement les informations intégrées. Nous étudierons les techniques de requêtage sémantique, les outils de visualisation d'ontologies, et les approches d'analyse avancée basées sur les ontologies. L'accent sera mis sur la façon dont ces technologies facilitent l'extraction de connaissances pertinentes et la prise de décisions, permettant ainsi aux experts de générer de nouvelles connaissances.

En conclusion de ce chapitre, nous synthétiserons les principales approches identifiées et mettrons en lumière les lacunes de la recherche actuelle par rapport à notre problématique spécifique. Cette synthèse nous permettra d'identifier les axes de recherche prometteurs et de positionner notre contribution dans le contexte de l'état de l'art existant.

Cette structure nous permettra d'explorer en détail les approches existantes pour chaque étape de la transformation des données en connaissances exploitables, en mettant l'accent sur les techniques d'IA et les méthodologies de gestion des connaissances les plus pertinentes pour notre contexte industriel. Nous examinerons comment ces différentes approches peuvent être combinées pour répondre à notre problématique d'enrichissement des jumeaux numériques.

3.2 De la donnée brute à l'information : utilisation de l'IA

Comme évoqué dans le Chapitre 2 développant les problématiques industrielles, l'environnement industriel moderne génère un volume sans précédent de données provenant de sources diverses telles que les capteurs de l'Internet des Objets Industriel (IIoT), les systèmes de production, les chaînes d'approvisionnement et les retours clients. Cette quantité massive de données s'accompagne d'une complexité croissante en termes de variété, de vélocité et de véracité, rendant leur analyse et leur exploration de plus en plus difficiles par des moyens traditionnels.(M. Chen et al., 2014) soulignent que la nature hétérogène de ces données, combinée à leur volume et à leur dynamisme, crée un défi majeur pour les approches conventionnelles de traitement de l'information.

Face à cette complexité, les méthodes manuelles ou basées sur des algorithmes simples s'avèrent rapidement insuffisantes, comme nous l'avons souligné précédemment. La multitude de paramètres à prendre en compte, les relations non linéaires entre les variables, et la nécessité de traiter les données en temps réel dans de nombreux cas, exigent des approches plus sophistiquées et automatisées. Selon (J. Lee et al., 2015), il devient crucial de développer des méthodes capables non seulement de gérer de grands volumes de données, mais aussi d'en extraire des informations pertinentes et exploitables de manière efficace.

Le traitement intelligent des données industrielles va au-delà de la simple automatisation. Il implique la capacité à gérer l'incertitude, à faire face à des données incomplètes ou bruitées, et à extraire des informations significatives à partir de sources hétérogènes. Cette intelligence dans le traitement des données est essentielle pour découvrir des schémas récurrents (« patterns ») complexes, identifier des anomalies subtiles, et générer des perspectives (« insights ») qui peuvent échapper à l'analyse humaine ou aux méthodes statistiques traditionnelles. Comme le soulignent (S. Wang et al., 2016), cette approche est fondamentale pour l'amélioration des processus de décision dans l'industrie 4.0.

Dans cette section, nous approfondirons l'exploration des techniques et approches qui permettent de relever ces défis, en mettant l'accent sur les méthodes capables de transformer efficacement des données brutes complexes en informations exploitables dans le contexte industriel. Nous examinerons comment ces approches peuvent être appliquées pour enrichir les jumeaux numériques et améliorer la prise de décision dans un environnement multi-modèle et multi-métiers, répondant ainsi aux enjeux identifiés dans notre analyse de la problématique industrielle.

3.2.1 Introduction à l'intelligence artificielle pour l'analyse de données

L'intelligence artificielle (« IA ») est un domaine en constante évolution, englobant diverses technologies et techniques telles que l'apprentissage automatique (« Machine Learning », ML) et l'apprentissage profond (« Deep Learning »), ainsi que des approches inspirées du monde naturel et du comportement humain. Ces méthodes novatrices améliorent la capacité des machines à résoudre des problèmes complexes et à s'adapter à diverses situations (Russell et al., 2010).

Les systèmes multi-agents, s'inspirant des comportements sociaux observés dans la nature, offrent une approche puissante pour résoudre des problèmes complexes. Grâce aux interactions entre différents agents autonomes, ces systèmes permettent de simuler des phénomènes tels que la coopération, la compétition et la négociation (Wooldridge, 2009). (Ferber, 1999) souligne

leur utilité particulière dans les contextes nécessitant une décentralisation et une répartition des tâches pour une résolution efficace des problèmes.

La logique floue, introduite par (Zadeh, 1965), s'inspire du raisonnement humain pour traiter l'incertitude et l'ambiguïté en utilisant des degrés d'appartenance plutôt que des valeurs binaires. (Klir & Yuan, 1995) mettent en évidence son efficacité dans le traitement de données incertaines ou incomplètes, permettant ainsi aux machines de prendre des décisions même en l'absence d'informations parfaites.

Les algorithmes génétiques, développés par (Holland, 1992), exploitent les principes de la sélection naturelle pour optimiser les solutions aux problèmes. En simulant des processus évolutifs tels que la sélection, la mutation et la reproduction, ces algorithmes convergent vers des solutions optimales ou quasi-optimales de manière efficace et adaptative (D. E. Goldberg & Holland, 1988). (M. Mitchell, 1996) souligne leur capacité à explorer efficacement de vastes espaces de solutions.

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle qui se concentre sur le développement d'algorithmes et de modèles permettant aux ordinateurs d'apprendre à partir de données et d'améliorer leurs performances sans être explicitement programmés (T. M. Mitchell, 1997). (Alpaydin, 2020) précise que le ML permet aux machines de détecter des schémas et des relations dans les données, d'effectuer des prédictions ou de prendre des décisions basées sur ces données, et de s'adapter au fil du temps en fonction de nouvelles données et expériences.

Les réseaux de neurones artificiels, inspirés du fonctionnement du système nerveux humain, permettent aux machines d'apprendre et de s'adapter en fonction des données d'entrée (McCulloch & Pitts, 1943). (LeCun et al., 2015) soulignent que ces réseaux sont au cœur de l'apprentissage profond, une méthode d'apprentissage automatique qui a révolutionné plusieurs domaines, tels que la vision par ordinateur, la reconnaissance vocale et le traitement automatique du langage naturel.

L'intégration et le développement de ces techniques inspirées de la nature et du comportement humain permettent à l'IA de continuer à étendre ses capacités et à repousser les limites de ce que les machines peuvent réaliser. (Nilsson, 2009) définit le but de l'intelligence artificielle comme permettant aux machines d'effectuer des tâches normalement réservées aux humains : apprendre, parler, raisonner, s'adapter à des situations complexes et réfléchir. (McCarthy et al., 1956), l'un des pionniers de l'IA, la considère comme une discipline de l'informatique visant à construire des programmes "intelligents".

Les défis de l'IA sont considérables et ses applications nombreuses. (Thrun et al., 2005) explorent son utilisation dans les robots et les véhicules autonomes, tandis que (Sheridan, 2016) examine son rôle dans la cobotique et les assistants personnels. (Provost & Fawcett, 2013) mettent en lumière son application dans l'analyse de données comportementales à des fins de marketing. (Kahneman, 2011) souligne son potentiel dans la prise de décision complexe basée sur des situations nouvelles et complexes. (Dounis & Caraiscos, 2009) étudient son utilisation dans l'optimisation de la gestion technique des bâtiments, tandis que (Hastie et al., 2009) se concentrent sur son application dans l'analyse massive de données et la production de prévisions.

L'utilisation de ces techniques permet de détecter des tendances, des relations et des corrélations dans les données, de prédire et planifier des événements futurs, et de mieux comprendre certains phénomènes (Chakrabarti et al., 2006); (Fayyad et al., 1996). (Mayer-Schönberger & Cukier, 2013) soulignent que grâce à l'analyse des Big Data, les industries peuvent bénéficier de nouveaux outils tels que des machines auto-apprenantes qui analysent les données de l'atelier de production ou d'un poste de travail, acquérant ainsi un savoir-faire mécanique quasi instantané. (Hilbert, 2016) met en évidence comment ces analyses et outils mécaniques favorisent la standardisation, la modularisation et la personnalisation de masse.

3.2.2 Apprentissage automatique

3.2.2.1 Définitions et fondements théoriques

L'apprentissage automatique est un domaine de l'intelligence artificielle qui se concentre sur le développement d'algorithmes capables d'apprendre et de s'améliorer à partir de données, sans être explicitement programmés pour chaque tâche spécifique. Cette approche s'inspire du processus d'apprentissage humain, où l'expertise s'accroît avec l'expérience (Alpaydin, 2020).

(Samuel, 1959) a été l'un des pionniers à définir l'apprentissage automatique comme un « domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés ». Cette définition initiale a été affinée au fil des années, notamment par (T. M. Mitchell, 1997) qui propose une définition plus formelle et largement acceptée :

« Un programme informatique apprend de l'expérience E par rapport à une classe de tâches T et une mesure de performance P, si sa performance pour les tâches de T, mesurée par P, s'améliore avec l'expérience E. »

Cette définition met en évidence trois composantes essentielles de l'apprentissage automatique :

1. La tâche (T) que le système doit accomplir
2. La mesure de performance (P) qui évalue la qualité de l'exécution de la tâche
3. L'expérience (E) à partir de laquelle le système apprend

Il est crucial de noter que dans le contexte de l'apprentissage automatique, le terme "expérience" fait référence aux données historiques disponibles pour la machine. Tout comme un expert humain s'améliore en accumulant de l'expérience au fil du temps, un algorithme d'apprentissage automatique s'améliore en traitant et en apprenant à partir d'un volume croissant de données historiques (Goodfellow et al., 2016).

L'apprentissage automatique peut être considéré comme une convergence entre les approches mathématiques, statistiques et informatiques. Il vise à créer des systèmes capables d'effectuer des tâches complexes difficiles à définir par des règles explicites, en apprenant à partir de données d'exemple, de manière analogue au processus d'apprentissage humain (Jordan & Mitchell, 2015).

L'apprentissage automatique repose sur le principe de l'amélioration des performances d'un système à travers l'expérience (T. M. Mitchell, 1997). Contrairement à la programmation traditionnelle où les règles sont explicitement codées, les systèmes d'apprentissage automatique sont conçus pour apprendre ces règles à partir des données (Domingos, 2012).

Le processus d'apprentissage peut être conceptualisé comme suit :

1. Entrées : les données d'entrée, généralement sous forme de vecteurs de caractéristiques, représentant les attributs pertinents du problème à résoudre.
2. Modèle : une représentation mathématique paramétrable qui transforme les entrées en sorties. Ce modèle peut prendre diverses formes, telles que des arbres de décision, des réseaux de neurones, ou des fonctions de régression (Hastie et al., 2009).
3. Fonction objectif : une mesure quantitative de la performance du modèle, souvent appelée fonction de coût ou de perte. Elle évalue l'écart entre les prédictions du modèle et les résultats attendus (Christopher M. Bishop, 2006).
4. Algorithme d'optimisation : une méthode pour ajuster les paramètres du modèle afin de minimiser la fonction objectif. La descente de gradient est un exemple courant d'algorithme d'optimisation (Ruder, 2017).
5. Sorties : les prédictions ou décisions générées par le modèle en réponse aux données d'entrée.

Le processus d'apprentissage consiste à itérer sur un ensemble de données d'entraînement, en ajustant progressivement les paramètres du modèle pour améliorer ses performances selon la fonction objectif choisie (Shalev-Shwartz & Ben-David, 2014).

3.2.2.2 Applications et avancées significatives

L'apprentissage automatique a connu des avancées remarquables ces dernières décennies, avec des applications dans de nombreux domaines industriels et scientifiques. Voici quelques exemples notables qui illustrent ces progrès :

1. Reconnaissance de caractères manuscrits : la base de données MNIST, introduite par (Lecun et al., 1998), est devenue un benchmark standard pour évaluer les algorithmes de reconnaissance de chiffres manuscrits. Les performances sur ce jeu de données se sont considérablement améliorées, passant d'un taux d'erreur d'environ 5% dans les travaux originaux à 0.21% avec les techniques les plus avancées (Ciregan et al., 2012). Cette amélioration significative démontre la capacité croissante des algorithmes d'apprentissage automatique à traiter des tâches de reconnaissance visuelle complexes.
2. Systèmes de recommandation : le concours Netflix Prize (Bennett & Lanning, 2007) a stimulé la recherche dans le domaine des systèmes de recommandation. L'équipe gagnante, BellKor's Pragmatic Chaos, a utilisé des techniques d'ensemble averaging, combinant plusieurs algorithmes pour améliorer les performances de prédiction (Koren et al., 2009). Ces avancées ont eu un impact significatif sur les systèmes de recommandation utilisés dans le commerce électronique, les plateformes de streaming et d'autres services personnalisés.

3. Applications médicales : l'apprentissage automatique joue un rôle croissant dans le domaine médical. (Esteva et al., 2017) ont démontré que les réseaux de neurones convolutifs peuvent atteindre des performances au niveau des dermatologues pour la classification des cancers de la peau. Dans le contexte des maladies neurodégénératives, (Myszczynska et al., 2020) ont montré le potentiel de l'apprentissage automatique pour le diagnostic précoce et l'interprétation d'images médicales. De plus, (Vaishya et al., 2020) ont mis en évidence son utilité pour prédire le comportement des patients et identifier précocement les infections, notamment dans le contexte de la pandémie de COVID-19.
4. Traitement du langage naturel : les techniques d'apprentissage automatique ont révolutionné l'analyse des sentiments et le traitement du langage naturel. (Devlin et al., 2019) ont introduit BERT (Bidirectional Encoder Representations from Transformers), un modèle qui a établi de nouveaux standards de performance dans diverses tâches de traitement du langage naturel. Des travaux comme ceux de (Barzenji, 2021) ont démontré l'efficacité de ces approches pour analyser les opinions et les sentiments exprimés dans les textes, avec des applications dans l'analyse de marché et l'évaluation de l'impact des campagnes publicitaires.
5. Filtrage de contenu : l'apprentissage automatique est largement utilisé pour le filtrage des courriers indésirables. Des études comme celles de (Subramaniam et al., 2010) et (Dada et al., 2019) ont montré l'efficacité de ces techniques pour identifier et bloquer les spams, s'adaptant continuellement aux nouvelles stratégies des spammeurs. Plus récemment, (Bhowmick & Hazarika, 2018) ont proposé des approches basées sur l'apprentissage profond pour améliorer encore la précision du filtrage des spams.

Ces exemples illustrent la diversité et l'impact de l'apprentissage automatique dans des domaines variés, de la reconnaissance de formes à l'analyse de données complexes en passant par la prise de décision automatisée. Ils soulignent également l'importance croissante de ces techniques pour l'industrie 4.0 et la transformation numérique des entreprises.

3.2.2.3 Taxonomie des approches d'apprentissage automatique

L'apprentissage automatique englobe une variété d'approches qui peuvent être catégorisées selon différents critères. Deux classifications principales sont couramment utilisées dans la littérature : la catégorisation basée sur le modèle d'apprentissage et celle basée sur le type de tâche à accomplir (Mohri et al., 2018); (Hastie et al., 2009).

3.2.2.3.1 Classification basée sur le modèle d'apprentissage

Cette classification, largement adoptée dans la communauté scientifique, distingue quatre catégories principales d'apprentissage automatique en fonction de la nature des données d'entraînement et de la manière dont l'algorithme interagit avec ces données :

1. Apprentissage supervisé

   Dans ce contexte, l'algorithme apprend à partir de paires d'entrées-sorties étiquetées pour établir une relation entre elles (Hastie et al., 2009).

Exemple : La détection de spam dans les e-mails illustre bien l'apprentissage supervisé. L'algorithme est entraîné sur un ensemble de données contenant des e-mails étiquetés comme

"spam" ou "non spam". Le modèle apprend à reconnaître les caractéristiques des e-mails indésirables et peut ensuite être utilisé pour classer de nouveaux e-mails (Dada et al., 2019).

2. Apprentissage non supervisé

   Dans ce cas, lors d'un apprentissage non supervisé, l'algorithme traite des données d'entrée non étiquetées et cherche à identifier des structures ou des motifs cachés dans les données (Ghahramani, 2004)

Exemple : La segmentation de la clientèle pour le marketing utilise souvent l'apprentissage non supervisé. Les algorithmes de clustering peuvent identifier des groupes de clients similaires en fonction de leurs comportements d'achat, préférences ou données démographiques, permettant ainsi de personnaliser les offres marketing (R. Xu & Wunsch, 2005).

3. Apprentissage semi-supervisé

   Ce type d'algorithme combine les éléments des apprentissages supervisé et non supervisé, particulièrement utile lorsque l'étiquetage est coûteux ou chronophage (Zhu, 2005).

Exemple : Dans la classification de documents, un ensemble de documents étiquetés et non étiquetés est utilisé. L'algorithme apprend d'abord des caractéristiques à partir des documents étiquetés, puis applique ces connaissances pour étiqueter les documents non étiquetés (Chapelle et al., 2010).

4. Apprentissage par renforcement

   Dans cette catégorie, l'algorithme interagit avec un environnement dynamique et apprend à prendre des décisions en fonction des récompenses ou des pénalités reçues (R. S. Sutton & Barto, 2018).

Exemple : La création d'agents intelligents pour les jeux vidéo utilise souvent l'apprentissage par renforcement. Les agents apprennent à naviguer et à interagir avec l'environnement du jeu en essayant différentes actions et en recevant des récompenses ou des pénalités (Mnih et al., 2015).

3.2.2.3.2 Classification basée sur le type de tâche

Cette taxonomie catégorise les algorithmes d'apprentissage automatique selon le type de problème qu'ils visent à résoudre. Les principales catégories sont :

1. Classification

   L'algorithme apprend à attribuer des étiquettes prédéfinies aux données d'entrée. Plusieurs sous-types se distinguent (Christopher M. Bishop, 2006); (Alpaydin, 2020) :

   a. Classification binaire : choix entre deux groupes disjoints.
   b. Classification multi-classes : plus de deux groupes disjoints.
   c. Classification multi-labels : possibilité d'attribuer plusieurs labels par entrée.

2. Régression

   L'algorithme estime les relations entre les variables d'entrée et de sortie pour prédire des valeurs continues (Hastie et al., 2009); (James et al., 2021).

3. Clustering

   L'algorithme cherche à regrouper les données en clusters basés sur leur similarité, sans étiquettes prédéfinies. Le nombre de clusters peut être un paramètre ou déterminé par l'algorithme (Jain, 2010); (Aggarwal & Reddy, 2013).

Il est important de noter que ces catégories ne sont pas mutuellement exclusives et que de nombreux problèmes du monde réel peuvent nécessiter une combinaison de ces approches. De plus, avec l'évolution rapide du domaine, de nouvelles méthodes et paradigmes émergent continuellement, élargissant et redéfinissant parfois ces catégories traditionnelles (LeCun et al., 2015); (Goodfellow et al., 2016).

Cette taxonomie fournit un cadre conceptuel pour comprendre et organiser les différentes méthodes d'apprentissage automatique, permettant aux experts de choisir l'approche la plus appropriée en fonction de la nature des données disponibles et des objectifs spécifiques de leur problème.

3.2.2.4 Préparation des données

La résolution des problèmes d'apprentissage automatique nécessite généralement une grande quantité de données. Ces données, souvent issues de sources massives et hétérogènes caractéristiques du Big Data, sont fréquemment bruitées, erronées ou incomplètes. Une phase de prétraitement est donc cruciale pour assurer la qualité et la pertinence des entrées fournies aux algorithmes d'apprentissage automatique (García et al., 2015); (S. B. Kotsiantis et al., 2006).

Le prétraitement des données vise à transformer les données brutes en un format adapté à l'entrée des algorithmes. Cette phase comprend plusieurs étapes importantes, qui peuvent être visualisées comme un processus séquentiel :

1. Nettoyage des données : cette étape implique l'identification et la correction des erreurs, des incohérences et des valeurs manquantes dans les données. Les techniques courantes incluent l'imputation des valeurs manquantes, la détection et la correction des valeurs aberrantes, et l'élimination des doublons (Rahm & Do, 2000). Par exemple, (Little & Rubin, 2019) proposent des méthodes statistiques avancées pour l'imputation des données manquantes.
2. Intégration des données : lorsque les données proviennent de sources multiples, il est nécessaire de les fusionner en un ensemble cohérent. Cette étape implique la consolidation des données, l'harmonisation des formats et la résolution des conflits. Elle peut inclure l'application de filtres pour éliminer les données redondantes, le calcul d'agrégats pour synthétiser l'information, et la gestion des mises à jour en temps réel des données provenant de sources continues (Lenzerini, 2002); (Doan et al., 2012). (Jahani et al., 2023) présentent des techniques avancées d'intégration de données hétérogènes dans le contexte du Big Data.
3. Transformation des données : les données brutes peuvent nécessiter des transformations pour être adaptées à l'entrée des algorithmes d'apprentissage automatique. Les techniques courantes incluent la normalisation, la standardisation, la discrétisation des variables continues et l'encodage des variables catégorielles (S. B. Kotsiantis et al., 2006).

(Q. Zhang et al., 2018) proposent des méthodes innovantes de transformation de données pour l'apprentissage profond.

4. Réduction de la dimensionnalité : les ensembles de données de grande dimension peuvent entraîner une complexité accrue et une diminution des performances des modèles d'apprentissage automatique, un phénomène connu sous le nom de "malédiction de la dimensionnalité" (Bellman, 1961). La réduction de la dimensionnalité vise à réduire le nombre de variables tout en préservant l'information pertinente. Les techniques courantes incluent l'analyse en composantes principales (PCA) et la sélection de caractéristiques (Guyon & Elisseeff, 2003). (Maaten & Hinton, 2008) ont, par exemple, introduit t-SNE, une technique de réduction de dimensionnalité particulièrement efficace pour la visualisation de données de haute dimension.
5. Équilibrage des classes : dans de nombreux problèmes de classification, les classes peuvent être déséquilibrées, ce qui peut biaiser l'apprentissage du modèle vers la classe majoritaire. Les techniques d'équilibrage des classes, telles que le sur-échantillonnage de la classe minoritaire ou le sous-échantillonnage de la classe majoritaire, peuvent être utilisées pour atténuer ce problème (Chawla et al., 2002). (Haibo He & Garcia, 2009) fournissent une revue complète des techniques d'apprentissage à partir de données déséquilibrées.

La sélection des caractéristiques (feature selection) est une étape particulièrement importante dans le prétraitement des données, visant à réduire la dimensionnalité et à améliorer la performance des modèles. Selon (Guyon & Elisseeff, 2003), les méthodes de sélection des caractéristiques peuvent être classées en trois catégories principales :

1. Méthodes de filtrage : ces méthodes évaluent la pertinence de chaque caractéristique individuellement par rapport aux données d'apprentissage, indépendamment de l'algorithme d'apprentissage choisi. Bien que rapides, elles ne prennent pas en compte les interactions potentielles entre les caractéristiques (Saeys et al., 2007). (J. Li et al., 2017) proposent des méthodes de filtrage avancées basées sur la théorie de l'information.
2. Méthodes d'enveloppement (wrapper) : ces méthodes utilisent un modèle prédictif pour évaluer des sous-ensembles de caractéristiques. Bien qu'elles puissent capturer les interactions entre les caractéristiques, elles sont généralement plus coûteuses en termes de calcul (Kohavi & John, 1997). (Maldonado & Weber, 2009) présentent une approche d'enveloppement basée sur les machines à vecteurs de support (SVM) pour la sélection de caractéristiques.
3. Méthodes intégrées (embedded) : ces méthodes intègrent la sélection des caractéristiques dans le processus d'apprentissage lui-même, offrant un compromis entre la rapidité des méthodes de filtrage et la prise en compte des interactions des méthodes d'enveloppement (Lal et al., 2006). (J. Tang et al., 2014) proposent une méthode intégrée basée sur l'apprentissage profond pour la sélection de caractéristiques.

Il est important de noter que le prétraitement des données est souvent un processus itératif et peut nécessiter une expertise du domaine pour être effectué efficacement. De plus, certaines entreprises explorent des approches innovantes pour améliorer la qualité de leurs données,

comme l'utilisation de plateformes de crowdsourcing telles qu'Amazon Mechanical Turk pour des tâches de validation de données ou d'étiquetage (Buhrmester et al., 2011). (Vaughan, 2018) fournit une analyse approfondie des avantages et des défis liés à l'utilisation du crowdsourcing pour la préparation des données en apprentissage automatique.

En somme, la préparation des données est une étape fondamentale dans le processus d'apprentissage automatique, ayant un impact significatif sur la performance et la fiabilité des modèles résultants. Une attention particulière à cette phase peut grandement améliorer la qualité des résultats obtenus par les algorithmes d'apprentissage automatique. Comme le soulignent (Domingos, 2012) et (NG, 2018), la qualité des données et leur préparation adéquate sont souvent plus importantes que le choix de l'algorithme d'apprentissage lui-même pour obtenir de bonnes performances.

3.2.2.5 Algorithmes d'optimisation

Les algorithmes d'optimisation jouent un rôle crucial dans le processus d'apprentissage automatique, visant à minimiser l'écart entre les sorties prédites par le modèle et les sorties désirées. Cette minimisation est réalisée en optimisant une fonction objectif, également appelée fonction de coût ou fonction de perte, qui quantifie cet écart (Goodfellow et al., 2016). Le choix de cette fonction dépend de la nature du problème et du type d'apprentissage utilisé.

3.2.2.5.1 Descente de gradient

L'algorithme de descente de gradient, introduit initialement dans le contexte de l'apprentissage automatique avec le modèle ADALINE (Widrow & Hoff, 1988), est une méthode fondamentale d'optimisation itérative (Figure 14). Son principe repose sur l'ajustement progressif des paramètres du modèle dans la direction opposée au gradient de la fonction de coût, permettant ainsi de converger vers un minimum local de cette fonction (Ruder, 2017).

La descente de gradient est particulièrement utile pour minimiser la fonction de coût en apprentissage automatique. Son principe fondamental repose sur l'idée de "descendre" progressivement vers le point le plus bas de la fonction de coût.

Le fonctionnement général peut être résumé ainsi :

1. Calcul du gradient : la direction de la plus forte augmentation de la fonction de coût est déterminée.
2. Déplacement opposé au gradient : l'algorithme ajuste les paramètres dans la direction opposée au gradient, avec une amplitude déterminée par le taux d'apprentissage.
3. Itération : ce processus est répété jusqu'à ce que le gradient devienne suffisamment petit, indiquant la proximité d'un minimum.

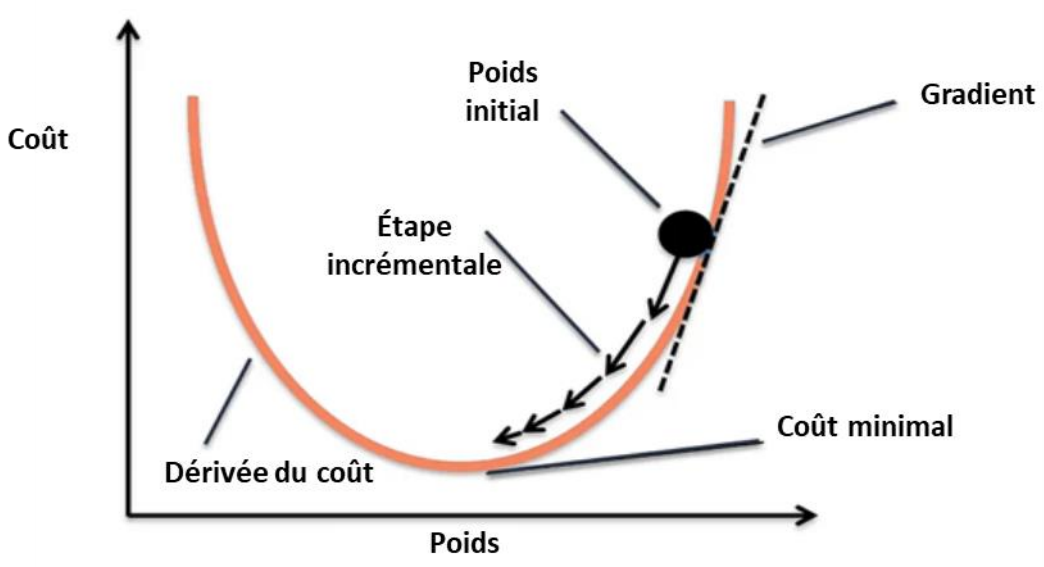


Figure 14 Illustration du processus de descente de gradient (Kapil, 2019)

Formellement, l'algorithme de descente de gradient peut être décrit comme suit :

1. Initialisation : choisir une valeur initiale w pour les paramètres du modèle.
2. Itération :
    a. Calculer la valeur de la fonction de coût $L(w)$
    b. Calculer le gradient $\frac{\partial L}{\partial w}$
    c. Mettre à jour les paramètres $w \leftarrow w - \eta \frac{\partial L}{\partial w}$ où η est le taux d'apprentissage.
3. Répéter l'étape 2 jusqu'à ce que $\left|\left|\frac{\partial L}{\partial w}\right|\right| < \varepsilon$ où ε est un seuil de tolérance prédéfini assez petit.

Sur la Figure 14, le point noir représente le poids initial (1), et les flèches montrent les étapes incrémentales (2) vers le minimum de la fonction de coût (3).

Le taux d'apprentissage η est un hyperparamètre crucial qui influence la convergence de l'algorithme. Un taux trop élevé peut entraîner des oscillations ou une divergence, tandis qu'un taux trop faible peut ralentir considérablement la convergence (L. Bottou et al., 2018). (Sutskever et al., 2013) ont démontré l'importance du choix judicieux du taux d'apprentissage pour la performance des modèles d'apprentissage profond.

3.2.2.5.2 Variantes et améliorations

Plusieurs variantes de la descente de gradient ont été développées pour améliorer sa performance et sa stabilité :

1. Descente de gradient stochastique (SGD) : cette variante utilise un sous-ensemble aléatoire des données à chaque itération, permettant une convergence plus rapide et une meilleure généralisation (L. Bottou, 2010). (Robbins & Monro, 1951) ont posé les bases théoriques de cette approche stochastique.

2. Descente de gradient par mini-lots : cette approche combine les avantages de la descente de gradient classique et de la SGD en utilisant des sous-ensembles de données de taille fixe (Goodfellow et al., 2016). (M. Li et al., 2014) ont analysé l'impact de la taille des mini-lots sur la convergence et la généralisation des modèles.
3. Optimiseurs adaptatifs : des algorithmes tels qu'Adam (Kingma & Ba, 2017), RMSprop (Tieleman & Hinton, 2012) et Adagrad (Duchi et al., 2011) adaptent dynamiquement le taux d'apprentissage pour chaque paramètre, améliorant ainsi la convergence et la stabilité de l'apprentissage. (Reddi et al., 2019) ont fourni une analyse approfondie des propriétés de convergence de ces optimiseurs adaptatifs.

Ces méthodes d'optimisation avancées ont contribué de manière significative à l'amélioration des performances des modèles d'apprentissage automatique, en particulier dans le contexte des réseaux de neurones profonds (Ruder, 2017). (Z. Zhang et al., 2022) ont proposé une étude comparative détaillée de ces différents optimiseurs dans le contexte de l'apprentissage profond.

3.2.2.5.3 Hyperparamètres et réglage

Les hyperparamètres, tels que le taux d'apprentissage, la taille des mini-lots ou le nombre d'époques d'entraînement, sont des paramètres définis avant l'apprentissage et non appris par le modèle lui-même. Leur optimisation est cruciale pour obtenir des performances optimales et peut être réalisée par des techniques telles que la recherche par grille, la recherche aléatoire ou l'optimisation bayésienne (Bergstra & Bengio, 2012). (Snoek et al., 2012) ont démontré l'efficacité de l'optimisation bayésienne pour le réglage des hyperparamètres dans les modèles d'apprentissage profond.

(Feurer & Hutter, 2019) ont fourni une revue exhaustive des techniques d'optimisation des hyperparamètres, soulignant l'importance de cette étape dans le processus d'apprentissage automatique. Ils ont notamment mis en évidence l'émergence de techniques d'optimisation automatique des hyperparamètres, ouvrant la voie à des systèmes d'apprentissage automatique plus autonomes.

Les algorithmes d'optimisation jouent un rôle central dans l'apprentissage automatique, permettant d'ajuster efficacement les paramètres des modèles pour minimiser l'erreur de prédiction. Le choix et le réglage appropriés de ces algorithmes et de leurs hyperparamètres sont essentiels pour obtenir des modèles performants et généralisables. Les avancées récentes dans ce domaine, telles que les optimiseurs adaptatifs et les techniques d'optimisation automatique des hyperparamètres, continuent d'améliorer la performance et l'efficacité des modèles d'apprentissage automatique, ouvrant de nouvelles perspectives pour leur application dans des domaines complexes.

3.2.2.6 Méthodes d'entraînement

L'entraînement des modèles d'apprentissage automatique peut être réalisé selon différentes approches, chacune présentant des avantages et des inconvénients spécifiques. Les trois principales méthodes d'entraînement sont l'entraînement par batch, l'entraînement par mini-batch et l'entraînement en ligne (ou stochastique) (Goodfellow et al., 2016). Le choix de la méthode d'entraînement appropriée peut avoir un impact significatif sur la performance et l'efficacité du modèle.

3.2.2.6.1 Entraînement par lot

L'entraînement par lot (batch), également appelé apprentissage hors ligne, utilise l'intégralité du jeu de données pour mettre à jour les paramètres du modèle en une seule itération (L. Bottou et al., 2018). Cette approche présente l'avantage de fournir une estimation précise du gradient de la fonction de coût, car elle prend en compte l'ensemble des données d'entraînement. Cependant, elle peut s'avérer coûteuse en termes de ressources computationnelles et de mémoire, particulièrement pour de grands jeux de données (LeCun et al., 2015).

Formellement, la mise à jour des paramètres θ à chaque itération t peut être exprimée par dans l'Équation 1.

Équation 1 Mise à jour des paramètres dans l'entraînement par lot

$$\theta(t+1) = \theta(t) - \eta \nabla L(\theta(t))$$

où η est le taux d'apprentissage et $\nabla L(\theta(t))$ est le gradient de la fonction de coût L calculé sur l'ensemble des données d'entraînement

Selon (L. Bottou et al., 2018), l'entraînement par batch offre une convergence stable et une estimation précise du gradient, mais peut être inefficace pour de très grands ensembles de données. De plus, comme le soulignent (Keskar et al., 2017), l'utilisation de grands batches peut conduire à une généralisation plus faible du modèle.

3.2.2.6.2 Entraînement en ligne (stochastique)

L'entraînement en ligne, également connu sous le nom d'apprentissage stochastique, met à jour les paramètres du modèle après le traitement de chaque exemple individuel du jeu de données (L. Bottou, 2010). Cette approche présente l'avantage d'être rapide et de nécessiter peu de mémoire, car seul un exemple est traité à la fois. De plus, elle peut aider à échapper aux minima locaux grâce aux fluctuations du gradient (L. E. Bottou, 1998). Cependant, les mises à jour des paramètres peuvent être bruyantes et moins stables en raison de l'estimation du gradient basée sur un seul exemple.

La mise à jour des paramètres pour un exemple x_i à l'itération i s'exprime par l'Équation 2.

Équation 2 Mise à jour des paramètres dans l'entraînement en ligne (stochastique)

$$\theta(t+1) = \theta(t) - \eta \nabla L_i(\theta(t))$$

où $\nabla L_i(\theta(t))$ est le gradient de la fonction de coût calculé sur l'exemple x_i

(L. Bottou & Bousquet, 2007) ont démontré que l'apprentissage stochastique peut atteindre une meilleure généralisation que l'apprentissage par batch dans certaines conditions, en particulier pour les grands ensembles de données.

3.2.2.6.3 Entraînement par mini-lot

L'entraînement par mini-lot (« mini-batch ») représente un compromis entre l'entraînement par lot et l'entraînement en ligne. Dans cette approche, le jeu de données est divisé en sous-ensembles (mini-batches) de taille fixe, et les paramètres du modèle sont mis à jour après le traitement de chaque mini-batch (Ruder, 2017). Cette méthode permet de réduire la variance des mises à jour des paramètres par rapport à l'entraînement en ligne, tout en offrant une convergence plus rapide que l'entraînement par batch (M. Li et al., 2014).

La mise à jour des paramètres pour un mini-batch B à l'itération t peut être exprimée par l'Équation 3.

Équation 3 Mise à jour des paramètres dans l'entraînement par mini-batch

$$\theta(t+1) = \theta(t) - \eta \nabla L_B(\theta(t))$$

où $\nabla L_B(\theta(t))$ est le gradient de la fonction de coût calculé sur le mini-batch B

(Masters & Luschi, 2018) ont mené une étude approfondie sur l'impact de la taille des mini-batches sur les performances des modèles d'apprentissage profond. Ils ont constaté que des mini-batches de taille relativement petite (32-512) conduisent généralement à de meilleures performances et à une meilleure généralisation.

3.2.2.6.4 Fonctions coût

Le choix de la fonction coût dépend de la nature du problème à résoudre. Parmi les fonctions coût couramment utilisées, il convient de citer :

1. L'erreur quadratique moyenne (MSE) : utilisée principalement pour les problèmes de régression, elle mesure la moyenne des carrés des différences entre les valeurs prédites et les valeurs réelles (Chai & Draxler, 2014). La formule de la MSE est donnée par l'Équation 4.

Équation 4 Formule de l'erreur quadratique moyenne

$$MSE = \frac{1}{n}\sum (y_i - \hat{y}_i)^2$$

où y_i sont les valeurs réelles et $\hat{y}_i$ les valeurs prédites

2. L'erreur absolue moyenne (MAE) : également utilisée pour la régression, elle mesure la moyenne des valeurs absolues des différences entre les prédictions et les valeurs réelles (Willmott & Matsuura, 2005). La MAE est exprimée par l'Équation 5.

Équation 5 Formule de l'erreur absolue moyenne

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i|$$

où y_i sont les valeurs réelles et $\hat{y}_i$ les valeurs prédites

3. L'entropie croisée : couramment employée pour les problèmes de classification, elle mesure la divergence entre la distribution de probabilité prédite et la distribution réelle (de Boer et al., 2005). La formule de l'entropie croisée est donnée par l'Équation 6.

Équation 6 Formule de l'entropie croisée

$$\mathrm{H(p,q)}=-\sum p(x)\log q(x)$$

où $p(x)$ est la distribution réelle et $q(x)$ la distribution prédite

(Janocha & Czarnecki, 2017) ont mené une étude comparative approfondie sur différentes fonctions de perte pour la classification, mettant en évidence les avantages et les inconvénients de chacune dans divers scénarios.

Le choix de la méthode d'entraînement et de la fonction de coût appropriée est crucial pour optimiser les performances du modèle d'apprentissage automatique et dépend souvent des caractéristiques spécifiques du problème à résoudre et des ressources computationnelles disponibles. Comme le souligne (L. N. Smith, 2017), l'adaptation dynamique des hyperparamètres

d'entraînement, tels que le taux d'apprentissage et la taille des mini-batches, peut conduire à des améliorations significatives des performances et de la vitesse de convergence.

3.2.2.7 Évaluation des modèles

Bien que la fonction d'erreur permette de quantifier l'écart entre les prédictions et les valeurs réelles, l'évaluation globale de la validité d'un modèle nécessite des métriques plus complexes. L'exactitude (accuracy), définie comme le rapport entre le nombre de prédictions correctes et le nombre total de prédictions, est souvent considérée comme une mesure intuitive de performance. Elle se calcule selon l'Équation 7.

Équation 7 Formule naive de l'exactitude

$$exactitude = \frac{nombre\ de\ prédictions\ correctes}{nombre\ total\ de\ prédiction}$$

Cependant, cette métrique présente des limitations significatives, particulièrement dans le contexte de données déséquilibrées (class-imbalanced data). Comme le soulignent (Haibo He & Garcia, 2009), dans un scénario où l'événement d'intérêt est rare, représentant moins de 1% des cas, un modèle naïf qui prédirait systématiquement la classe majoritaire atteindrait une exactitude supérieure à 99%, malgré son absence totale de valeur prédictive pour la classe minoritaire.

Cette problématique souligne la nécessité d'utiliser des métriques d'évaluation plus robustes et informatives. La matrice de confusion, un outil fondamental en apprentissage automatique, offre une base pour dériver des métriques plus nuancées (Fawcett, 2006). Elle se présente sous la forme du Tableau 4.

Tableau 4 Matrice de confusion

	Réalité : oui	Réalité : non
Prédiction : oui	Vrai Positif (VP, true positive)	Faux Positif (FP, false positive)
Prédiction : non	Faux Négatif (FN, false negative)	Vrai Négatif (VN, true negative)

Cette représentation permet une analyse plus détaillée des performances du modèle, en distinguant les différents types d'erreurs et de prédictions correctes. À partir de cette matrice, il est possible de dériver des métriques plus sophistiquées, offrant une évaluation plus complète et nuancée de la performance du modèle.

1. Exactitude (Accuracy) :

L'exactitude, définie par l'Équation 8, représente la proportion globale de prédictions correctes.

Équation 8 Formule de l'exactitude pour la classification binaire

$$exactitude = \frac{(VP + VN)}{(VP + VN + FP + FN)}$$

Bien qu'intuitive, cette métrique peut être trompeuse dans le cas de classes déséquilibrées, comme l'ont démontré (Haibo He & Garcia, 2009) dans leur étude approfondie sur l'apprentissage à partir de données déséquilibrées.

2. Précision (Precision) :

La précision, illustrée par l'Équation 9 mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives.

Équation 9 Formule de la précision pour la classification binaire

$$précision = \frac{VP}{(VP + FP)}$$

Cette métrique est particulièrement utile lorsque le coût des faux positifs est élevé. (Powers & Ailab, 2011) ont fourni une analyse détaillée de l'importance de la précision dans divers contextes d'application.

3. Rappel (Recall) :

Également appelé sensibilité ou taux de vrais positifs, le rappel, présenté dans l'Équation 10, quantifie la proportion d'instances positives correctement identifiées.

Équation 10 Formule du rappel pour la classification binaire

$$rappel = \frac{VP}{(VP + FN)}$$

Le rappel est crucial lorsque la détection des cas positifs est primordiale, comme dans certaines applications médicales. (Saito & Rehmsmeier, 2015) ont souligné l'importance du rappel dans leur étude comparative des courbes précision-rappel.

4. Score F1 :

Le score F1, donné par l'Équation 11, est la moyenne harmonique de la précision et du rappel, offrant un équilibre entre ces deux métriques :

Équation 11 Formule du score F1 pour la classification binaire

$$F1 = 2 \times \frac{(précision \times rappel)}{(précision + rappel)}$$

Cette métrique est particulièrement utile lorsqu'un équilibre entre précision et rappel est nécessaire. (Chicco & Jurman, 2020) ont fourni une analyse approfondie du score F1 et de ses variantes dans le contexte de l'évaluation des modèles de classification binaire.

Il est important de noter qu'il existe souvent un compromis entre précision et rappel, comme illustré par la Figure 15. Un modèle optimisé pour une haute précision peut avoir un faible rappel, et vice versa. Le choix de la métrique à privilégier dépend du contexte spécifique de l'application et des coûts associés aux différents types d'erreurs, comme l'a souligné (Fawcett, 2006) dans son étude fondamentale sur l'utilisation et l'interprétation des courbes ROC.

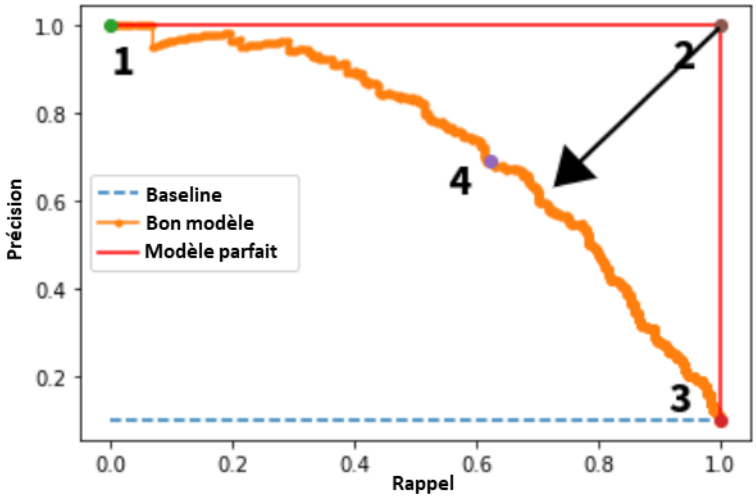


Figure 15 Exemple de courbe précision-rappel (Bonnet, 2023)

Ces métriques fournissent une base solide pour l'évaluation des modèles de classification, permettant une analyse nuancée de leurs performances. Elles seront utilisées dans cette thèse pour évaluer et comparer les performances des modèles développés, offrant ainsi une perspective complète sur leur efficacité.

De plus, il est important de considérer d'autres techniques d'évaluation complémentaires, telles que la validation croisée et les courbes ROC (Receiver Operating Characteristic). La validation croisée, comme l'ont décrit (Kohavi, 1995) et (Arlot & Celisse, 2010), permet d'estimer la capacité de généralisation du modèle en le testant sur différents sous-ensembles de données. Les courbes ROC, quant à elles, offrent une visualisation graphique de la performance du classificateur pour différents seuils de décision, comme l'ont expliqué en détail (Bradley, 1997) et (Hanley & McNeil, 1982).

Enfin, dans le contexte de l'apprentissage automatique appliqué à des domaines spécifiques, il est crucial de considérer des métriques d'évaluation adaptées au domaine. Par exemple, dans le domaine médical, (Obuchowski et al., 2004) ont souligné l'importance d'utiliser des métriques qui prennent en compte la gravité des erreurs de diagnostic. De même, dans le domaine de la finance, (López de Prado, 2018) a proposé des métriques spécifiques pour évaluer les modèles de trading algorithmique.

3.2.2.8 Synthèse

L'apprentissage automatique classique, tel que nous l'avons exploré jusqu'à présent, offre un ensemble puissant d'outils et de techniques pour extraire des informations et faire des prédictions à partir de données. Les méthodes que nous avons examinées, allant de la préparation des données aux algorithmes d'optimisation et aux techniques d'évaluation des modèles, ont démontré leur efficacité dans de nombreux domaines d'application.

Cependant, malgré ces avancées significatives, l'apprentissage automatique classique présente certaines limitations qui peuvent entraver son efficacité face à des problèmes complexes et des ensembles de données massifs et hautement dimensionnels. Comme l'ont souligné (LeCun et al., 2015), ces limitations incluent :

1. La difficulté à traiter des données brutes et non structurées : les méthodes classiques nécessitent souvent une ingénierie des caractéristiques manuelle et experte, ce qui peut être chronophage et sujet à des biais humains.
2. Le manque de capacité à capturer des relations complexes et non linéaires : les modèles traditionnels peuvent avoir du mal à saisir des motifs subtils et des interactions complexes entre les variables.
3. La difficulté à généraliser à partir d'un nombre limité d'exemples : les approches classiques peuvent nécessiter de grandes quantités de données étiquetées pour atteindre des performances satisfaisantes.
4. La sensibilité au fléau de la dimensionnalité : avec l'augmentation du nombre de caractéristiques, les modèles classiques peuvent devenir inefficaces ou sujets au sur-apprentissage.

Face à ces défis, et motivés par le besoin de traiter des problèmes toujours plus complexes dans des domaines tels que la vision par ordinateur, le traitement du langage naturel et la robotique, les chercheurs se sont tournés vers des approches plus avancées. C'est dans ce contexte que l'apprentissage profond, ou "deep learning", a émergé comme une extension puissante de l'apprentissage automatique classique.

L'apprentissage profond, comme l'ont décrit (Goodfellow et al., 2016), propose de surmonter ces limitations en utilisant des architectures de réseaux de neurones multicouches capables d'apprendre automatiquement des représentations hiérarchiques des données. Cette approche permet de traiter directement des données brutes, de capturer des relations complexes et non linéaires, et de généraliser efficacement à partir d'un nombre limité d'exemples.

Dans la section suivante, nous explorerons en détail les principes fondamentaux de l'apprentissage profond, ses architectures les plus courantes, et ses applications révolutionnaires dans divers domaines. Nous verrons comment cette approche repousse les frontières de l'intelligence artificielle et ouvre de nouvelles perspectives pour résoudre des problèmes complexes qui étaient auparavant hors de portée des méthodes d'apprentissage automatique classiques.

3.2.3 Apprentissage profond

3.2.3.1 Introduction à l'apprentissage profond

L'apprentissage profond, ou "deep learning" en anglais, représente une avancée majeure dans le domaine de l'apprentissage automatique. Cette approche se distingue par l'utilisation de réseaux de neurones artificiels complexes et multicouches, capables d'apprendre des représentations hiérarchiques des données (LeCun et al., 2015); (LeCun et al., 2015). S'inspirant du fonctionnement du cerveau humain, l'apprentissage profond traite l'information à travers de multiples niveaux d'abstraction, permettant ainsi d'aborder des problèmes complexes avec une efficacité remarquable.

L'essor de l'apprentissage profond est intimement lié à trois facteurs clés :

1. La disponibilité massive de données : l'avènement de l'ère du Big Data a fourni la matière première nécessaire à l'entraînement de modèles complexes. (Cisco, 2018) souligne que l'entrée dans l'ère du Zettabyte, où le trafic Internet global a dépassé un Zettabyte, a marqué un tournant décisif dans la disponibilité des données. De plus, la prolifération de

l'Internet des Objets (IoT) a considérablement élargi les sources de données exploitables (Yasmeen, 2018); (Gubbi et al., 2013).

2. L'augmentation de la puissance de calcul : les avancées en matière de hardware, notamment l'utilisation des processeurs graphiques (GPU), ont permis de surmonter les limitations computationnelles qui freinaient auparavant le développement de modèles complexes. (Oh & Jung, 2004) ainsi que (Raina et al., 2009) ont démontré que les GPU, avec leur architecture massivement parallèle, sont particulièrement adaptés aux opérations matricielles intensives caractéristiques de l'apprentissage profond. Des architectures comme CUDA de NVIDIA ont facilité l'exploitation efficace de cette puissance de calcul pour l'entraînement de réseaux de neurones profonds (Nickolls et al., 2008).
3. Les avancées algorithmiques : le développement de nouvelles architectures de réseaux, de techniques d'optimisation et de régularisation a permis d'améliorer significativement les performances et la stabilité des modèles d'apprentissage profond (Schmidhuber, 2015); (Bengio et al., 2013).

L'apprentissage profond a démontré son efficacité dans une variété de domaines, notamment :

- La vision par ordinateur : (Krizhevsky et al., 2012) et (He et al., 2016) ont réalisé des avancées significatives dans la reconnaissance d'objets, la détection de visages et la segmentation d'images.
- Le traitement du langage naturel : (Devlin et al., 2019) et (Vaswani et al., 2017) ont révolutionné la traduction automatique, l'analyse de sentiment et la génération de texte.
- La reconnaissance vocale : (H. Zhao et al., 2018) et (Graves et al., 2013) ont amélioré considérablement la conversion de la parole en texte et la synthèse vocale.
- La robotique : (Levine et al., 2016) et (Tai et al., 2017) ont fait progresser la perception de l'environnement et la planification de trajectoires.
- Les systèmes de recommandation : (S. Zhang et al., 2019) et (Covington et al., 2016) ont perfectionné la personnalisation de contenu et la prédiction de préférences utilisateur.

Une caractéristique fondamentale de l'apprentissage profond est sa capacité à apprendre automatiquement des représentations hiérarchiques des données (Bengio et al., 2013). Contrairement aux approches traditionnelles qui nécessitent une ingénierie manuelle des caractéristiques, les modèles d'apprentissage profond peuvent extraire automatiquement des caractéristiques pertinentes à partir des données brutes (LeCun et al., 2015). Cette capacité est particulièrement précieuse pour traiter des données complexes et non structurées, comme les images, le texte ou les signaux audio (Deng & Yu, 2014).

L'architecture typique d'un réseau de neurones profond comprend plusieurs couches de neurones artificiels interconnectés (Schmidhuber, 2015). Chaque couche apprend à représenter les données à un niveau d'abstraction différent. (Zeiler & Fergus, 2014) ont démontré que dans le cas de la reconnaissance d'images, les premières couches peuvent apprendre à détecter des contours simples, tandis que les couches suivantes combinent ces informations pour reconnaître des formes plus complexes, jusqu'à la reconnaissance d'objets entiers.

Malgré ses succès remarquables, l'apprentissage profond présente également des défis importants :

- Besoin en données : (C. Sun et al., 2017) soulignent que les modèles d'apprentissage profond nécessitent généralement de grandes quantités de données étiquetées pour atteindre des performances optimales, ce qui peut être coûteux et chronophage à obtenir dans certains domaines.
- Interprétabilité : (Lipton, 2018) et (Rudin, 2019) mettent en évidence la difficulté d'interpréter les décisions des modèles d'apprentissage profond, ce qui peut être problématique dans des applications critiques comme la médecine ou la finance.
- Généralisation : (C. Zhang et al., 2021) soulignent le défi d'assurer que les modèles généralisent bien à des données non vues pendant l'entraînement, en particulier lorsque les distributions des données de test diffèrent significativement de celles d'entraînement.
- Ressources computationnelles : (Strubell et al., 2019) mettent en lumière les importantes ressources computationnelles nécessaires pour l'entraînement de modèles d'apprentissage profond complexes, ce qui peut limiter leur accessibilité.

Malgré ces défis, l'apprentissage profond continue d'évoluer rapidement, avec des développements prometteurs dans des domaines tels que l'apprentissage par transfert (Pan & Yang, 2010), l'apprentissage par renforcement profond (Mnih et al., 2015), et les modèles génératifs (Goodfellow et al., 2014). Ces avancées ouvrent de nouvelles perspectives pour l'application de l'apprentissage profond dans des contextes industriels complexes, comme évoqués précédemment.

Dans les sections suivantes, nous explorerons plus en détail les architectures spécifiques de réseaux de neurones profonds et leur application potentielle à notre problématique de recherche.

3.2.3.2 Réseaux de neurones artificiels

Les réseaux de neurones artificiels (RNA) constituent le fondement des méthodes d'apprentissage profond. S'inspirant du fonctionnement du cerveau humain, ces réseaux sont composés de multiples couches de neurones interconnectés, capables de traiter et de transformer l'information de manière hiérarchique (Goodfellow et al., 2016). Un RNA est un ensemble interconnecté de neurones artificiels, chacun étant une unité de traitement élémentaire imitant le fonctionnement des neurones biologiques.

3.2.3.2.1 Structure d'un neurone artificiel

Un neurone artificiel, dont la structure est illustrée dans la Figure 16, se compose des éléments suivants :

1. Entrées : Un ensemble de valeurs d'entrée $x_1, x_2, \ldots, x_n$.
2. Poids synaptiques : Chaque entrée x_i est associée à un poids w_i.
3. Biais : Une valeur b, souvent représentée comme une entrée supplémentaire x_0 avec $w_0 = b$.
4. Fonction de sommation : Calcule la somme pondérée des entrées.
5. Fonction d'activation : Une fonction non-linéaire φ appliquée à la somme pondérée.
6. Sortie : La valeur y produite par le neurone.

Mathématiquement, la sortie y d'un neurone peut être exprimée par l'Équation 12 et schématisé par la Figure 16.

Équation 12 Formule générale de la sortie d'un neurone artificiel

$$y = \varphi\left(\sum_{i=0}^{n} w_i x_i\right)$$
$$w_0 = b$$

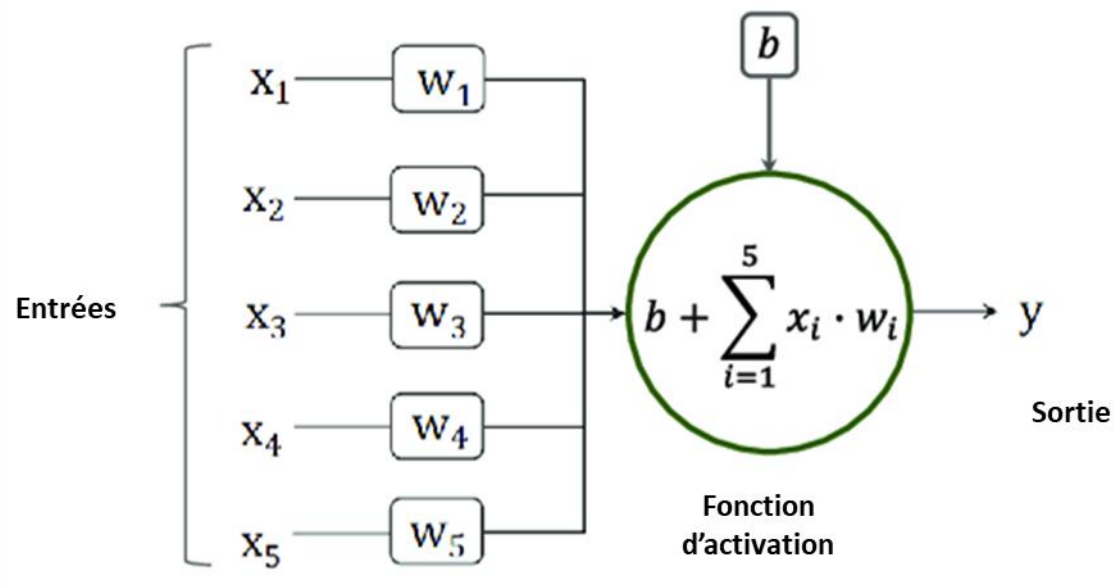


Figure 16 Schéma de principe d'un neurone artificiel (Lemus et al., 2021)

3.2.3.2.2 Le perceptron : un cas particulier

Le perceptron, introduit par (Rosenblatt, 1958), est un type spécifique de neurone artificiel utilisant la fonction signe comme fonction d'activation, calculée par l'Équation 13.

Équation 13 Fonction d'activation du perceptron (fonction signe)

$$\varphi(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

Le perceptron agit comme un classificateur binaire, divisant l'espace d'entrée en deux catégories. (Minsky & Papert, 1969) ont démontré que le perceptron est un modèle linéaire, capable de résoudre des problèmes de classification linéairement séparables, tels que les fonctions logiques AND et OR. Cependant, il est incapable de traiter des problèmes non linéairement séparables, comme la fonction XOR.

La principale limitation du perceptron simple réside dans son incapacité à modéliser des fonctions non linéaires. Cette restriction a constitué un obstacle majeur dans le développement initial des réseaux de neurones artificiels. Pour surmonter cette limitation, il a été nécessaire de développer des architectures plus complexes, comprenant plusieurs couches de neurones et des fonctions d'activation non linéaires.

Ces avancées ont conduit à l'émergence des réseaux de neurones multicouches et, par la suite, aux architectures d'apprentissage profond actuelles. Comme le soulignent (LeCun et al., 2015), ces architectures plus avancées permettent de modéliser des relations complexes et non linéaires dans les données, ouvrant la voie à une large gamme d'applications en apprentissage automatique et en intelligence artificielle. (Bengio et al., 2013) ont démontré que ces architectures profondes sont capables d'apprendre des représentations hiérarchiques des

données, ce qui les rend particulièrement efficaces pour traiter des problèmes complexes dans divers domaines tels que la vision par ordinateur, le traitement du langage naturel et la robotique.

3.2.3.3 Architectures des réseaux de neurones artificiels

Pour aborder des problèmes plus complexes que ceux traités par un simple perceptron, les réseaux de neurones artificiels (RNA) multicouches ont été développés. Ces architectures plus sophistiquées permettent de modéliser des relations non linéaires complexes dans les données, ouvrant ainsi la voie à une large gamme d'applications en apprentissage automatique et en intelligence artificielle (LeCun et al., 2015).

Un RNA typique est composé d'une succession de couches de neurones interconnectés, comme illustré dans la Figure 17. Cette architecture comprend généralement :

1. Une couche d'entrée (en bleu) : reçoit les données brutes ou les caractéristiques initiales.
2. Une ou plusieurs couches cachées (en jaune) : effectuent des transformations non linéaires successives des données.
3. Une couche de sortie (en vert) : produit la prédiction ou la décision finale du réseau.

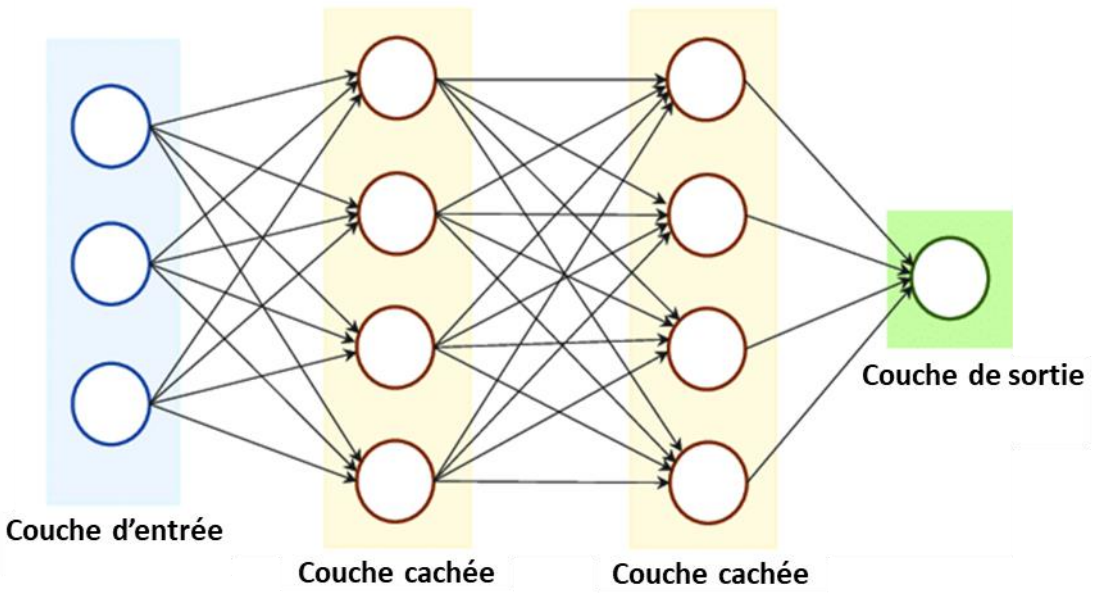


Figure 17 Schéma de principe d'un réseau de neurones artificiels (Lemus et al., 2021)

Cette structure en couches s'inspire de l'organisation hiérarchique observée dans le cortex cérébral, où l'information est traitée de manière séquentielle à travers différentes régions du cerveau (Felleman & Van Essen, 1991). (Bengio et al., 2013) ont démontré que ces architectures profondes sont capables d'apprendre des représentations hiérarchiques des données, ce qui les rend particulièrement efficaces pour traiter des problèmes complexes dans divers domaines tels que la vision par ordinateur, le traitement du langage naturel et la robotique.

3.2.3.3.1 Apprentissage hiérarchique

Le concept fondamental de l'apprentissage profond repose sur la notion de représentations hiérarchiques de l'information (Bengio et al., 2013). Chaque couche du réseau apprend à extraire

des caractéristiques de plus en plus abstraites et complexes des données d'entrée. Par exemple, dans le traitement d'images :

- Les premières couches peuvent détecter des contours simples.
- Les couches intermédiaires combinent ces contours pour identifier des formes plus complexes.
- Les couches profondes peuvent reconnaître des objets entiers ou des concepts abstraits.

(Zeiler & Fergus, 2014) ont visualisé ce processus d'apprentissage hiérarchique dans les réseaux convolutifs, démontrant comment les différentes couches apprennent progressivement des représentations de plus en plus complexes des images. Cette capacité d'apprentissage hiérarchique permet aux modèles d'apprentissage profond de traiter efficacement des données complexes et non structurées, surpassant souvent les approches traditionnelles de l'intelligence artificielle dans des tâches telles que la reconnaissance d'images ou le traitement du langage naturel (LeCun et al., 2015).

3.2.3.3.2 Paramètres et hyperparamètres

Les réseaux de neurones artificiels sont caractérisés par deux types de variables ajustables :

1. Paramètres : ce sont les variables internes du modèle, ajustées durant l'apprentissage. Ils comprennent :
    a. Les poids synaptiques (w_i)
    b. Les biais (b)
    c. Les paramètres des fonctions d'activation (φ)
2. Hyperparamètres : ce sont des variables externes au modèle, définies avant l'entraînement. Ils incluent c
    a. Le nombre de couches cachées
    b. Le nombre de neurones par couche
    c. Le taux d'apprentissage
    d. La fonction de coût
    e. L'algorithme d'optimisation

L'optimisation des hyperparamètres est cruciale pour les performances du modèle et fait souvent l'objet de recherches approfondies. (Bergstra & Bengio, 2012) ont démontré que la recherche aléatoire des hyperparamètres peut être plus efficace que la recherche par grille traditionnelle. Plus récemment, (Feurer & Hutter, 2019) ont proposé des méthodes d'optimisation automatique des hyperparamètres basées sur l'apprentissage par renforcement, ouvrant la voie à des systèmes d'apprentissage automatique plus autonomes.

3.2.3.3.3 Processus d'apprentissage

L'entraînement d'un RNA implique l'ajustement itératif des paramètres du modèle (poids et biais) pour minimiser une fonction de coût prédéfinie. Ce processus, généralement réalisé par rétropropagation du gradient (Rumelhart et al., 1986), vise à améliorer progressivement la capacité du réseau à faire des prédictions précises sur les données d'entraînement tout en maintenant une bonne généralisation sur des données non vues.

Cependant, l'augmentation du nombre de couches, bien qu'elle permette potentiellement une modélisation plus fine des données, peut également conduire à des problèmes de surapprentissage, où le modèle mémorise les données d'entraînement au détriment de sa capacité de généralisation (Goodfellow et al., 2016). (C. Zhang et al., 2021) ont mis en évidence ce phénomène, montrant que les réseaux profonds peuvent parfaitement ajuster des données aléatoires, soulignant ainsi l'importance des techniques de régularisation.

Des techniques de régularisation, telles que le dropout (Srivastava et al., 2014) ou la normalisation par lots (Ioffe & Szegedy, 2015), ont été développées pour atténuer ces problèmes et améliorer la robustesse des modèles profonds. Le dropout, en particulier, s'est révélé être une technique efficace pour prévenir le surapprentissage en introduisant une forme de régularisation stochastique. La normalisation par lots, quant à elle, a permis d'accélérer considérablement l'entraînement des réseaux profonds en stabilisant la distribution des activations à travers les couches.

Ces avancées dans l'architecture et l'entraînement des réseaux de neurones artificiels ont ouvert la voie à des modèles toujours plus profonds et plus performants, capables de traiter des tâches de plus en plus complexes. Cependant, comme le soulignent (Marcus, 2018) et (Pearl, 2018), des défis importants subsistent, notamment en termes d'interprétabilité des modèles et de leur capacité à généraliser au-delà des données d'entraînement.

3.2.3.4 Fonctions d'activation

Les fonctions d'activation jouent un rôle crucial dans les réseaux de neurones artificiels en introduisant des non-linéarités essentielles à la modélisation de relations complexes dans les données (Nwankpa et al., 2020). Elles déterminent la sortie d'un neurone en fonction de son entrée pondérée, permettant ainsi au réseau d'apprendre des représentations riches et non linéaires.

Plusieurs fonctions d'activation sont couramment utilisées, chacune ayant ses caractéristiques et domaines d'application spécifiques. Les représentations graphiques des fonctions d'activation et leurs derivées utilisées dans cette section sont basées sur les illustrations fournies par (Anushruthika, 2023).

1. Sigmoïde :

Équation 14 Fonction sigmoïde et son gradient

$$S(x) = \frac{1}{1+e^{-x}} \qquad S'(x) = S(x) \times (1 - S(x))$$

La fonction sigmoïde, dont l'équation et le gradient sont donnés par l'Équation 14, est représentée graphiquement par la Figure 18. Cette fonction, bornée entre 0 et 1, est particulièrement adaptée aux problèmes de classification binaire et à la production de probabilités, comme l'ont démontré (Han & Moraga, 1995) dans leurs travaux pionniers sur les réseaux de neurones.

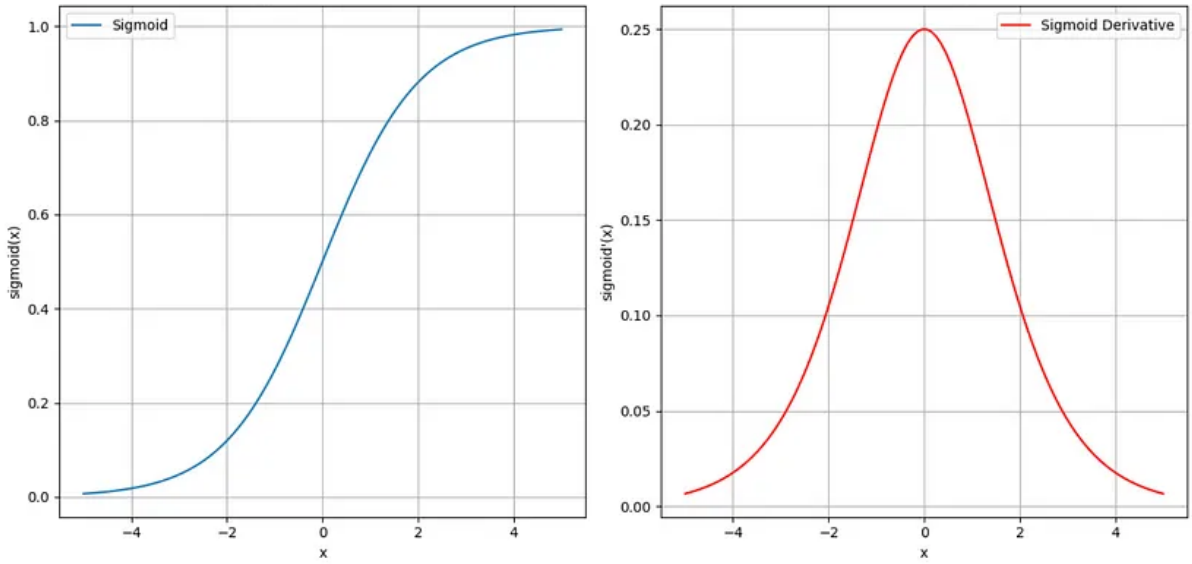


Figure 18 Représentation graphique de la fonction sigmoïde et de son gradient

La fonction sigmoïde trouve son application principale dans la couche de sortie des réseaux de neurones pour les problèmes de classification binaire, tels que la détection de spam dans les e-mails, comme l'ont illustré (Goodfellow et al., 2016).

2. Tangente hyperbolique (tanh) :

Équation 15 Fonction tangente hyperbolique et son gradient

$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \qquad tanh'(x) = 1 - tanh\text{^}2(x)$$

La fonction tangente hyperbolique (tanh), présentée dans l'Équation 15 et illustrée par la Figure 19, est souvent employée dans les couches cachées des réseaux de neurones. (Olgac & Karlik, 2011) ont mis en évidence ses avantages par rapport à la sigmoïde dans certaines architectures de réseaux.

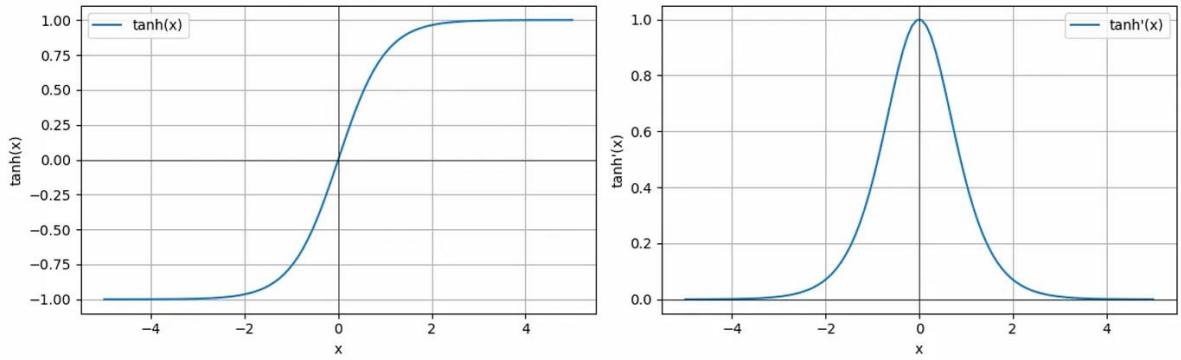


Figure 19 Représentation graphique de la fonction tanh et de son gradient

La fonction tanh est fréquemment utilisée dans les réseaux de neurones récurrents (RNN) pour le traitement du langage naturel, notamment dans des tâches complexes comme la traduction automatique (Sutskever et al., 2014).

3. ReLU (Rectified Linear Unit) :

Équation 16 Fonction ReLU et son gradient

$$ReLU(x) = max(0, x) \qquad ReLU'(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

La fonction ReLU (Rectified Linear Unit), introduite par (Nair & Hinton, 2010), est devenue prédominante dans les architectures profondes. Son équation et son gradient sont donnés par l'Équation 16, et sa représentation graphique est illustrée par la Figure 20.

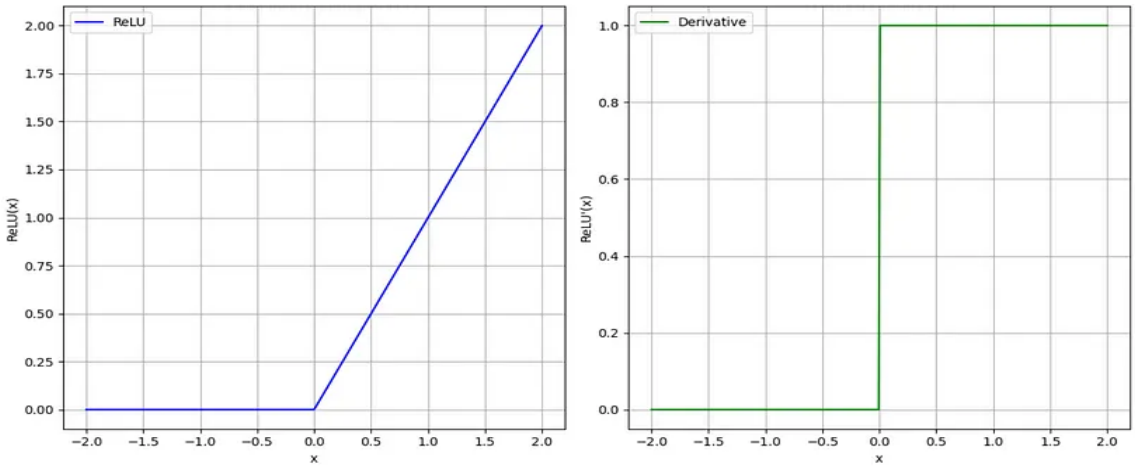


Figure 20 Représentation graphique de la fonction ReLU et de son gradient

La ReLU est largement utilisée dans les réseaux convolutifs (CNN) pour la reconnaissance d'images et la vision par ordinateur, comme l'ont démontré (Krizhevsky et al., 2012) dans leur travail sur ImageNet.

4. Leaky ReLU :

Équation 17 Fonction Leaky ReLU et son gradient

$$Leaky\ ReLU(x) = max(\alpha x, x) \qquad LeakyReLU'(x) = \begin{cases} 1, & x > 0 \\ \alpha, & x \leq 0 \end{cases}$$

où α est un petit nombre positif (généralement 0,01)

La fonction Leaky ReLU, une variante de la ReLU proposée par (Maas et al., 2013), est définie par l'Équation 17 et représentée graphiquement dans la Figure 21. Elle permet de conserver un faible gradient pour les entrées négatives, résolvant partiellement le problème des "neurones morts".

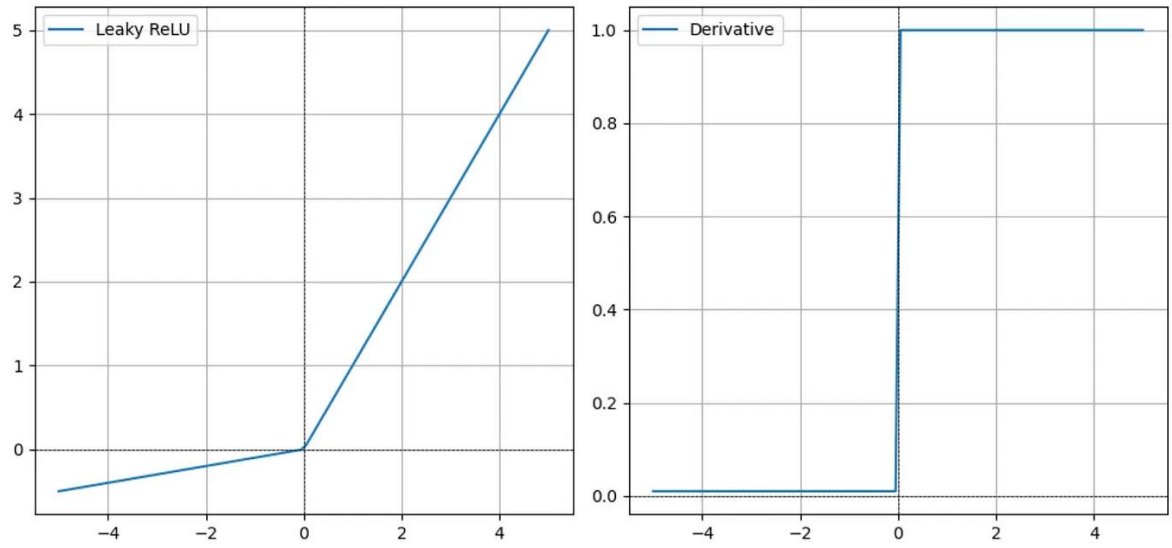


Figure 21 Représentation graphique de la fonction Leaky ReLU et de son gradient

Cette fonction est utilisée comme alternative à la ReLU dans les tâches de vision par ordinateur avancées, comme la détection d'objets en temps réel (He et al., 2015).

5. Softmax :

Équation 18 Fonction Softmax et son gradient

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad \frac{\partial \sigma(z)_i}{\partial z_j} = \begin{cases} \sigma(z)_i\big(1 - \sigma(z)_j\big), i = j \\ -\sigma(z)_j \sigma(z)_i, i \neq j \end{cases}$$

La fonction Softmax, généralement utilisée dans la couche de sortie pour les problèmes de classification multiclasse, est définie par l'Équation 18 et représentée graphiquement dans la Figure 22. Cette fonction, introduite par (Bridle, 1990), normalise un vecteur de scores en une distribution de probabilités. Son gradient est également présenté dans la même équation pour une compréhension complète.

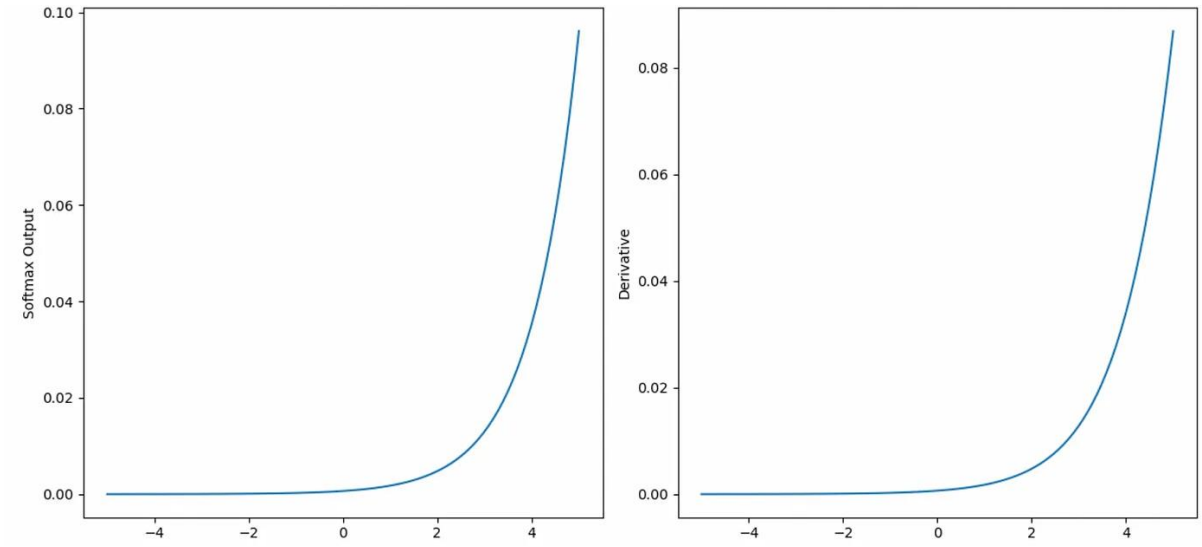


Figure 22 Représentation graphique de la fonction Softmax et de son gradient

La fonction Softmax est couramment utilisée dans la couche de sortie des réseaux pour la classification multiclasse, comme la reconnaissance de caractères manuscrits. Son efficacité dans ce domaine a été démontrée par (Lecun et al., 1998) dans leurs travaux sur les réseaux convolutifs pour la reconnaissance de chiffres manuscrits.

Chacune de ces fonctions d'activation joue un rôle crucial dans l'apprentissage et la performance des réseaux de neurones. Le choix de la fonction d'activation appropriée dépend de la nature du problème, de l'architecture du réseau et des caractéristiques spécifiques de la tâche à accomplir. Comme l'ont souligné (Ramachandran et al., 2017) dans leur étude comparative des fonctions d'activation, il n'existe pas de solution universelle, et l'expérimentation reste souvent nécessaire pour déterminer la meilleure fonction d'activation pour une application donnée.

3.2.3.5 Processus d'apprentissage

Le processus d'apprentissage dans les réseaux de neurones profonds se déroule en deux phases principales :

1. Propagation avant (forward pass) :

   Dans cette phase, les données d'entrée sont propagées à travers le réseau, couche par couche, jusqu'à la sortie. Chaque neurone applique sa fonction d'activation à la somme pondérée de ses entrées, produisant ainsi une prédiction (Rumelhart et al., 1986).

2. Rétropropagation du gradient (backpropagation) :

   Introduite par (Rumelhart et al., 1986) et formalisée par (Schmidhuber, 1992), cette phase consiste à calculer le gradient de la fonction de coût par rapport aux paramètres du réseau. Ce gradient est ensuite utilisé pour ajuster les poids et les biais du réseau dans le sens opposé au gradient, minimisant ainsi l'erreur entre les prédictions du réseau et les sorties désirées.

L'objectif de ce processus itératif est de minimiser une fonction de coût prédéfinie, généralement par la méthode de la descente de gradient ou ses variantes (L. Bottou, 2010). Cette approche permet au réseau d'apprendre progressivement des représentations de plus en plus adaptées aux données d'entraînement, tout en maintenant une capacité de généralisation à des données non vues.

La Figure 23 illustre de manière synthétique le fonctionnement d'un réseau de neurones artificiels et son processus d'apprentissage. Il se compose de trois parties principales qui mettent en lumière les concepts clés abordés précédemment.

La partie A présente l'architecture typique d'un réseau de neurones multicouche, comprenant une couche d'entrée, une couche cachée et une couche de sortie. Les connexions entre les neurones, représentées par des flèches, sont associées à des poids (w) qui sont ajustés lors de l'apprentissage.

La partie B détaille le fonctionnement d'un neurone individuel, mettant en évidence le processus de somme pondérée des entrées suivi de l'application d'une fonction d'activation (ici, la fonction ReLU). Cette illustration permet de comprendre comment chaque neurone traite l'information qu'il reçoit.

Enfin, la partie C représente graphiquement le processus d'optimisation par descente de gradient. Elle montre comment les poids du réseau sont ajustés itérativement pour minimiser la fonction de coût L(w), illustrant la recherche d'un optimum global à travers des optimums locaux.

La Figure 23 offre ainsi une vue d'ensemble du processus de propagation avant (forward propagation) où l'information traverse le réseau de l'entrée vers la sortie, et de la rétropropagation (backward propagation) qui permet d'ajuster les poids en fonction de l'erreur constatée.

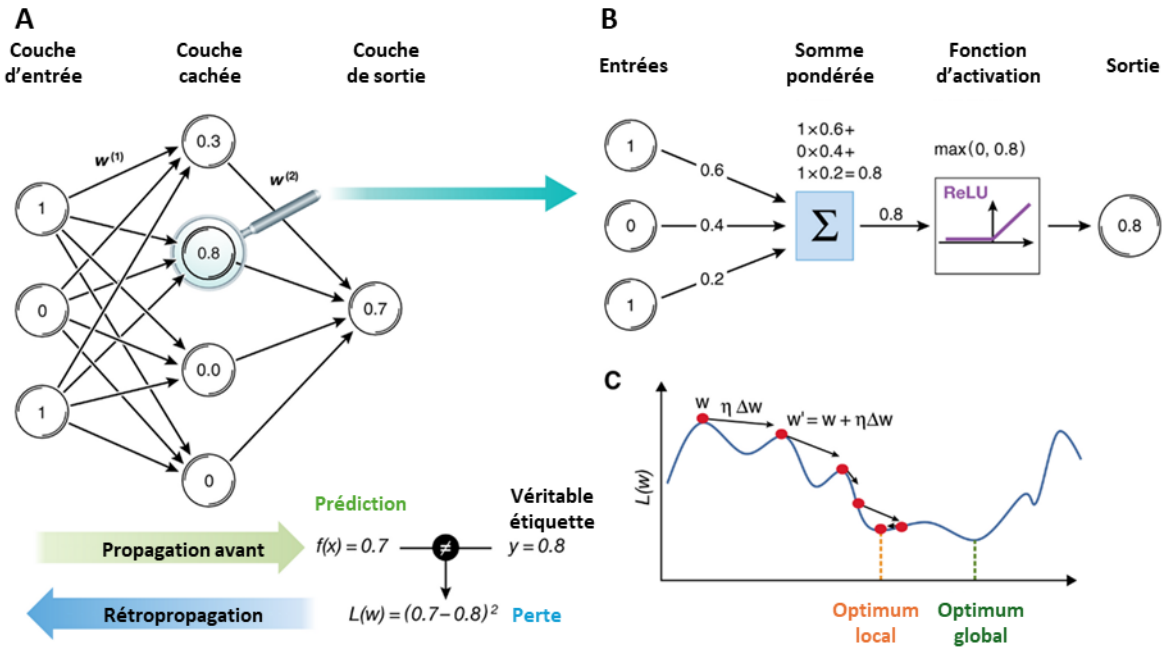


Figure 23 Architecture et fonctionnement d'un réseau de neurones artificiels (Angermueller et al., 2016)

L'efficacité de l'apprentissage dépend fortement du choix judicieux des fonctions d'activation, de l'architecture du réseau, et des hyperparamètres d'optimisation. Ces choix doivent être adaptés à la nature spécifique du problème traité et aux caractéristiques des données disponibles (Goodfellow et al., 2016). Par exemple, (He et al., 2015) ont démontré que l'initialisation des poids en tenant compte de la fonction d'activation utilisée peut grandement améliorer la convergence et les performances des réseaux profonds.

De plus, des techniques avancées comme la normalisation par lots (Ioffe & Szegedy, 2015) et le dropout (Srivastava et al., 2014) ont été développées pour améliorer la stabilité et la généralisation de l'apprentissage dans les réseaux profonds. Ces méthodes, combinées à des architectures de plus en plus sophistiquées, ont permis des avancées significatives dans divers domaines d'application de l'apprentissage profond.

3.2.3.6 Réseaux de neurones en graphes

Les Graph Neural Networks (GNN) représentent une extension significative des réseaux de neurones profonds, spécifiquement conçue pour traiter des données structurées sous forme de graphes (Scarselli et al., 2009). Cette approche est particulièrement pertinente pour le traitement des données relatives aux ontologies, car elle permet d'exploiter efficacement la structure des

graphes de connaissances, comme l'ont démontré (Ristoski & Paulheim, 2016) dans leur étude sur l'apprentissage sémantique.

3.2.3.6.1 Principes fondamentaux des GNN

Les GNN opèrent sur des graphes, structures de données composées de nœuds (ou sommets) et d'arêtes (ou liens) connectant ces nœuds. Leur fonctionnement repose sur des calculs locaux effectués sur chaque nœud du graphe, suivis d'une propagation de ces informations aux nœuds voisins. Ce processus itératif permet de capturer les dépendances de voisinage à différentes échelles, générant ainsi des représentations vectorielles pour les nœuds, les arêtes ou l'ensemble du graphe (Zhou et al., 2020). La Figure 24 illustre l'architecture et la visualisation d'un réseau de neurones en graphe (GNN), où (a) montre la structure d'un réseau convolutionnel en graphe avec ses couches d'entrée, cachées et de sortie, et (b) présente une visualisation des activations de la couche cachée.

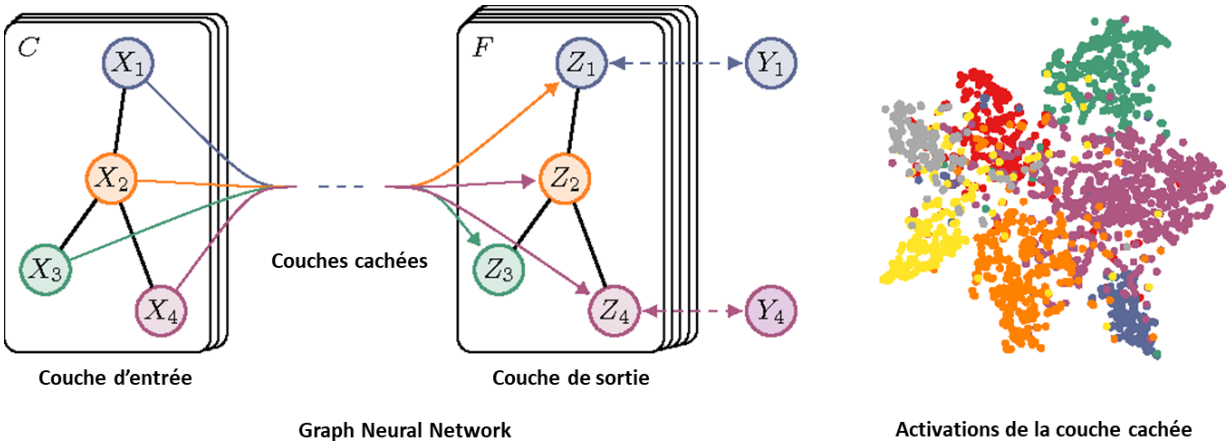


Figure 24 Architecture et visualisation d'un GNN (Kipf & Welling, 2017)

3.2.3.6.2 Applications des GNN

Les GNN ont démontré leur efficacité dans divers domaines, comme le soulignent (Zhou et al., 2020) et (Wu et al., 2021). En chimie, (Gilmer et al., 2017) ont utilisé les GNN pour prédire les propriétés des molécules, tandis qu'en biologie, (Zitnik & Leskovec, 2017) ont appliqué ces modèles à l'analyse des réseaux de régulation des gènes. Dans le domaine des réseaux sociaux, (Hamilton et al., 2017a) ont montré l'excellence des GNN dans la détection de communautés. (R. Ying et al., 2018) ont démontré leur efficacité dans les systèmes de recommandation basés sur des graphes, et (G. Li et al., 2019) ont exploité les GNN en vision par ordinateur pour la segmentation d'images.

Dans le contexte spécifique des ontologies et des graphes de connaissances, les GNN ont montré leur utilité dans plusieurs applications clés :

1. Complétion de graphes de connaissances : (Schlichtkrull et al., 2018) ont utilisé des GNN pour prédire des liens manquants dans des graphes de connaissances, améliorant ainsi la complétude des ontologies.
2. Classification d'entités : (H. Wang et al., 2019) ont appliqué des GNN pour classifier des entités dans des graphes de connaissances, permettant une meilleure organisation des concepts dans les ontologies.

3. Alignement d'ontologies : (Z. Sun et al., 2017) ont proposé une approche basée sur les GNN pour aligner différentes ontologies, facilitant l'interopérabilité entre différents systèmes de connaissances.
4. Raisonnement sur les graphes de connaissances : (C. Zhang et al., 2019) ont utilisé des GNN pour effectuer des inférences complexes sur des graphes de connaissances, améliorant ainsi les capacités de raisonnement basées sur les ontologies.
5. Extraction de relations sémantiques : (Vashishth et al., 2019) ont développé une méthode utilisant des GNN pour extraire des relations sémantiques complexes à partir de textes, enrichissant ainsi les ontologies existantes.

3.2.3.6.3 Défis et perspectives de recherche

Malgré leurs succès, les GNN font face à plusieurs défis dans le contexte du traitement des graphes de connaissances :

1. Scalabilité : le traitement de graphes de connaissances de grande taille reste un défi. Des approches comme celle de (W.-L. Chiang et al., 2019) proposent des méthodes d'échantillonnage pour améliorer l'efficacité computationnelle des GNN sur de grands graphes.
2. Gestion de l'hétérogénéité : les graphes de connaissances sont souvent hétérogènes, avec différents types de nœuds et d'arêtes. (C. Zhang et al., 2019) ont proposé des GNN hétérogènes pour mieux gérer cette complexité.
3. Interprétabilité : l'interprétation des décisions des GNN est cruciale dans le contexte des ontologies. (Z. Ying et al., 2019) ont développé des techniques d'explication pour les GNN, améliorant leur transparence.
4. Intégration de connaissances externes : l'incorporation de connaissances externes dans les GNN pour améliorer le traitement des ontologies reste un domaine de recherche actif (Z. Liu et al., 2019).
5. Apprentissage continu : la mise à jour dynamique des modèles GNN pour refléter l'évolution des ontologies est un défi important, abordé par des travaux comme ceux de (Sankar et al., 2020).

Ces applications variées témoignent de la capacité des GNN à modéliser efficacement les relations complexes et non euclidiennes entre les entités, faisant d'eux une méthode puissante et polyvalente pour le traitement des données structurées en graphes, y compris dans le contexte des ontologies et des graphes de connaissances. Comme l'ont souligné (Battaglia et al., 2018), les GNN offrent un cadre flexible pour l'apprentissage relationnel, ouvrant de nouvelles perspectives pour l'exploitation des connaissances structurées.

3.2.3.7 Synthèse: « Deep Learning » vs « Machine Learning » classique

Le « Deep Learning » et le « Machine Learning » classique représentent deux approches distinctes de l'apprentissage automatique, chacune présentant des avantages et des inconvénients spécifiques. Cette synthèse vise à comparer ces deux approches en mettant en évidence leurs principales caractéristiques, forces et limitations, dans le contexte de notre problématique d'enrichissement des jumeaux numériques par l'intégration de données provenant de sources externes, notamment de l'Internet des Objets Industriel (IIoT).

Avantages du « Deep Learning » :

- Traitement de données complexes :

Les modèles de « Deep Learning », en particulier les réseaux de neurones profonds, excellent dans le traitement de données complexes et non structurées telles que les signaux issus de capteurs IIoT et les données hétérogènes des jumeaux numériques (LeCun et al., 2015). Leur capacité à apprendre automatiquement des représentations hiérarchiques permet de capturer des caractéristiques de haut niveau difficiles à modéliser avec des méthodes traditionnelles, ce qui est particulièrement pertinent pour notre problématique multi-modèle et multi-métiers.

- Performance supérieure :

Dans de nombreux domaines liés à notre problématique, notamment le traitement de séries temporelles issues de capteurs IIoT et l'analyse de données multidimensionnelles des jumeaux numériques, les modèles de Deep Learning surpassent significativement les méthodes de « Machine Learning » classiques (He et al., 2016); (Devlin et al., 2019). Cette supériorité est cruciale pour extraire des informations pertinentes de nos ensembles de données complexes.

- Réduction de l'ingénierie des caractéristiques :

Les réseaux de neurones profonds peuvent apprendre automatiquement les caractéristiques pertinentes à partir des données brutes de l'IIoT et des jumeaux numériques, réduisant ainsi le besoin d'une ingénierie manuelle des caractéristiques, souvent coûteuse en temps et en expertise (Bengio et al., 2013). Cette capacité est particulièrement avantageuse dans notre contexte où les sources de données sont diverses et évolutives.

Limitations du « Deep Learning » :

- Dépendance aux grandes quantités de données :

Les modèles de « Deep Learning » nécessitent généralement de vastes ensembles de données pour être efficaces. Avec des données limitées, ils sont sujets au surapprentissage, compromettant leur capacité de généralisation (C. Sun et al., 2017). Dans notre cas, l'abondance des données IIoT et des jumeaux numériques atténue cette limitation.

- Coûts computationnels élevés :

L'entraînement et l'inférence des modèles de « Deep Learning », en particulier pour les réseaux profonds, requièrent des ressources informatiques importantes, entraînant des coûts élevés en termes de matériel et de temps de calcul (Strubell et al., 2019). Cependant, dans un contexte industriel, ces coûts peuvent être justifiés par les gains potentiels en termes de performance et d'efficacité opérationnelle.

- Complexité et opacité :

Les modèles de « Deep Learning » sont souvent considérés comme des « boîtes noires » en raison de leur complexité, rendant difficile l'interprétation de leurs décisions. Cela soulève des questions d'explicabilité et de responsabilité, particulièrement critiques dans certains domaines d'application (Lipton, 2018); (Rudin, 2019). Dans notre contexte industriel, des efforts supplémentaires seront nécessaires pour assurer l'interprétabilité des résultats.

Avantages du « Machine Learning » classique :

- Efficacité avec des données limitées :

Les méthodes classiques, telles que les arbres de décision ou les machines à vecteurs de support, peuvent être plus efficaces lorsque les données d'entraînement sont limitées (Hastie et al., 2009). Cependant, dans notre cas, la disponibilité de grandes quantités de données IIoT réduit l'importance de cet avantage.

- Interprétabilité :

Les modèles de « Machine Learning » classique, comme les arbres de décision ou les modèles linéaires, sont généralement plus simples et plus faciles à interpréter, facilitant la compréhension et l'explication de leurs décisions (Molnar, 2020). Bien que cet aspect soit important, la complexité de notre problématique peut nécessiter des modèles plus sophistiqués.

- Efficience computationnelle :

Ces méthodes sont souvent plus rapides à entraîner et moins exigeantes en ressources computationnelles, les rendant plus adaptées à des applications avec des contraintes de ressources (S. Kotsiantis, 2007). Toutefois, dans un environnement industriel, les ressources computationnelles sont généralement disponibles pour supporter des modèles plus complexes si nécessaire.

Limitations du « Machine Learning » classique :

- Difficulté à traiter des données non structurées :

Les méthodes classiques peuvent avoir du mal à gérer efficacement les données non structurées ou de haute dimension, qui sont courantes dans les environnements IIoT et les jumeaux numériques (Domingos, 2012).

- Performance limitée sur des tâches complexes :

Pour des problèmes impliquant des relations non linéaires complexes ou des interactions subtiles entre variables, les méthodes classiques peuvent atteindre leurs limites en termes de performance (Jordan & Mitchell, 2015).

- Nécessité d'une ingénierie des caractéristiques manuelle :

Contrairement au « Deep Learning », les méthodes classiques nécessitent souvent une sélection et une ingénierie manuelle des caractéristiques, ce qui peut être chronophage et moins adapté à des environnements dynamiques comme le nôtre (Guyon & Elisseeff, 2003).

En conclusion, pour notre problématique d'enrichissement des jumeaux numériques par l'intégration de données provenant de sources externes, notamment de l'IIoT, nous optons pour l'utilisation du « Deep Learning. » Ce choix est motivé par plusieurs facteurs clés :

1. La complexité et l'hétérogénéité des données : les données issues de l'IIoT et des jumeaux numériques sont souvent non structurées, de haute dimension et hétérogènes. Le « Deep Learning » est particulièrement adapté pour traiter ce type de données complexes sans nécessiter une ingénierie manuelle des caractéristiques extensive.

2. Le volume important de données : l'environnement industriel génère de grandes quantités de données, ce qui est favorable à l'utilisation de modèles de Deep Learning qui excellent dans l'exploitation de vastes ensembles de données.
3. La capacité à capturer des relations complexes : notre problématique multi-modèle et multi-métiers implique des interactions complexes entre différentes sources de données. Les architectures profondes du Deep Learning sont capables de modéliser ces relations non linéaires de manière plus efficace que les méthodes classiques.
4. L'adaptabilité aux changements dynamiques : l'environnement industriel est en constante évolution. La capacité du « Deep Learning » à apprendre continuellement et à s'adapter à de nouvelles données est un atout majeur pour notre application.
5. Les performances supérieures : dans des domaines similaires au nôtre, tels que l'analyse de séries temporelles industrielles et la modélisation de systèmes complexes, le « Deep Learning » a démontré des performances supérieures aux méthodes classiques.

Bien que nous reconnaissions les défis liés à l'interprétabilité et aux coûts computationnels du Deep Learning, nous estimons que les avantages en termes de performance et de capacité à traiter des données complexes l'emportent dans notre contexte. Pour atténuer ces limitations, nous prévoyons d'intégrer des techniques d'interprétabilité spécifiques au « Deep Learning » et d'optimiser l'utilisation des ressources computationnelles disponibles dans l'environnement industriel.

Dans le chapitre suivant, nous explorerons donc en détail les architectures de Deep Learning spécifiques, telles que les réseaux de neurones convolutifs (CNN) pour le traitement des données spatiales, les réseaux de neurones récurrents (RNN) pour les séries temporelles, et les « Graph Neural Networks » (GNN) pour modéliser les relations complexes dans les jumeaux numériques et les ontologies. Cette approche nous permettra de développer une solution robuste et performante pour l'enrichissement des jumeaux numériques dans un contexte industriel complexe et dynamique.

3.2.4 Techniques de plongement (« embedding »)

3.2.4.1 Introduction au traitement automatique du langage naturel (TALN)

Le traitement automatique du langage naturel (TALN), ou « Natural Language Processing » (NLP) en anglais, est un domaine interdisciplinaire à l'intersection de l'intelligence artificielle, de la linguistique computationnelle et de l'informatique. Son objectif principal est de développer des systèmes capables de comprendre, d'interpréter et de générer le langage humain sous ses diverses formes (Jurafsky & Martin, 2024). L'importance du TALN s'est considérablement accrue ces dernières années, en raison de l'explosion des données textuelles disponibles et de la nécessité croissante d'interagir avec les machines de manière naturelle.

Dans ce contexte, les ontologies jouent un rôle crucial en fournissant une structure formelle et une représentation explicite des connaissances dans des domaines spécifiques (Staab & Studer, 2009). Comme l'ont souligné (Guarino et al., 2009), les ontologies offrent un cadre sémantique riche qui peut être exploité pour améliorer la compréhension et l'interprétation du langage naturel par les machines. Elles permettent de formaliser les relations entre les concepts, facilitant ainsi l'analyse sémantique des textes.

L'un des défis majeurs du TALN est la représentation efficace des mots et des phrases d'une manière qui soit à la fois compréhensible par les machines et capable de capturer les nuances sémantiques et syntaxiques du langage. Comme l'ont démontré (Collobert & Weston, 2008), les représentations traditionnelles des mots, telles que les encodages one-hot, sont souvent insuffisantes pour capturer la richesse sémantique du langage. C'est dans ce contexte que les techniques de plongement lexical, ou « word embeddings », ont émergé comme une avancée significative (Mikolov, Chen, et al., 2013).

Les techniques de plongement lexical visent à représenter les mots dans un espace vectoriel continu de faible dimension, où les relations sémantiques entre les mots sont préservées. Cette approche a révolutionné de nombreuses tâches du TALN, comme l'ont démontré (Pennington et al., 2014) avec leur modèle GloVe. Ces représentations vectorielles permettent aux modèles d'apprentissage automatique de mieux comprendre le contexte et les nuances du langage, améliorant ainsi les performances sur diverses tâches telles que la traduction automatique, l'analyse de sentiment, et la réponse aux questions.

3.2.4.2 Représentations traditionnelles des mots

Avant l'avènement des techniques de plongement lexical, les approches traditionnelles de représentation des mots dans le TALN reposaient principalement sur des méthodes discrètes. Parmi ces méthodes, l'encodage « one-hot » était largement utilisé en raison de sa simplicité (Jurafsky & Martin, 2024).

L'encodage « one-hot » représente chaque mot par un vecteur binaire de grande dimension, généralement égale à la taille du vocabulaire. Dans ce vecteur, tous les éléments sont à 0, sauf celui correspondant à l'index du mot dans le vocabulaire, qui est à 1. Par exemple, pour un vocabulaire de quatre mots {maison, chat, chien, voiture}, les vecteurs « one-hot » seraient :

maison : [1, 0, 0, 0]

chat : [0, 1, 0, 0]

chien : [0, 0, 1, 0]

voiture : [0, 0, 0, 1]

Bien que simple à mettre en œuvre, cette approche présente plusieurs limitations significatives :

1. Dimensionnalité élevée : pour un vocabulaire de taille n, chaque mot est représenté par un vecteur de dimension n. Cette haute dimensionnalité peut entraîner des problèmes de stockage et de calcul, en particulier pour de grands vocabulaires (Y. Goldberg, 2017). Comme l'ont souligné (Bengio et al., 2003), cette dimensionnalité élevée peut également conduire au problème de la « malédiction de la dimensionnalité », rendant difficile l'apprentissage de modèles statistiques robustes.
2. Absence de sémantique : les vecteurs « one-hot » sont orthogonaux entre eux, ne capturant aucune information sur les relations sémantiques entre les mots. Par exemple, "chat" et "chien" sont sémantiquement plus proches que "chat" et "voiture", mais cette relation n'est pas reflétée dans leurs représentations « one-hot » (Mikolov, Sutskever, et al., 2013). Cette limitation a été particulièrement mise en évidence par (Turney & Pantel, 2010) dans leur étude sur la sémantique distributionnelle.

3. Sparsité : ces vecteurs sont extrêmement creux (« sparse »), avec une seule valeur non nulle. Cette sparsité peut limiter l'efficacité de certains algorithmes d'apprentissage automatique (Bengio et al., 2003). Comme l'ont démontré (Blei et al., 2003) dans leur travail sur l'allocation de Dirichlet latente, la sparsité peut rendre difficile la modélisation de thèmes latents dans les textes.
4. Incapacité à gérer les mots hors vocabulaire : cette méthode ne peut pas représenter des mots qui n'étaient pas dans le vocabulaire initial, limitant ainsi la flexibilité du système (Bojanowski et al., 2017). Cette limitation est particulièrement problématique dans des domaines spécialisés ou pour des langues morphologiquement riches, comme l'ont souligné (Sennrich et al., 2016) dans leur travail sur la traduction automatique neuronale.

Les limitations des représentations traditionnelles ont motivé le développement de techniques de plongement lexical plus avancées, capables de capturer la sémantique des mots dans des espaces vectoriels denses et de dimension réduite. Ces nouvelles approches, que nous explorerons dans les sections suivantes, ont significativement amélioré les performances des modèles de TALN sur une variété de tâches, comme l'ont démontré (Devlin et al., 2019) avec leur modèle BERT.

3.2.4.3 Plongement lexical (Word embeddings)

Le plongement lexical, ou « word embedding » en anglais, représente une avancée significative dans le domaine du traitement automatique du langage naturel (TALN). Cette technique vise à représenter les mots sous forme de vecteurs denses dans un espace continu de faible dimension, capturant ainsi les relations sémantiques et syntaxiques entre les mots de manière plus efficace que les méthodes traditionnelles (Mikolov, Chen, et al., 2013). Cette approche a révolutionné la compréhension du langage par les modèles d'apprentissage automatique, ouvrant la voie à des performances améliorées dans diverses tâches du TALN.

Deux méthodes principales ont marqué l'évolution de ce domaine : Word2Vec et GloVe. Ces approches, bien que différentes dans leur conception, partagent l'objectif commun de générer des représentations vectorielles des mots qui reflètent leur signification et leur usage dans le langage.

3.2.4.3.1 Word2Vec

Word2Vec, introduit par (Mikolov, Chen, et al., 2013), est une méthode d'apprentissage non supervisé qui génère des représentations vectorielles des mots en exploitant leur contexte d'apparition. Cette approche s'appuie sur l'hypothèse distributionnelle, formulée initialement par (Harris, 1954), selon laquelle les mots apparaissant dans des contextes similaires tendent à avoir des significations proches. Cette hypothèse a été largement validée et exploitée dans le domaine de la linguistique computationnelle (Turney & Pantel, 2010).

Word2Vec propose deux architectures distinctes : le Continuous Bag-of-Words (CBOW) et le Skip-Gram. Ces deux modèles diffèrent dans leur approche de la prédiction, mais partagent le même objectif fondamental de capturer les relations sémantiques entre les mots, comme illustré dans la Figure 25.

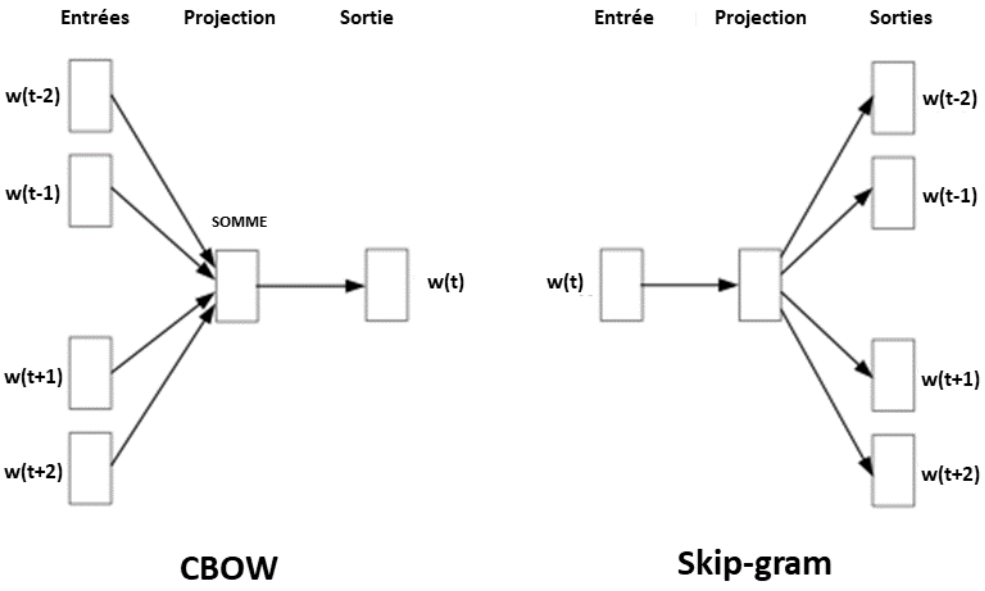


Figure 25 Architectures CBOW et Skip-gram de Word2Vec (Mikolov, Chen, et al., 2013)

L'architecture CBOW, représentée à gauche dans la Figure 25, vise à prédire un mot cible à partir de son contexte, c'est-à-dire les mots qui l'entourent dans une fenêtre de taille fixe. Cette approche permet au modèle d'apprendre des représentations vectorielles qui capturent le sens d'un mot en fonction de son environnement lexical, comme l'ont expliqué (Y. Goldberg & Levy, 2014) dans leur analyse approfondie du modèle. Le CBOW est particulièrement efficace pour traiter de grands corpus de texte et tend à mieux performer sur les mots fréquents.

Le modèle Skip-Gram, illustré à droite dans la Figure 25, inverse la tâche de prédiction en cherchant à prédire le contexte d'un mot à partir du mot lui-même. Cette approche s'est révélée particulièrement efficace pour capturer les relations sémantiques fines entre les mots, comme l'ont démontré (Mikolov, Sutskever, et al., 2013) dans leurs expériences sur diverses tâches linguistiques. Le Skip-Gram est souvent préféré pour les petits corpus ou lorsque l'intérêt se porte particulièrement sur les mots rares.

(Baroni et al., 2014) ont mené une étude comparative approfondie des modèles de représentation vectorielle, y compris Word2Vec, et ont constaté que ces modèles prédictifs surpassent généralement les méthodes traditionnelles basées sur le comptage pour une variété de tâches sémantiques.

Les deux architectures présentent des avantages et des inconvénients selon la taille du corpus et la distribution des mots. CBOW est généralement plus rapide à entraîner et plus performant pour les mots fréquents, tandis que Skip-Gram est plus adapté aux mots rares et aux petits corpus (Mikolov, Chen, et al., 2013). Le choix entre ces deux modèles dépend donc souvent des caractéristiques spécifiques de la tâche et des données disponibles.

De plus, des recherches récentes, comme celles de (Bojanowski et al., 2017), ont étendu le modèle Word2Vec pour prendre en compte la structure interne des mots, permettant ainsi de générer des représentations vectorielles pour des mots hors vocabulaire et d'améliorer les performances sur les langues morphologiquement riches.

La Figure 26 illustre les relations sémantiques capturées par les représentations vectorielles continues, démontrant la capacité de ces modèles à encoder des informations linguistiques complexes dans un espace vectoriel de faible dimension.

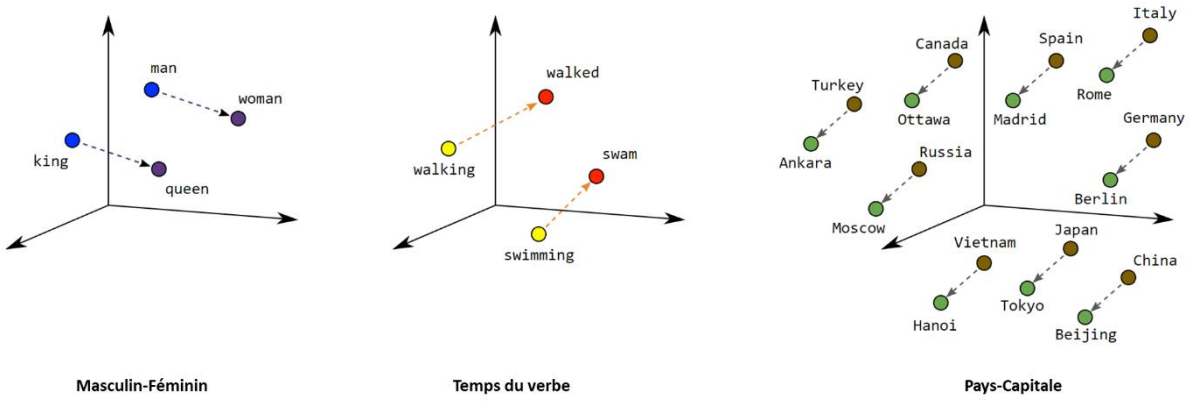


Figure 26 Exemples de relations sémantiques capturées par les représentations vectorielles (Google Developers, s. d.)

La Figure 26 met en évidence la puissance des représentations vectorielles continues, telles que celles générées par Word2Vec, pour capturer diverses relations sémantiques. Trois types de relations y sont observables : male-female, verb tense, et country-capital. Ces exemples illustrent comment les modèles de représentation vectorielle parviennent à encoder des informations linguistiques subtiles et complexes dans un espace multidimensionnel, permettant ainsi des opérations arithmétiques sur le sens des mots. Cette capacité à représenter et à manipuler le sens des mots de manière mathématique ouvre de nombreuses possibilités pour l'analyse sémantique et les applications en traitement du langage naturel.

3.2.4.3.2 GloVe (Global Vectors for Word Representation)

GloVe, proposé par (Pennington et al., 2014), représente une approche novatrice dans le domaine des représentations vectorielles de mots. Contrairement à Word2Vec qui se concentre sur les contextes locaux, GloVe adopte une perspective plus globale en exploitant l'information de cooccurrence des mots dans l'ensemble du corpus.

Le principe fondamental de GloVe repose sur l'idée que les relations sémantiques entre les mots peuvent être capturées en analysant leur fréquence d'apparition conjointe dans le texte. Cette approche permet de saisir des nuances linguistiques subtiles qui pourraient échapper à des méthodes plus localisées. Comme l'expliquent (Y. Goldberg & Levy, 2014), cette méthode offre un équilibre intéressant entre l'exploitation des statistiques globales du corpus et la préservation des informations contextuelles locales.

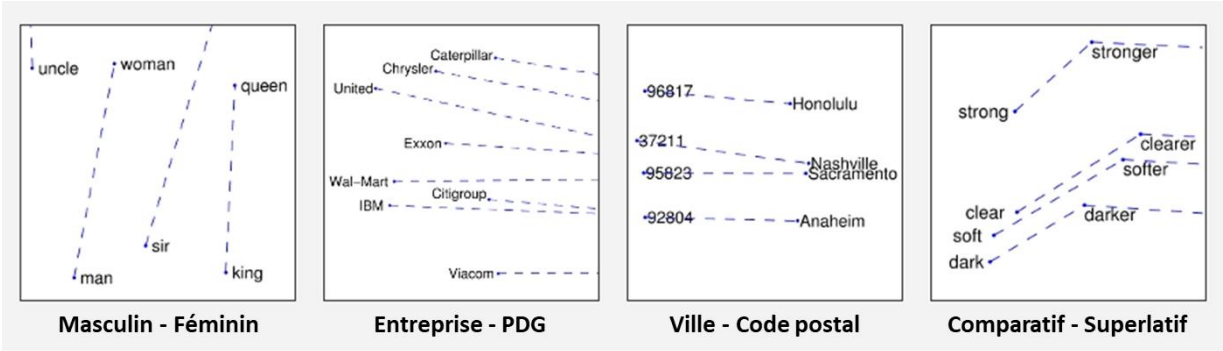


Figure 27 Représentations vectorielles et relations sémantiques capturées par GloVe (Pennington et al., 2014)

La Figure 27 illustre la capacité de GloVe à capturer diverses relations sémantiques et syntaxiques entre les mots. Des exemples de relations analogiques (comme homme-femme, roi-reine), des associations hiérarchiques (entreprise-PDG), des correspondances géographiques (ville-code postal), et des gradations sémantiques (comparatif-superlatif) y sont observables. Cette visualisation démontre la richesse et la versatilité des représentations vectorielles générées par GloVe.

Le processus d'apprentissage de GloVe vise à créer des vecteurs de mots qui, lorsqu'ils sont combinés mathématiquement, reflètent fidèlement les patterns de cooccurrence observés dans le corpus. Cette approche permet de capturer à la fois les relations linéaires et non linéaires entre les mots dans l'espace vectoriel, offrant ainsi une représentation plus riche et nuancée du langage (Pennington et al., 2014).

Un aspect crucial de l'algorithme GloVe est sa fonction de pondération, qui joue un rôle essentiel dans l'équilibrage de l'importance accordée aux différentes paires de mots lors de l'apprentissage. Cette fonction permet d'atténuer l'impact des cooccurrences très fréquentes ou très rares, assurant ainsi une représentation plus équilibrée et robuste du vocabulaire. Comme le soulignent (Schnabel et al., 2015), cette caractéristique contribue significativement à la qualité des vecteurs produits par GloVe.

Les évaluations empiriques ont démontré que GloVe performe particulièrement bien sur des tâches d'analogie et de similarité sémantique. Par exemple, (Levy et al., 2015) ont constaté que GloVe surpasse souvent d'autres méthodes de plongement lexical sur diverses tâches d'évaluation linguistique. Ces résultats soulignent la capacité de GloVe à capturer des nuances sémantiques fines et des relations complexes entre les mots.

Les techniques de plongement lexical comme Word2Vec et GloVe ont significativement amélioré les performances des modèles de TALN sur une variété de tâches, en fournissant des représentations vectorielles denses et sémantiquement riches des mots. Comme l'observent (Young et al., 2018), ces représentations ont permis une meilleure compréhension et modélisation du langage par les systèmes d'intelligence artificielle, améliorant les performances sur une variété de tâches de TALN.

3.2.4.4 Intérêts et limitations des « word embeddings »

Les « word embeddings » ont révolutionné le traitement automatique du langage naturel (TALN) en offrant des représentations vectorielles denses et sémantiquement riches des mots. Cette section examine leurs avantages, applications et limitations.

Avantages des « word embeddings » :

1. Capture de la sémantique et des relations entre les mots : les « word embeddings » permettent de représenter les similarités et les nuances de sens entre les mots dans un espace vectoriel continu. (Mikolov, Sutskever, et al., 2013) ont démontré que les mots ayant des significations proches ou apparaissant dans des contextes similaires ont tendance à avoir des vecteurs proches dans cet espace. Cette propriété facilite grandement l'utilisation de ces représentations dans divers modèles de TALN. Par exemple, (Levy & Goldberg, 2014) ont approfondi cette notion en montrant comment différentes métriques de similarité dans l'espace vectoriel correspondent à différentes relations sémantiques entre les mots.
2. Opérations arithmétiques sémantiquement significatives : une caractéristique remarquable des « word embeddings » est la possibilité d'effectuer des opérations arithmétiques sur les vecteurs, produisant des résultats sémantiquement cohérents. L'exemple classique "roi" - "homme" + "femme" ≈ "reine", mis en évidence par (Mikolov, Sutskever, et al., 2013), illustre comment ces représentations parviennent à capturer et à quantifier des relations sémantiques complexes entre les mots. (Gittens et al., 2017) ont fourni une analyse théorique de ce phénomène, expliquant pourquoi ces opérations vectorielles reflètent des relations sémantiques.
3. Indépendance linguistique : les techniques de création de « word embeddings » sont généralement indépendantes de la langue, comme l'ont montré (Bojanowski et al., 2017) avec FastText. Cette propriété permet d'appliquer des approches similaires pour générer des « embeddings » dans différentes langues, facilitant ainsi l'adaptation et l'extension des modèles de TALN à divers contextes linguistiques. (Ruder et al., 2019) ont exploré plus en détail les méthodes d'apprentissage de représentations multilingues, ouvrant la voie à des applications de TALN véritablement multilingues.

Applications majeures :

Les « word embeddings » ont trouvé de nombreuses applications dans le domaine du TALN :

1. Traduction automatique : (Hill et al., 2015) ont montré comment les « embeddings » améliorent la représentation des mots dans les langues source et cible, augmentant ainsi les performances des modèles de traduction. (Artetxe et al., 2018) ont poussé cette idée plus loin en développant des méthodes pour aligner les espaces vectoriels de différentes langues sans supervision, permettant ainsi une traduction automatique non supervisée.
2. Génération de texte : (Lebret et al., 2016) ont utilisé les « embeddings » pour améliorer la compréhension du contexte et la production de textes plus cohérents et pertinents. Plus récemment, (Radford et al., 2019) ont montré comment des modèles pré-entraînés sur de vastes corpus, utilisant des représentations contextuelles basées sur les « embeddings », peuvent générer des textes de haute qualité sur une variété de sujets.
3. Classification de texte : (D. Tang et al., 2015) ont démontré l'efficacité des « embeddings » pour représenter les documents dans des tâches telles que l'analyse de sentiment ou la

détection de thèmes. (Howard & Ruder, 2018) ont étendu cette approche avec leur méthode ULMFiT, qui utilise des « embeddings » pré-entraînés pour améliorer significativement les performances de classification sur une variété de tâches.

4. Réponse aux questions : (Bordes et al., 2014) ont exploité les « embeddings » pour améliorer la compréhension des requêtes et des documents, permettant des réponses plus pertinentes. (D. Chen et al., 2017) ont poussé cette approche plus loin en développant des modèles de lecture et de compréhension basés sur des « embeddings » contextuels, capables de répondre à des questions complexes sur des textes longs.
5. Détection d'anomalies : (W. E. Zhang et al., 2020) ont utilisé les « embeddings » pour identifier des mots ou phrases ne correspondant pas au contexte général d'un document, améliorant ainsi la qualité et la sécurité des textes analysés. (Ruff et al., 2021) ont étendu cette approche en proposant des méthodes de détection d'anomalies basées sur l'apprentissage profond et les « embeddings », applicables à une variété de domaines au-delà du TALN.

Limitations et défis :

Malgré leurs nombreux avantages, les « word embeddings » présentent certaines limitations :

1. Polysémie et homonymie : Comme l'ont souligné (Faruqui et al., 2016), les « embeddings » classiques ne peuvent pas représenter efficacement les différentes significations d'un mot polysémique, chaque mot étant associé à un seul vecteur. (Peters et al., 2018) ont proposé une solution partielle à ce problème avec leur modèle ELMo, qui génère des représentations contextuelles des mots.
2. Biais et stéréotypes : (Bolukbasi et al., 2016) ont mis en évidence que les « embeddings » peuvent apprendre et perpétuer des biais présents dans les données d'entraînement, conduisant à des représentations potentiellement discriminatoires. (Caliskan et al., 2017) ont approfondi cette question en montrant comment les biais culturels et historiques se reflètent dans les « embeddings », soulignant l'importance de développer des méthodes pour atténuer ces biais.
3. Évolution du langage : (Rodriguez & Spirling, 2022) ont souligné la difficulté des embeddings statiques à s'adapter aux changements linguistiques, limitant leur efficacité pour des données textuelles récentes ou en évolution rapide. (Z. Yao et al., 2018) ont proposé des méthodes pour créer des « embeddings » dynamiques qui évoluent avec le temps, offrant une solution potentielle à ce problème.
4. Évaluation et interprétation : (Gladkova & Drozd, 2016) ont mis en lumière les défis liés à l'évaluation objective des performances des word embeddings et à l'interprétation de leurs résultats dans différentes tâches. (Schnabel et al., 2015) ont proposé un cadre d'évaluation plus complet pour les « embeddings », prenant en compte une variété de tâches et de métriques.

Approches récentes et perspectives :

Pour surmonter certaines de ces limitations, des approches plus avancées ont été développées :

1. « Debiasing techniques » : des méthodes comme celles proposées par (J. Zhao et al., 2018) visent à réduire les biais dans les « embeddings », améliorant ainsi leur équité et leur applicabilité dans des contextes sensibles. (Gonen & Goldberg, 2019) ont cependant

montré que ces techniques peuvent être insuffisantes, soulignant la nécessité de recherches continues dans ce domaine.

2. « Embeddings » multimodaux : l'intégration d'informations provenant de différentes modalités (texte, image, son) dans les « embeddings », comme proposé par (Kiela et al., 2016), ouvre de nouvelles perspectives pour des représentations plus riches et plus robustes. (Baltrušaitis et al., 2019) ont fourni une revue complète des techniques d'apprentissage multimodal, soulignant le potentiel de ces approches pour améliorer la compréhension du langage naturel.
3. Modèles de langue contextuels : des modèles comme BERT (Devlin et al., 2019) et GPT (Radford et al., 2018) génèrent des représentations dynamiques des mots en fonction de leur contexte d'utilisation, surmontant ainsi certaines limitations des « embeddings » statiques. Ces modèles ont établi de nouveaux standards de performance dans de nombreuses tâches de TALN, comme l'ont montré (A. Wang et al., 2018) dans leur benchmark GLUE.

Pour conclure, les « word embeddings » ont marqué une avancée significative dans le domaine du TALN, offrant des représentations puissantes et flexibles des mots. Bien qu'ils présentent certaines limitations, les développements récents et en cours promettent de surmonter ces défis, ouvrant la voie à des applications encore plus sophistiquées et performantes dans le traitement du langage naturel. Comme l'ont souligné (Q. Liu et al., 2020) dans leur revue des progrès récents en TALN, l'évolution des techniques de représentation du langage continue de repousser les frontières de ce que les machines peuvent accomplir en termes de compréhension et de génération du langage naturel.

3.2.4.5 Extension des techniques de plongement

Les techniques de plongement ont connu une évolution significative avec l'introduction des "paragraph embeddings", une extension visant à générer des représentations vectorielles pour des unités textuelles plus larges telles que les phrases, les paragraphes et les documents entiers. Cette approche, qui cherche à capturer la sémantique et la structure de textes plus longs, élargit considérablement le champ d'application des techniques de plongement dans le traitement automatique du langage naturel (TALN).

Le modèle Doc2Vec, introduit par (Le & Mikolov, 2014), représente une avancée notable dans ce domaine. Ce modèle introduit un "vecteur document" en complément des vecteurs de mots, permettant ainsi de générer une représentation vectorielle unique pour des unités textuelles plus grandes. Doc2Vec peut être implémenté selon des architectures similaires à celles utilisées dans Word2Vec, à savoir CBOW (Continuous Bag-of-Words) ou Skip-Gram. Comme l'ont souligné (Lau & Baldwin, 2016), cette approche permet de capturer efficacement le contexte global d'un document tout en préservant les relations sémantiques entre les mots.

Les "paragraph embeddings" présentent plusieurs avantages significatifs :

1. Capture d'informations contextuelles plus larges : cette capacité est cruciale pour des tâches complexes telles que la classification de texte ou la génération de contenu. Comme l'ont démontré (Dai et al., 2015), ces représentations permettent de mieux saisir la cohérence thématique et la structure argumentative des textes longs.
2. Représentation compacte d'unités textuelles plus grandes : en encodant des documents entiers dans un espace vectoriel de dimension réduite, ces techniques facilitent

considérablement le traitement et l'analyse des données textuelles à grande échelle. (M. Chen, 2017) a montré comment cette propriété peut être exploitée pour améliorer les performances des systèmes de recherche d'information.

3. Comparaison sémantique efficace : la possibilité de comparer directement différents documents, phrases ou paragraphes dans l'espace vectoriel ouvre de nouvelles perspectives pour des tâches telles que le regroupement de documents ou la recherche d'informations similaires. Les travaux de (Kusner et al., 2015) sur la distance de transport de mots (Word Mover's Distance) illustrent l'efficacité de ces représentations pour mesurer la similarité sémantique entre documents.

Cependant, ces techniques présentent également certaines limitations qu'il convient de prendre en compte :

1. Risque de perte d'informations locales : pour des textes longs et complexes, il existe un risque de dilution des informations spécifiques ou localement importantes dans la représentation globale. (Arora et al., 2017) ont proposé des méthodes pour atténuer ce problème en utilisant une décomposition en composantes principales pondérée.
2. Difficulté d'interprétation : la nature de haute dimension et non structurée des représentations vectorielles peut rendre leur interprétation et leur analyse difficiles pour les humains. Des travaux récents, comme ceux de (Iyyer et al., 2015), explorent des moyens de rendre ces représentations plus interprétables tout en préservant leur pouvoir prédictif.

Dans le contexte de notre problématique d'enrichissement des jumeaux numériques, les « paragraph embeddings » offrent des perspectives particulièrement intéressantes. Ils pourraient être utilisés pour représenter de manière compacte et sémantiquement riche des descriptions textuelles de composants, de processus ou de systèmes entiers au sein des jumeaux numériques. Cette approche pourrait faciliter l'intégration de données textuelles non structurées dans les modèles, améliorant ainsi la capacité des jumeaux numériques à capturer et à exploiter des informations complexes provenant de diverses sources.

Par exemple, (Zheng et al., 2018) ont montré comment les « embeddings » de documents peuvent être utilisés pour améliorer la représentation et l'analyse de données techniques complexes dans le domaine de l'ingénierie. De même, (X. Li et al., 2019) ont exploré l'utilisation de ces techniques pour intégrer des informations textuelles dans des modèles de maintenance prédictive, démontrant ainsi leur potentiel pour enrichir les jumeaux numériques avec des données non structurées.

3.2.4.6 Conclusion sur les techniques de plongement

Les techniques de plongement, depuis les « word embeddings » jusqu'aux "paragraph embeddings", ont indéniablement révolutionné le traitement automatique du langage naturel en offrant des représentations vectorielles denses et sémantiquement riches des unités textuelles. Ces méthodes ont permis des avancées significatives dans diverses tâches de TALN, de la traduction automatique à l'analyse de sentiment, en passant par la génération de texte. Comme l'ont souligné (Young et al., 2018) dans leur revue exhaustive, ces techniques ont transformé la manière dont nous abordons le traitement du langage naturel, ouvrant la voie à des applications toujours plus sophistiquées.

Dans le cadre de notre recherche sur l'enrichissement des jumeaux numériques, ces techniques offrent des opportunités prometteuses pour intégrer efficacement des données textuelles non structurées dans les modèles. Elles pourraient permettre de représenter de manière compacte et sémantiquement pertinente des descriptions textuelles de composants, de processus ou de systèmes entiers, facilitant ainsi leur intégration et leur exploitation dans un environnement multi-modèle et multi-métiers. Les travaux de (Tao, Sui, Liu, Qi, et al., 2019) sur l'intégration de données textuelles dans les jumeaux numériques pour l'industrie 4.0 illustrent le potentiel de ces approches.

Cependant, l'utilisation de ces techniques dans le contexte des jumeaux numériques soulève également de nouveaux défis. Il sera crucial d'adapter ces méthodes pour qu'elles puissent capturer efficacement les spécificités du domaine industriel, tout en assurant leur interopérabilité avec d'autres types de données et modèles utilisés dans les jumeaux numériques. Comme l'ont souligné (Qi et al., 2021), l'intégration de connaissances spécifiques au domaine dans les représentations vectorielles reste un défi important. De plus, l'interprétabilité de ces représentations vectorielles dans un contexte industriel, un aspect crucial pour la prise de décision et l'analyse des processus, demeure un défi significatif à relever.

Les développements futurs dans ce domaine pourraient se concentrer sur l'élaboration de techniques spécifiques au domaine industriel, capables de capturer non seulement la sémantique générale du texte, mais aussi les relations et les concepts propres aux systèmes industriels complexes. Les travaux récents de (Singla et al., 2019) sur les embeddings spécifiques au domaine pour l'Internet des Objets Industriel (IIoT) ouvrent des pistes prometteuses dans cette direction. L'intégration de ces représentations avec d'autres types de données (numériques, catégorielles, temporelles) au sein des jumeaux numériques constitue également une piste de recherche prometteuse pour améliorer l'interopérabilité et la performance globale des systèmes.

En conclusion, bien que les techniques de plongement offrent des opportunités significatives pour l'enrichissement des jumeaux numériques, leur application efficace dans ce contexte nécessitera des adaptations et des développements spécifiques. La combinaison de ces techniques avec d'autres approches d'intelligence artificielle, telles que l'apprentissage par transfert ou l'apprentissage fédéré, comme suggéré par (J. Wang et al., 2019), pourrait ouvrir de nouvelles voies pour relever les défis d'interopérabilité et d'adaptabilité dans les environnements industriels complexes.

3.2.5 Synthèse : techniques d'IA retenues pour l'analyse des données industrielles

L'exploration des différentes approches d'intelligence artificielle (IA) pour l'analyse des données industrielles, en particulier celles issues de l'Internet Industriel des Objets (IIoT), nous a permis d'identifier les techniques les plus prometteuses pour notre problématique d'enrichissement des jumeaux numériques. Cette synthèse vise à mettre en évidence les méthodes que nous retiendrons pour la suite de nos travaux, en se concentrant sur leur pertinence et leur applicabilité dans notre contexte spécifique.

1. « Deep Learning » pour l'extraction de caractéristiques

Parmi les techniques d'IA explorées, le deep learning se distingue par sa capacité à extraire automatiquement des caractéristiques pertinentes à partir de données brutes et complexes.

Cette propriété est particulièrement intéressante pour notre problématique, car elle permet de traiter efficacement les données hétérogènes et non structurées issues de l'IIoT sans nécessiter une ingénierie manuelle des caractéristiques extensive. Comme l'ont démontré (LeCun et al., 2015), les architectures de deep learning, telles que les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents (RNN), excellent dans l'apprentissage de représentations hiérarchiques à partir de grandes quantités de données.

2. « Graph Neural Networks » (GNN) pour le traitement des structures en graphe

Les « Graph Neural Networks » (GNN) émergent comme une technique particulièrement adaptée à notre contexte. Comme l'ont souligné (Scarselli et al., 2009) et (Zhou et al., 2020), les GNN permettent de traiter efficacement des données structurées sous forme de graphes, ce qui les rend particulièrement pertinents pour l'analyse des ontologies et des graphes de connaissances. Cette capacité s'aligne parfaitement avec notre objectif d'intégrer et d'exploiter des informations structurées dans les jumeaux numériques.

3. Scalabilité pour les données massives de l'IIoT

Les approches de « deep learning », y compris les GNN, ont démontré leur capacité à passer à l'échelle pour traiter de grands volumes de données. Cette caractéristique est cruciale pour notre problématique, étant donné la nature massive et continue des flux de données générés par l'IIoT. Les travaux de (W.-L. Chiang et al., 2019) sur l'optimisation des GNN pour les grands graphes illustrent les progrès réalisés dans ce domaine, rendant ces techniques applicables à notre contexte industriel.

4. Techniques de plongement pour l'intégration de données textuelles

Les techniques de plongement (« embeddings »), en particulier les word embeddings et les paragraph embeddings, offrent des moyens puissants pour intégrer des données textuelles non structurées dans nos modèles. Ces méthodes, comme Word2Vec (Mikolov, Sutskever, et al., 2013) et Doc2Vec (Le & Mikolov, 2014), permettront de représenter de manière compacte et sémantiquement riche les descriptions textuelles des composants et des processus industriels, facilitant leur intégration dans les jumeaux numériques.

Au vu de ces considérations, nous orienterons nos travaux futurs vers l'exploration approfondie et l'application des techniques de deep learning, en mettant l'accent sur les Graph Neural Networks pour le traitement des structures en graphe et les techniques de plongement pour l'intégration des données textuelles. Ces approches nous permettront de relever les défis liés à l'hétérogénéité des données, à la scalabilité et à l'intégration de connaissances structurées dans les jumeaux numériques.

Dans les chapitres suivants, nous nous concentrerons sur l'adaptation de ces techniques à notre contexte spécifique d'enrichissement des jumeaux numériques dans un environnement multi-modèle et multi-métiers. Nous explorerons notamment :

- L'utilisation des GNN pour modéliser et exploiter les relations complexes au sein des jumeaux numériques.
- L'intégration des techniques de plongement pour enrichir les jumeaux numériques avec des informations textuelles non structurées.

- Le développement de méthodes pour assurer l'interopérabilité entre les différentes représentations de données au sein des jumeaux numériques.

Ces axes de recherche nous permettront de développer une approche innovante pour l'enrichissement des jumeaux numériques, capable de tirer pleinement parti des données massives et hétérogènes générées par l'IIoT.

3.3 De l'information à la structure : construction d'ontologies

3.3.1 Rappel sur les ontologies

Les ontologies Propriétés : Elles décrivent les caractéristiques et les relations des entités, se divisant en deux catégories principales.

a) Propriétés d'objet :
   Elles modélisent les relations entre instances de classes. Par exemple :
   - "faitPartieDeProcessus" pourrait lier un "Équipement" à un "Processus de fabrication"
   - "estConnectéÀ" pourrait relier différents "Capteurs" ou "Systèmes de contrôle"
   - "produit" pourrait connecter une "Ligne de production" à un "Produit"

b) Propriétés de données :
   Elles associent des valeurs littérales aux instances. Par exemple :
   - "aPourCapacitéProduction" pour une "Ligne de production"
   - "aPourPrécision" pour un "Capteur"
   - "aPourDateDeMiseEnService" pour un équipement spécifique

Comme le soulignent (Guarino et al., 2009), ces composants fondamentaux permettent de créer une représentation riche et nuancée du domaine, facilitant l'intégration et l'exploitation des connaissances dans des systèmes complexes tels que les jumeaux numériques.

3.3.1.1 Modèle des triplets

Au cœur de la représentation des connaissances dans les ontologies se trouve le modèle des triplets sujet-prédicat-objet. Cette structure simple mais puissante permet d'exprimer des déclarations formelles sur le domaine modélisé (Allemang et al., 2020) :

- Sujet : L'entité décrite (une classe ou une instance)
- Prédicat : La propriété ou la relation
- Objet : La valeur ou l'entité liée

Par exemple, dans le contexte industriel :

- (Robot_A1, aPourPrécision, "0.01mm")
- (Ligne_Production_01, estSituéeDans, "Usine_Paris")
- (Capteur, mesure, Température)

Ce modèle offre une flexibilité importante, permettant de représenter aussi bien des faits simples que des relations complexes entre entités. Comme l'ont démontré (Bizer et al., 2009) dans leurs travaux sur le Web sémantique, cette approche facilite l'intégration et l'interrogation de données provenant de sources hétérogènes, un aspect crucial pour notre problématique d'enrichissement.

3.3.1.2 Logiques de description

Les logiques de description fournissent un cadre formel rigoureux pour représenter et raisonner sur les connaissances ontologiques (Baader, 2003). Elles offrent un ensemble riche de constructeurs et d'opérateurs permettant d'exprimer des concepts et des relations complexes.

3.3.1.2.1 Constructeurs de base

Les constructeurs logiques fondamentaux incluent :

- ⊓ (and) : Intersection
- ⊔ (or) : Union
- ¬ (not) : Négation

Ces opérateurs permettent de combiner et de raffiner les concepts. Par exemple :

- (Robot) ⊓ ¬(EnMaintenance)

Cette expression désigne l'ensemble des robots qui ne sont pas en maintenance, c'est-à-dire les robots opérationnels.

3.3.1.2.2 Quantificateurs et restrictions

Les logiques de description offrent également des quantificateurs et des restrictions pour exprimer des relations plus nuancées (Horrocks et al., 2006):

- ∃ (some) : Quantificateur existentiel
- ∀ (only) : Quantificateur universel
- ≥ (min) : Restriction de cardinalité minimale
- ≤ (max) : Restriction de cardinalité maximale
- − (inverse) : Relation inverse

Ces constructeurs permettent d'exprimer des concepts plus complexes. Par exemple :

- ∃utiliséDans.ProcessusFabrication

Cette expression désigne tout ce qui est utilisé dans au moins un processus de fabrication.

- ≥3 équipéDe.Capteur

Cette expression désigne tout ce qui est équipé d'au moins 3 capteurs.

3.3.1.2.3 Relations entre concepts

La relation de subsomption (⊑) est fondamentale dans les ontologies, permettant d'exprimer des hiérarchies de concepts :

- RobotSoudeur ⊑ Robot

Cette relation indique qu'un robot soudeur est un type de robot.

Cette relation s'applique également aux rôles (propriétés) :

- contrôle ⊑ interagitAvec

Cette relation indique que tout ce qui contrôle quelque chose interagit nécessairement avec cette chose.

3.3.1.2.4 Propriétés des rôles

Les rôles (ou propriétés) dans les ontologies peuvent avoir diverses caractéristiques, enrichissant leur sémantique (Horrocks & Sattler, 2004) :

- Transitivité
- Fonctionnalité
- Réflexivité
- Symétrie

Ces propriétés permettent des inférences logiques supplémentaires, renforçant la puissance expressive de l'ontologie. Par exemple, la transitivité de la relation "faitPartieDeProcessus" permettrait de déduire automatiquement qu'un sous-composant d'un équipement fait également partie du processus dans lequel cet équipement est utilisé.

3.3.1.2.5 Raisonnement et inférence

L'un des avantages majeurs des logiques de description réside dans leur capacité à supporter le raisonnement automatisé, comme l'a souligné (Horrocks, 2008). Cette fonctionnalité permet non seulement d'enrichir automatiquement la base de connaissances, mais aussi de vérifier sa cohérence.

La transitivité de la relation de subsomption, en particulier, offre un mécanisme puissant pour déduire de nouvelles relations. Prenons un exemple concret dans le domaine industriel :

- Si : RobotSoudeurLaser ⊑ RobotSoudeur
- Et : RobotSoudeur ⊑ ÉquipementFabrication
- Alors on peut inférer : RobotSoudeurLaser ⊑ ÉquipementFabrication

Cette inférence automatique permet d'enrichir la base de connaissances du jumeau numérique sans intervention manuelle, assurant ainsi une représentation plus complète et cohérente de l'environnement industriel.

(Baader et al., 2017) ont démontré que ces capacités d'inférence vont au-delà de la simple déduction de relations hiérarchiques. Elles permettent également de :

1. Vérifier la consistance de l'ontologie : Détecter les contradictions potentielles dans la modélisation des processus industriels.
2. Classifier automatiquement les instances : Attribuer automatiquement des équipements ou des processus à leurs catégories appropriées.
3. Répondre à des requêtes complexes : Extraire des informations implicites cruciales pour la prise de décision dans l'environnement industriel.

(Sirin et al., 2007) ont souligné l'importance des raisonneurs OWL, tels que Pellet, dans l'exploitation de ces capacités d'inférence. Ces outils permettent d'automatiser le processus de raisonnement, facilitant ainsi la maintenance et l'évolution des ontologies complexes.

3.3.1.3 Variantes des logiques de description

Les logiques de description englobent une famille de formalismes, chacun offrant un équilibre spécifique entre expressivité et complexité computationnelle. Cette diversité permet de choisir le formalisme le plus adapté aux besoins spécifiques du domaine modélisé et aux contraintes de performance du système. Parmi les variantes les plus significatives, on peut citer :

1. ALC (Attributive Language with Complements) : Introduite par (Schmidt-Schauß & Smolka, 1991), cette logique de base offre un point de départ solide pour la modélisation ontologique. Elle inclut les constructeurs fondamentaux tels que la conjonction, la disjonction, la négation, et les quantificateurs existentiels et universels.
2. SHOIN(D) : Cette variante plus expressive, décrite par (Horrocks et al., 2003), forme la base du langage OWL-DL (Web Ontology Language - Description Logic). Elle ajoute des fonctionnalités cruciales pour la modélisation complexe, telles que :
   - Hiérarchies de rôles (S)
   - Restrictions de cardinalité (N)
   - Propriétés inverses (I)
   - Nominaux (O)
   - Types de données (D)
3. SROIQ(D) : Une extension encore plus puissante, proposée par (Horrocks et al., 2006), qui ajoute :
   - Chaînes de rôles (R)
   - Axiomes complexes sur les rôles (Q)

Ces variantes plus expressives permettent de modéliser des aspects complexes des systèmes industriels, tels que les contraintes de production, les interdépendances entre équipements, ou les flux de processus sophistiqués.

Le choix de la logique appropriée dépend de plusieurs facteurs, comme l'ont souligné (Krötzsch et al., 2013) :

1. Complexité du domaine : des domaines industriels hautement complexes peuvent nécessiter des logiques plus expressives comme SROIQ(D).
2. Besoins en inférence : Certaines tâches de raisonnement peuvent être indécidables ou computationnellement coûteuses dans des logiques très expressives.
3. Performance requise : les logiques moins expressives comme ALC offrent généralement de meilleures performances de raisonnement.
4. Outils disponibles : la disponibilité de raisonneurs efficaces pour une logique donnée peut influencer le choix.

3.3.1.4 Intérêts dans le contexte industriel

Dans le cadre de notre problématique d'enrichissement des jumeaux numériques et d'amélioration de l'interopérabilité, les ontologies et les logiques de description offrent plusieurs avantages cruciaux :

1. Modélisation formelle : elles permettent de représenter de manière explicite et non ambiguë les concepts, relations et contraintes du domaine industriel. Comme l'ont démontré (Lemaignan et al., 2006) dans leur travail sur MASON (Manufacturing's Semantics ONtology), cette formalisation fournit une base solide pour la construction de modèles précis et cohérents.
2. Raisonnement automatisé : les mécanismes d'inférence facilitent la découverte de nouvelles connaissances et la vérification de la cohérence des modèles. (Panetto et al., 2012) ont souligné l'importance de ces capacités pour maintenir l'intégrité des modèles dans des environnements industriels dynamiques.

3. Interopérabilité sémantique : en fournissant un vocabulaire partagé et une structure conceptuelle commune, les ontologies favorisent l'intégration et l'échange de données entre différents systèmes et domaines d'expertise. (Chungoora et al., 2013) ont démontré l'efficacité de cette approche pour améliorer l'interopérabilité dans les systèmes de fabrication.
4. Flexibilité et extensibilité : les ontologies peuvent être facilement étendues pour intégrer de nouveaux concepts ou relations, s'adaptant ainsi à l'évolution rapide des technologies et des besoins industriels. Cette caractéristique est particulièrement importante dans le contexte de l'Industrie 4.0, comme l'ont souligné (Thoben et al., 2017).
5. Support pour l'analyse des données : en structurant les connaissances du domaine, les ontologies peuvent guider l'interprétation et l'analyse des données issues de l'IoT et d'autres sources. (Tao, Zhang, Liu, et al., 2019) ont montré comment cette approche peut enrichir les capacités analytiques des jumeaux numériques, permettant une prise de décision plus informée et plus rapide.

L'utilisation des ontologies et des logiques de description dans le contexte des jumeaux numériques ouvre des perspectives prometteuses pour améliorer la représentation, l'intégration et l'exploitation des connaissances industrielles complexes. Comme l'ont démontré (Uhlemann et al., 2017), elles fournissent un cadre puissant pour relever les défis d'interopérabilité et d'enrichissement des modèles dans des environnements multi-modèles et multi-métiers, contribuant ainsi à la réalisation de systèmes industriels plus intelligents, adaptatifs et efficaces.

3.3.2 Approches de développement d'ontologies sans ontologie initiale

Le développement d'ontologies en l'absence d'une ontologie initiale, également appelé découverte complète, représente un défi majeur dans le domaine de l'ingénierie des connaissances. Cette section examine les fondements théoriques, les langages de représentation, les méthodologies et les bonnes pratiques associés à cette démarche complexe mais cruciale.

3.3.2.1 Fondements théoriques et langages de représentation

Les ontologies s'appuient sur des bases logiques rigoureuses, principalement les logiques de description. Comme l'ont souligné (Baader, 2003), ces logiques offrent un cadre formel pour la représentation des connaissances tout en préservant la décidabilité, un avantage significatif par rapport à la logique du premier ordre. Cette caractéristique est particulièrement importante dans le contexte des systèmes industriels complexes, où la capacité à raisonner efficacement sur les connaissances représentées est cruciale.

Dans la pratique, deux langages de représentation se sont imposés comme standards :

1. RDF Schema (RDFS) : introduit par le W3C (Lassila & Swick, 1999), RDFS fournit un vocabulaire de base pour décrire les propriétés et les classes des ressources RDF. (Antoniou & Van Harmelen, 2004) ont démontré son utilité pour la modélisation de connaissances simples dans divers domaines, y compris l'industrie.
2. Web Ontology Language (OWL) : évolution de RDFS, OWL offre une expressivité accrue. Sa dernière version, OWL 2, présentée par (Hitzler et al., 2012), permet une modélisation plus fine des connaissances complexes. (Horrocks et al., 2003) ont mis en évidence la puissance d'OWL pour représenter des systèmes sophistiqués.

3.3.2.2 Approches philosophiques et méthodologiques

Le développement d'ontologies s'appuie sur des approches philosophiques et méthodologiques spécifiques :

1. Réalisme ontologique : cette approche, défendue par (B. Smith & Ceusters, 2010), postule l'existence d'entités objectives indépendantes de notre perception. Ils soulignent que dans le contexte des ontologies, le réalisme ontologique se réfère à une méthodologie qui suppose que les termes utilisés dans l'ontologie se réfèrent à des entités réelles ou des "universaux" dans le monde. Cette approche est particulièrement pertinente pour la modélisation de systèmes industriels physiques.
2. Basic Formal Ontology (BFO) : proposée par (Arp et al., 2015), la BFO sert d'ontologie de haut niveau, offrant un cadre neutre pour la représentation des distinctions de domaine. Comme l'ont démontré (Keet et al., 2015), l'utilisation de la BFO peut faciliter l'intégration de différentes ontologies de domaine.
3. Architecture multi-niveaux : cette approche, recommandée par (Guarino, 1998), structure les ontologies en niveaux hiérarchiques, partant d'une ontologie de haut niveau vers des ontologies de domaine plus spécifiques. (Borgo & Leitão, 2007) ont appliqué avec succès cette approche dans le domaine de la fabrication, démontrant son efficacité pour gérer la complexité des systèmes industriels.

3.3.2.3 Méthodologies de développement

Plusieurs méthodologies structurées ont été proposées pour guider le processus de développement d'ontologies c

1. METHONTOLOGY : développée par (Fernández-López et al., 1997), cette méthodologie s'inspire des processus de développement logiciel et d'ingénierie des connaissances. Elle propose un cycle de vie basé sur des prototypes évolutifs et des techniques spécifiques pour chaque étape du développement. (Corcho et al., 2003) ont démontré son efficacité dans le développement d'ontologies pour des domaines techniques complexes.
2. On-To-Knowledge : (Staab et al., 2001) ont proposé cette méthodologie axée sur l'application, qui met l'accent sur l'analyse des scénarios d'utilisation et l'évaluation continue. (Sure et al., 2004) ont appliqué cette approche avec succès dans le développement d'ontologies pour la gestion des connaissances en entreprise.
3. NeOn Methodology : (Suárez-Figueroa et al., 2012) ont développé cette approche flexible, qui prend en compte divers scénarios de développement, y compris la réutilisation et la réingénierie d'ontologies existantes. (Espinoza-Arias et al., 2019) ont démontré son efficacité dans le développement d'ontologies pour l'Internet des Objets.

3.3.2.4 Bonnes pratiques et défis

Le développement d'ontologies à partir de zéro présente plusieurs défis et nécessite l'adoption de bonnes pratiques :

1. Modularité : (Parent & Spaccapietra, 2009) soulignent que la conception d'ontologies modulaires facilite la maintenance, la réutilisation et l'interopérabilité. Cette approche est particulièrement recommandée pour les grands domaines industriels, où la combinaison de plusieurs petites ontologies peut couvrir l'ensemble du domaine de

manière plus efficace. (Keet, 2018) a fourni des lignes directrices détaillées pour la modularisation des ontologies dans des domaines complexes.

2. Réutilisation : (Simperl, 2009) met en avant l'importance de la réutilisation d'ontologies existantes pour améliorer l'interopérabilité et réduire les efforts de développement. Dans le contexte industriel, (Panetto et al., 2012) ont démontré comment la réutilisation d'ontologies peut faciliter l'intégration de systèmes hétérogènes.
3. Évaluation continue : (Vrandečić, 2009) insiste sur la nécessité d'une évaluation rigoureuse et continue des ontologies tout au long de leur cycle de vie. (Poveda-Villalón et al., 2014) ont proposé des méthodes d'évaluation pour les ontologies, mettant l'accent sur la cohérence et la complétude.
4. Gestion de la complexité : pour les domaines vastes et complexes, la gestion de la portée et de la granularité de l'ontologie reste un défi majeur, comme le soulignent (Hepp & Roman, 2007). Les ontologies volumineuses et diversifiées présentent des inconvénients importants, tels que la difficulté à trouver des informations à travers plusieurs ontologies en raison de la duplication du contenu et des problèmes de modification et de gouvernance. (Obrst et al., 2012) ont proposé des stratégies pour gérer cette complexité dans le contexte des systèmes de cybersécurité.
5. Interopérabilité : pour assurer une meilleure interopérabilité, une approche architecturale multi-niveaux est suggérée, qui commence par une seule ontologie de haut niveau et s'étend aux ontologies de niveau intermédiaire et inférieur. (Chungoora et al., 2013) ont démontré l'efficacité de cette approche pour améliorer l'interopérabilité dans les systèmes de fabrication.

3.3.2.5 Synthèse

En conclusion, le développement d'ontologies en mode découverte est un processus complexe qui nécessite une approche méthodique, s'appuyant sur des fondements théoriques solides et des bonnes pratiques établies. La combinaison judicieuse de ces éléments est essentielle pour créer des ontologies robustes, interopérables et adaptées aux besoins spécifiques des domaines d'application industriels. Le développement de petites ontologies modulaires et interchangeables, avec des lignes directrices pour circonscrire la portée horizontale du contenu, apparaît comme une stratégie prometteuse pour relever les défis de l'ingénierie ontologique moderne dans le contexte de l'Industrie 4.0 et des jumeaux numériques.

Dans le cadre de notre recherche sur l'enrichissement des jumeaux numériques et l'amélioration de l'interopérabilité dans un environnement multi-modèle et multi-métiers, ces approches offrent des avantages cruciaux. Elles permettent une représentation formelle des connaissances complexes, assurent la flexibilité nécessaire pour s'adapter à l'évolution rapide des technologies industrielles, et favorisent l'interopérabilité sémantique entre différents systèmes et domaines d'expertise. De plus, le support pour le raisonnement automatisé qu'elles offrent est essentiel pour enrichir dynamiquement les jumeaux numériques et gérer la complexité inhérente aux systèmes industriels modernes.

L'adoption de ces méthodologies de développement d'ontologies nous fournira une base solide pour intégrer efficacement les données provenant de sources externes, notamment de l'Internet des Objets Industriel (IIoT), tout en assurant la cohérence et l'interopérabilité des modèles. Cette approche structurée et flexible sera cruciale pour réaliser notre objectif d'enrichissement des

jumeaux numériques et d'amélioration de la performance globale des systèmes industriels dans un environnement en constante évolution.

3.3.3 Construction d'ontologies en présence d'ontologies existante

La construction d'ontologies en présence d'ontologies existantes représente une approche stratégique pour le développement de représentations de connaissances cohérentes et interopérables. Cette méthode s'appuie sur des fondements théoriques solides et des pratiques éprouvées, offrant des avantages significatifs en termes d'efficacité et de qualité des ontologies résultantes.

3.3.3.1 Approches de développement

Le développement d'ontologies peut suivre deux trajectoires principales : top-down et bottom-up. Comme l'ont souligné (Uschold & Gruninger, 1996), l'approche top-down débute avec les concepts généraux pour progresser vers les spécifiques, tandis que l'approche bottom-up suit le chemin inverse. Le choix entre ces approches dépend souvent du contexte et des objectifs du projet ontologique.

L'approche top-down, particulièrement adaptée lorsqu'une compréhension structurée du domaine existe déjà, permet d'assurer une cohérence globale de l'ontologie. Elle facilite l'intégration de concepts importants dès les premières étapes du développement. En revanche, l'approche bottom-up, comme l'ont noté (Fernández-López & Gómez-Pérez, 2002), est privilégiée lorsque la compréhension préalable du domaine est limitée ou lorsque l'ontologie doit être construite à partir de ressources existantes. Cette méthode favorise l'exploration et la découverte de connaissances spécifiques au domaine.

Dans la pratique, une approche hybride combinant les méthodologies top-down et bottom-up est souvent adoptée. (Pinto & Martins, 2004) ont démontré que cette approche mixte permet de bénéficier des avantages des deux méthodes, assurant à la fois une cohérence globale et une exploration approfondie des connaissances spécifiques au domaine.

3.3.3.2 Types d'ontologies

La classification des ontologies proposée par (Guarino, 1998) distingue deux types principaux : les ontologies de haut niveau et les ontologies de domaine. Les ontologies de haut niveau fournissent un cadre de modélisation général applicable à divers domaines, facilitant ainsi l'interopérabilité entre ontologies. Les ontologies de domaine, quant à elles, se concentrent sur la modélisation des connaissances spécifiques à un domaine particulier.

(Mascardi et al., 2007) ont affiné cette classification en proposant trois catégories d'ontologies selon leur niveau d'abstraction :

1. Les ontologies de fondation, telles que la Basic Formal Ontology (BFO) et la Suggested Upper Merged Ontology (SUMO), qui s'appliquent à des concepts et relations générales utilisables dans des domaines hétérogènes.
2. Les ontologies de référence, comme la Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), qui sont spécifiques à certains domaines mais réutilisables au sein de ces domaines.
3. Les ontologies applicatives, conçues pour des tâches uniques et souvent construites en s'appuyant sur des ontologies de haut niveau ou de référence.

3.3.3.3 Bonnes pratiques de développement

Pour garantir la pérennité et l'efficacité des ontologies, (B. Smith et al., 2007) ont souligné l'importance de créer des termes et des définitions compréhensibles et valides. Cela implique plusieurs étapes cruciales :

- Identification et évaluation des ontologies existantes
- Création d'une liste préliminaire de termes
- Recherche d'un consensus maximal avec les experts du domaine
- Prévention des erreurs ontologiques courantes

Les auteurs insistent sur l'importance d'organiser les ontologies dans une hiérarchie avec des liens d'hérédité uniques entre les nœuds. Les définitions doivent être basées sur des caractéristiques essentielles, éviter la circularité et assurer l'intelligibilité.

(Reiter, 1978) a mis en avant l'importance de l'hypothèse du monde ouvert dans le développement d'ontologies. Cette approche, qui stipule que la véracité d'une proposition peut être vraie indépendamment de sa reconnaissance comme telle, se distingue de l'hypothèse du monde fermé et offre une flexibilité cruciale dans la représentation des connaissances.

3.3.3.4 Approches d'alignement et d'intégration

La fusion ou la combinaison d'ontologies existantes peut s'avérer plus avantageuse que la création d'une nouvelle ontologie à partir de zéro. (N. Noy & Musen, 2000a) ont démontré que cette approche permet de capitaliser sur les connaissances déjà représentées, réduisant ainsi le temps et les efforts de développement tout en favorisant l'interopérabilité sémantique.

Deux approches particulièrement intéressantes pour l'alignement et l'intégration d'ontologies sont l'utilisation de "backbone ontologies" et d'ontologies pivot

1. Les "backbone ontologies", ou ontologies d'épine dorsale, fournissent une structure commune pour intégrer et relier plusieurs ontologies de domaine spécifique. (Duong et al., 2010) ont souligné leur rôle crucial dans la facilitation de l'interopérabilité sémantique entre différents domaines de connaissance. Des exemples notables incluent la Basic Formal Ontology (BFO), DOLCE, et SUMO, chacune ayant ses propres principes de conception et expressivité adaptés à des besoins spécifiques (Mascardi et al., 2007).
2. Les ontologies pivot, comme l'ont expliqué (Euzenat & Shvaiko, 2013), agissent comme intermédiaires entre les ontologies de domaines spécifiques pour faciliter leur alignement et leur interopérabilité. Cette approche est particulièrement utile pour aligner des ontologies de domaines très différents ou présentant des différences significatives dans leur structure et terminologie (N. F. Noy, 2004).

3.3.3.5 Synthèse

En conclusion, la construction d'ontologies en s'appuyant sur des ontologies existantes offre des avantages considérables pour l'enrichissement des jumeaux numériques dans un contexte industriel complexe. Cette approche permet une réutilisation efficace des connaissances déjà formalisées, favorise l'interopérabilité entre différents systèmes et métiers, et réduit significativement le temps et les ressources nécessaires au développement ontologique. Dans le cadre de l'Industrie 4.0 et des environnements multi-modèles, l'utilisation de "backbone

ontologies" et d'ontologies pivot s'avère particulièrement pertinente pour aligner et intégrer des ontologies provenant de divers domaines industriels.

L'intégration de techniques d'intelligence artificielle (IA) dans ce processus ouvre de nouvelles perspectives prometteuses. Les algorithmes d'apprentissage automatique, notamment les techniques de traitement du langage naturel et d'apprentissage profond, peuvent être utilisés pour automatiser certains aspects de la construction et de l'alignement des ontologies. Par exemple, les modèles de plongement lexical (word embeddings) peuvent aider à identifier des relations sémantiques entre concepts issus de différentes ontologies, tandis que les réseaux de neurones en graphes (« Graph Neural Networks ») peuvent être exploités pour analyser et aligner les structures ontologiques complexes. Ces approches basées sur l'IA peuvent significativement accélérer le processus d'intégration des connaissances et améliorer la précision des alignements ontologiques, tout en s'adaptant dynamiquement aux évolutions des systèmes industriels.

Cependant, l'application de ces approches dans un environnement industriel complexe nécessite une attention particulière à la cohérence sémantique et à la gestion des conflits potentiels entre les différentes ontologies. Le choix de l'approche la plus appropriée dépendra des caractéristiques spécifiques du projet d'enrichissement des jumeaux numériques, des domaines industriels concernés, et des objectifs d'interopérabilité visés. En combinant judicieusement les méthodologies de construction d'ontologies avec les techniques d'IA, les industries peuvent améliorer significativement la capacité de leurs jumeaux numériques à intégrer, analyser et exploiter des données provenant de sources diverses, conduisant ainsi à une meilleure prise de décision, une optimisation des processus et une augmentation de la performance globale des systèmes industriels.

3.4 De la structure à l'intégration : alignement et intégration d'ontologies

3.4.1 Défis de l'alignement d'ontologies

L'alignement d'ontologies joue un rôle crucial dans l'établissement de l'interopérabilité sémantique entre différents domaines ou systèmes, particulièrement dans le contexte de l'Industrie 4.0. Bien que chaque domaine bénéficie d'une ontologie dédiée, cette spécificité peut paradoxalement limiter l'interopérabilité aux domaines pour lesquels ces ontologies ont été conçues, allant ainsi à l'encontre du principe fondamental d'interopérabilité de l'Industrie 4.0 (Euzenat & Shvaiko, 2013).

L'alignement manuel d'ontologies est une approche visant à résoudre ce problème. Cette méthode implique l'intervention directe d'experts du domaine pour identifier et établir des correspondances entre les éléments de différentes ontologies. (N. Noy & Musen, 2000b) ont décrit le processus d'alignement manuel comme comprenant généralement plusieurs étapes clés :

1. Examen détaillé des ontologies à aligner, nécessitant une compréhension approfondie de leur structure, concepts, relations et axiomes.
2. Identification des correspondances potentielles entre les éléments des ontologies, incluant des équivalences exactes, des relations de subsomption, ou des correspondances plus complexes.

3. Formalisation des alignements identifiés, potentiellement en utilisant des langages spécifiques comme EDOAL (Expressive and Declarative Ontology Alignment Language) (David et al., 2011).

(Euzenat & Shvaiko, 2013) soulignent que l'alignement manuel présente l'avantage de produire des résultats de haute qualité, bénéficiant de l'expertise humaine pour capturer des nuances sémantiques subtiles. Cependant, cette approche comporte plusieurs limitations significatives :

1. Temps et ressources : (Otero-Cerdeira et al., 2015) ont mis en évidence que l'alignement manuel est extrêmement chronophage, en particulier pour des ontologies larges et complexes. Le temps nécessaire augmente de manière quadratique avec le nombre d'ontologies à aligner, car chaque paire d'ontologies doit être examinée séparément. En effet, le nombre de paires dans un ensemble de n ontologies est $\frac{n(n-1)}{2}$, illustrant la croissance rapide de la complexité avec l'augmentation du nombre d'ontologies.
2. Erreurs et incohérences : (Cheatham et al., 2015) ont souligné que l'alignement manuel est sujet à l'erreur humaine et peut souffrir d'incohérences, en particulier lorsque plusieurs experts sont impliqués.
3. Maintenance : (Euzenat & Zimmermann, 2007) ont noté que chaque mise à jour d'une ontologie source nécessite une révision des alignements, ce qui peut devenir un processus fastidieux et coûteux à long terme.
4. Subjectivité : 'expertise individuelle peut influencer les résultats, conduisant potentiellement à des alignements biaisés ou incomplets, comme l'ont souligné (Dragisic et al., 2016).

Malgré ces limitations, l'alignement manuel reste une approche importante, souvent utilisée en combinaison avec des méthodes semi-automatiques ou automatiques pour valider et affiner les résultats (Paulheim et al., 2013). Dans le contexte de l'amélioration de l'interopérabilité dans des environnements multi-modèles et multi-métiers, l'alignement manuel peut jouer un rôle crucial dans l'établissement initial de correspondances sémantiques précises entre différents domaines d'expertise.

En conclusion, bien que l'alignement manuel d'ontologies offre certains avantages en termes de précision et de capture des nuances sémantiques, il présente des limitations significatives en termes de temps, de ressources et d'évolutivité. Ces contraintes sont particulièrement problématiques dans le contexte de l'Industrie 4.0 et des jumeaux numériques, où la rapidité, l'adaptabilité et l'automatisation sont essentielles.

Par conséquent, dans la suite de notre étude, nous nous concentrerons sur les méthodes d'alignement automatique et semi-automatique d'ontologies, qui sont plus cohérentes avec notre approche globale d'enrichissement des jumeaux numériques. Ces méthodes, s'appuyant sur des techniques d'intelligence artificielle et d'apprentissage automatique, offrent des solutions plus adaptées aux défis de l'interopérabilité sémantique dans des environnements multi-modèles et multi-métiers complexes et dynamiques.

3.4.2 Méthodes d'alignement d'ontologies

L'alignement d'ontologies joue un rôle crucial dans l'établissement de l'interopérabilité sémantique entre différents systèmes ou domaines de connaissances. Cette section examine les principales méthodes d'alignement, en commençant par clarifier les termes clés et en présentant une classification des approches, avant de détailler les techniques spécifiques et les défis actuels.

3.4.2.1 Définitions fondamentales

(Euzenat & Shvaiko, 2013) ont établi un cadre conceptuel essentiel pour comprendre les processus liés à l'alignement d'ontologies, fournissant des définitions précises pour plusieurs termes clés :

- Alignement d'ontologies (ontology alignment) : il s'agit d'un ensemble de correspondances entre plusieurs ontologies. Ces correspondances peuvent être simples, établissant des liens entre des entités individuelles, ou complexes, reliant des groupes d'entités ou des sous-structures ontologiques.
  Par exemple, dans un contexte industriel, un alignement pourrait lier le concept "Machine" d'une ontologie de production à "Équipement" dans une ontologie de maintenance.
- Correspondance d'ontologies (ontology matching) : ce terme englobe deux types de relations :
  - Les prédicats de similarité (matching) qui quantifient le degré de ressemblance entre des entités.
  - Les axiomes logiques (mapping) qui établissent des relations formelles entre les entités.

  Par exemple, un matching pourrait indiquer que "Produit" et "Article" ont une similarité de 0,8, tandis qu'un mapping pourrait affirmer que "Produit" est équivalent à "Article".

Il est important de noter que dans la littérature, les termes "matching" et "mapping" sont souvent utilisés de manière interchangeable, ce qui peut parfois prêter à confusion.

- Fusion d'ontologies (ontology merging) : ce processus vise à combiner plusieurs ontologies en une seule, intégrant les connaissances de toutes les ontologies fusionnées. Les fusions avancées peuvent aller au-delà de la simple union des ontologies en incluant des axiomes supplémentaires qui définissent explicitement les relations entre les concepts des ontologies d'origine. Par exemple, fusionner une ontologie de production et une ontologie de qualité pourrait nécessiter l'ajout d'axiomes reliant les processus de fabrication aux critères de qualité correspondants.
- Traduction d'ontologies (ontology translation) : ce processus implique la modification de la sémantique sous-jacente d'une connaissance pour passer de la sémantique d'une ontologie source à celle d'une ontologie cible. L'objectif est de préserver intégralement le sens original sans perte d'information. Par exemple, traduire une ontologie de maintenance préventive développée pour l'industrie automobile vers une ontologie utilisée dans l'aéronautique nécessiterait une

adaptation fine des concepts et relations pour refléter les spécificités de chaque domaine.

Ces définitions fournissent une base solide pour comprendre et analyser les différentes approches et techniques utilisées dans l'alignement d'ontologies.

3.4.2.2 Classification des approches d'alignement

La classification des approches d'alignement a évolué pour refléter la complexité croissante des techniques utilisées. (Rahm & Bernstein, 2001) ont proposé une classification initiale distinguant principalement les approches basées sur les éléments individuels et celles basées sur la structure des ontologies Figure 28.

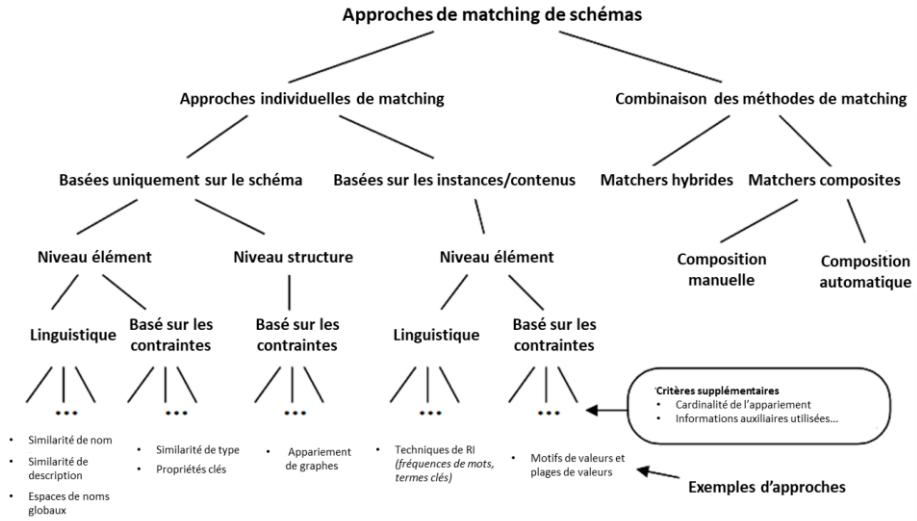


Figure 28 Classification initiale des approches d'alignements proposée par (Rahm & Bernstein, 2001)

(Euzenat & Shvaiko, 2013) ont affiné la classification initiale des approches d'alignement d'ontologies en introduisant des catégories plus détaillées montrées dans la Figure 29.

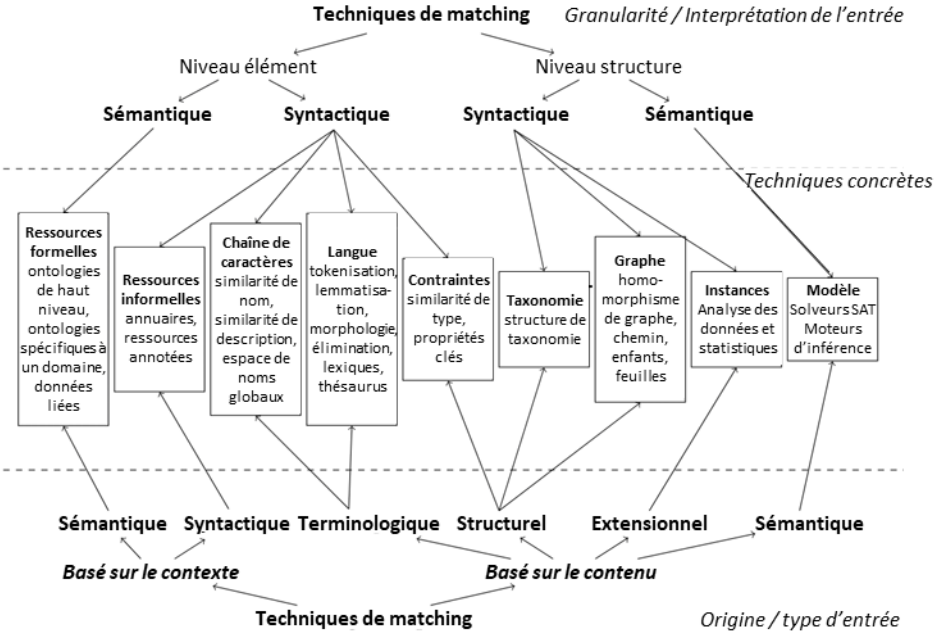


Figure 29 Evolution de la classification précédente pour les ontologies par (Euzenat & Shvaiko, 2013)

Cette classification multidimensionnelle permet une compréhension plus nuancée des différentes techniques d'alignement et de leurs applications potentielles :

1. Niveau de granularité :
    - Approches au niveau des éléments (Element-level) :
      Ces méthodes considèrent chaque élément de l'ontologie séparément, ignorant les relations entre les éléments au sein d'une même ontologie. Elles se concentrent sur les propriétés intrinsèques des entités, telles que leurs noms, descriptions ou types de données (Euzenat & Shvaiko, 2013). Par exemple, la comparaison de chaînes de caractères ou l'utilisation de techniques de traitement du langage naturel pour analyser les labels des entités (Cheatham & Hitzler, 2013).
    - Approches au niveau de la structure (Structure-level) :
      Ces techniques prennent en compte les relations entre les éléments de l'ontologie et la façon dont ces relations sont organisées. Elles exploitent la structure hiérarchique ou les relations sémantiques entre les concepts pour identifier des correspondances (Euzenat & Shvaiko, 2013). Par exemple, l'utilisation de techniques de similarité de graphes ou l'analyse des relations taxonomiques (Melnik et al., 2002).
2. Type d'entrée :
    - Approches basées sur le contenu (Content-based) :
      Ces méthodes utilisent uniquement les informations internes aux ontologies à aligner, telles que les labels, les commentaires, les propriétés et la structure des ontologies (Euzenat & Shvaiko, 2013). Elles sont particulièrement utiles lorsque les ontologies sont riches en métadonnées et en relations internes (Faria et al., 2013).
    - Approches basées sur le contexte (Context-based) :
      Ces techniques exploitent des ressources externes, formelles ou informelles, pour enrichir le processus d'alignement. Elles peuvent utiliser des ressources linguistiques comme WordNet, des ontologies de domaine existantes, ou même des données d'instance pour améliorer la précision de l'alignement (Euzenat & Shvaiko, 2013); (Otero-Cerdeira et al., 2015).
3. Technique d'alignement :
    - Approches syntactiques :
      Ces méthodes se basent sur des techniques d'alignement syntactique, comme la comparaison de chaînes de caractères, l'utilisation de mesures de similarité lexicale, ou l'alignement de graphes basé sur la structure (Euzenat & Shvaiko, 2013). Elles sont généralement plus rapides et plus simples à mettre en œuvre, mais peuvent manquer de précision pour des alignements complexes (Cheatham & Hitzler, 2013).
    - Approches sémantiques :
      Ces techniques utilisent des méthodes spécifiques aux ontologies, comme la réutilisation d'ontologies existantes, l'utilisation de raisonneurs sémantiques, ou l'exploitation de relations sémantiques définies formellement (Euzenat & Shvaiko, 2013). Elles peuvent produire des alignements plus précis et plus riches

sémantiquement, mais sont souvent plus complexes et plus coûteuses en termes de calcul (Jiménez-Ruiz & Cuenca Grau, 2011).

Cette classification multidimensionnelle offre un cadre complet pour comprendre et catégoriser les diverses approches d'alignement d'ontologies. Elle permet aux chercheurs et aux praticiens de choisir les méthodes les plus appropriées en fonction des caractéristiques spécifiques des ontologies à aligner et des objectifs de l'alignement.

3.4.2.3 Principales méthodes d'alignement

1. Alignement terminologique :

Cette approche se concentre sur la similarité lexicale entre les noms des entités dans différentes ontologies. Elle peut utiliser des techniques simples comme la comparaison de chaînes de caractères (par exemple, la distance de Levenshtein) ou des méthodes plus avancées comme l'analyse linguistique. Les techniques d'analyse linguistique peuvent inclure la tokenisation, la lemmatisation, et l'utilisation de thésaurus ou de bases de données lexicales comme WordNet pour identifier des synonymes et des relations sémantiques (Euzenat & Shvaiko, 2013). Par exemple, (Cheatham & Hitzler, 2013) ont démontré l'efficacité des techniques de traitement du langage naturel pour améliorer la précision de l'alignement terminologique.

2. Alignement structurel :

Cette méthode exploite la structure hiérarchique des ontologies pour identifier des correspondances. Elle compare les relations entre les concepts, leurs propriétés, et leur position dans la hiérarchie ontologique. Les techniques d'alignement structurel peuvent inclure la comparaison des voisinages des concepts, l'analyse des chemins dans la hiérarchie, ou l'utilisation d'algorithmes de similarité de graphes (N. Noy & Musen, 2001). (Melnik et al., 2002) ont proposé l'algorithme Similarity Flooding, qui propage la similarité à travers la structure des ontologies, démontrant l'efficacité de l'approche structurelle.

3. Alignement sémantique :

Cette approche utilise la sémantique formelle des ontologies, souvent exprimée en logique de description, pour identifier des correspondances basées sur le sens plutôt que sur la forme. Elle peut impliquer l'utilisation de raisonneurs logiques pour inférer des relations implicites entre les concepts. Les techniques d'alignement sémantique peuvent inclure la vérification de la satisfiabilité des correspondances, la propagation des contraintes logiques, et l'utilisation de patterns d'axiomes (Jiménez-Ruiz & Cuenca Grau, 2011). (Stoilos et al., 2005) ont développé une mesure de similarité sémantique basée sur la logique de description, démontrant l'importance de l'alignement sémantique pour capturer des correspondances complexes.

4. Alignement basé sur les instances :

Cette méthode compare les instances (individus) des concepts dans différentes ontologies pour établir des correspondances. Elle est particulièrement utile lorsque les ontologies partagent des instances communes ou lorsque les instances ont des

attributs similaires. Les techniques d'alignement basé sur les instances peuvent inclure la comparaison des valeurs de propriétés, l'analyse de la distribution des instances, et l'utilisation de méthodes d'apprentissage automatique pour classifier les instances (Isaac et al., 2007). (Thor et al., 2007) ont proposé une approche d'apprentissage actif pour l'alignement basé sur les instances, démontrant son efficacité dans des scénarios avec peu de données d'entraînement.

5. Alignement hybride :

Les approches hybrides combinent plusieurs des méthodes mentionnées ci-dessus pour tirer parti de leurs forces respectives et compenser leurs faiblesses. Ces approches peuvent utiliser des stratégies de pondération ou de vote pour combiner les résultats de différentes méthodes d'alignement, ou employer des techniques d'apprentissage automatique pour optimiser la combinaison des différentes approches (Cruz et al., 2009). Par exemple, (Doan et al., 2004) ont développé le système GLUE, qui combine des techniques d'apprentissage automatique avec des contraintes de domaine pour réaliser un alignement hybride efficace.

3.4.2.4 Approches basées sur l'apprentissage profond

Récemment, les techniques d'apprentissage profond ont émergé comme une approche prometteuse pour l'alignement d'ontologies, en particulier pour l'appariement d'entités. (Barlaug & Gulla, 2021) ont réalisé une étude approfondie sur l'utilisation des réseaux de neurones pour l'alignement d'entités, mettant en lumière les avancées significatives dans ce domaine.

Principales caractéristiques des approches basées sur l'apprentissage profond :

1. Représentation hiérarchique puissante :
   La principale contribution de ces méthodes est leur capacité à apprendre des représentations hiérarchiques complexes à partir du texte, permettant une compréhension plus nuancée des entités et de leurs relations. Par exemple, les modèles de langage pré-entraînés comme BERT (Devlin et al., 2019) peuvent capturer des contextes sémantiques riches, améliorant ainsi la précision de l'alignement.
2. Flexibilité dans le processus d'alignement :
   Ces méthodes peuvent être appliquées à différentes étapes du processus traditionnel d'appariement d'entités, offrant une grande flexibilité dans leur utilisation. (Kolyvakis et al., 2018) ont démontré comment les techniques d'apprentissage profond peuvent être utilisées pour générer des représentations d'entités, calculer des similarités, et même effectuer directement l'alignement.
3. Gestion de la qualité des données :
   Les approches d'apprentissage profond se sont montrées particulièrement efficaces pour gérer les problèmes de qualité des données, un défi persistant dans l'alignement d'entités. Les architectures comme les autoencodeurs débruitants (Vincent et al., 2010) peuvent aider à traiter les données bruitées ou incomplètes, améliorant ainsi la robustesse de l'alignement.

4. Adaptation aux différentes étapes :
   Les techniques de réseaux neuronaux peuvent être adaptées pour aborder spécifiquement différentes étapes du processus d'alignement, utilisant des architectures variées selon les besoins. Par exemple, les réseaux siamois (Bromley et al., 1993) sont particulièrement efficaces pour comparer des paires d'entités, tandis que les modèles d'attention (Vaswani et al., 2017) peuvent être utilisés pour capturer des relations complexes entre les entités.

Avantages par rapport aux méthodes traditionnelles :

1. Capacité à traiter des données textuelles complexes et non structurées :
   Les modèles d'apprentissage profond, en particulier les architectures basées sur les transformers comme BERT (Devlin et al., 2019), excellent dans le traitement de texte non structuré, permettant une meilleure compréhension du contexte et des nuances sémantiques.
2. Meilleure gestion des ambiguïtés et des variations linguistiques :
   Les techniques d'apprentissage profond peuvent capturer des variations subtiles dans le langage et résoudre des ambiguïtés contextuelles, ce qui est crucial pour l'alignement précis des ontologies dans des domaines spécialisés (Y. Li et al., 2020).
3. Potentiel d'amélioration continue :
   Grâce à leur capacité d'apprentissage sur de grandes quantités de données, ces modèles peuvent s'améliorer continuellement, s'adaptant à de nouveaux domaines ou à l'évolution des ontologies existantes (Z. Sun et al., 2017).
4. Capacité à capturer des relations complexes :
   Les architectures avancées comme les « Graph Neural Networks » (GNNs) peuvent modéliser efficacement les structures complexes des ontologies, permettant une meilleure compréhension des relations entre les entités (Z. Wang et al., 2018).

Défis et limitations :

Malgré leurs avantages, les approches basées sur l'apprentissage profond présentent également des défis :

1. Besoin de grandes quantités de données d'entraînement :
   L'efficacité de ces modèles dépend souvent de la disponibilité de vastes ensembles de données d'entraînement, qui peuvent être difficiles à obtenir dans certains domaines spécialisés (Kolyvakis et al., 2018).
2. Interprétabilité limitée :
   La nature "boîte noire" de nombreux modèles d'apprentissage profond peut rendre difficile l'interprétation et la justification des alignements proposés, un aspect crucial dans certains domaines d'application (Guidotti et al., 2018).
3. Coût computationnel :
   L'entraînement et l'inférence des modèles d'apprentissage profond peuvent être computationnellement intensifs, posant des défis pour l'alignement en temps réel ou à grande échelle (Strubell et al., 2019).

Perspectives futures :

L'intégration de ces techniques d'apprentissage profond avec les méthodes traditionnelles d'alignement d'ontologies pourrait conduire à des systèmes hybrides plus robustes et efficaces. Ces systèmes pourraient combiner la puissance de représentation des modèles d'apprentissage profond avec les connaissances du domaine et les règles logiques des approches traditionnelles, offrant ainsi une solution plus complète pour gérer la complexité croissante des ontologies.

De plus, les développements récents dans le domaine de l'apprentissage par transfert (H. Yao et al., 2020) et de l'apprentissage few-shot (Y. Wang et al., 2020) ouvrent des perspectives prometteuses pour l'application de ces techniques dans des domaines où les données d'entraînement sont limitées, un scénario courant dans l'alignement d'ontologies spécialisées.

En conclusion, les approches basées sur l'apprentissage profond représentent une avancée significative dans le domaine de l'alignement d'ontologies. Leur intégration judicieuse avec les méthodes traditionnelles et leur adaptation continue aux spécificités des environnements complexes ouvrent la voie à des systèmes d'alignement plus performants et adaptables.

3.4.3 Synthèse

En conclusion, le choix de l'approche d'alignement d'ontologies la plus appropriée pour notre problématique d'enrichissement des jumeaux numériques dans un environnement multi-modèle et multi-métiers n'est pas évident a priori. La complexité de notre contexte industriel, caractérisé par des données hétérogènes et des domaines d'expertise variés, rend difficile la sélection d'une méthode unique sans une exploration approfondie.

Face à cette complexité, il apparaît nécessaire d'explorer des outils d'alignement automatique, qui pourraient offrir la flexibilité et l'efficacité requises pour traiter notre problème à grande échelle. Ces outils, s'appuyant sur diverses techniques allant des approches traditionnelles aux méthodes d'apprentissage profond, pourraient nous permettre d'évaluer la faisabilité de l'alignement dans notre contexte spécifique.

Cette exploration nous aidera à identifier les défis concrets auxquels nous serons confrontés, tels que la gestion de la qualité des données, l'interprétabilité des résultats, ou encore les contraintes computationnelles. Elle nous permettra également d'évaluer dans quelle mesure ces outils peuvent être adaptés ou combinés pour répondre à nos besoins particuliers.

Ainsi, notre prochaine étape consistera à mener des expérimentations avec différents outils d'alignement automatique, en les appliquant à des échantillons représentatifs de nos ontologies industrielles. Cette démarche empirique nous fournira des insights précieux sur la pertinence et l'efficacité de ces approches dans notre contexte, et nous guidera vers la solution la plus adaptée pour relever les défis spécifiques de notre projet d'enrichissement des jumeaux numériques.

3.5 De l'intégration à la proposition de connaissances : techniques d'exploration et d'exploitation d'ontologies

Une fois les ontologies construites et alignées, l'étape cruciale d'exploration et d'exploitation des connaissances représentées devient essentielle, particulièrement dans le contexte des jumeaux numériques. Plusieurs approches ont été développées pour faciliter cette exploration, chacune

présentant des avantages et des limitations spécifiques. Dans cette section, nous nous concentrerons sur quatre approches principales :

1. Édition directe du fichier OWL :
   Cette approche consiste à manipuler directement le fichier OWL (Web Ontology Language) à l'aide d'un éditeur de texte. OWL est un langage de représentation des connaissances conçu pour définir et instancier des ontologies web (Hitzler et al., 2012, p. 2). Cette méthode offre un contrôle total sur le contenu de l'ontologie, permettant une manipulation fine des concepts, relations et axiomes. Cependant, elle nécessite une expertise approfondie en OWL et présente un risque élevé d'erreurs syntaxiques ou sémantiques.
2. Éditeurs d'ontologies :
   Des outils spécialisés comme Protégé (Musen, 2015) offrent une interface graphique pour la création, l'édition et l'exploration d'ontologies. Ces éditeurs facilitent la navigation dans la hiérarchie des concepts, la visualisation des relations, et souvent intègrent des fonctionnalités de raisonnement et de vérification de cohérence. Ils rendent l'exploration plus intuitive pour les utilisateurs non-experts tout en offrant des fonctionnalités avancées pour les spécialistes.
3. Outils de visualisation légers :
   Des applications web comme WebVOWL (Lohmann et al., 2016) permettent une visualisation interactive des ontologies. Ces outils transforment les structures ontologiques complexes en représentations graphiques intuitives, facilitant la compréhension des relations entre concepts. Ils sont particulièrement utiles pour des ontologies de taille modérée et pour une exploration rapide, offrant souvent des fonctionnalités de filtrage et de recherche pour naviguer efficacement dans l'ontologie.
4. Bases de données de graphes :
   Pour les ontologies de grande taille et les analyses complexes, l'utilisation de bases de données de graphes comme Neo4j ou Amazon Neptune offre des capacités avancées d'exploration et d'interrogation (Robinson et al., 2015). Ces solutions permettent de stocker efficacement les structures ontologiques sous forme de graphes, facilitant les requêtes complexes et les analyses de relations (Angles & Gutierrez, 2008). Elles sont particulièrement adaptées pour gérer des graphes de connaissances volumineux et pour réaliser des requêtes impliquant des parcours de graphes complexes.

Pour choisir l'approche la plus appropriée pour l'exploration et l'exploitation des ontologies dans le contexte de notre problématique d'enrichissement des jumeaux numériques, plusieurs facteurs clés doivent être pris en compte. Ces facteurs sont :

1. Complexité et taille des ontologies :
   - La taille et la complexité des ontologies utilisées pour représenter les jumeaux numériques dans un environnement industriel multi-modèle et multi-métiers sont susceptibles d'être importantes.
   - Cela favoriserait l'utilisation de bases de données de graphes, capables de gérer efficacement de grandes quantités de données et des relations complexes.
2. Diversité des utilisateurs :
   - Les jumeaux numériques seront probablement utilisés par divers profils, allant des experts en ontologies aux ingénieurs de domaine et aux opérateurs.

- Cela nécessite des interfaces d'exploration adaptées à différents niveaux d'expertise, suggérant une combinaison d'outils de visualisation légers pour les utilisateurs non-experts et d'éditeurs d'ontologies plus avancés pour les spécialistes.

3. Besoins d'intégration :
   - L'intégration des ontologies avec d'autres systèmes industriels et sources de données (comme l'IIoT) est cruciale.
   - Cela pourrait favoriser l'utilisation de bases de données de graphes ou d'APIs spécialisées pour faciliter l'interopérabilité.
4. Exigences de performance :
   - Dans un environnement industriel, la rapidité d'accès et d'analyse des données est souvent critique.
   - Les bases de données de graphes offrent généralement de meilleures performances pour les requêtes complexes sur de grandes ontologies.
5. Besoins de visualisation :
   - La capacité à visualiser rapidement et intuitivement les relations entre différents éléments du jumeau numérique est importante pour la prise de décision.
   - Cela suggère l'utilité des outils de visualisation légers, potentiellement en combinaison avec des bases de données de graphes pour la gestion des données sous-jacentes.
6. Flexibilité et évolutivité :
   - Les jumeaux numériques et les ontologies associées sont susceptibles d'évoluer au fil du temps.
   - Cela nécessite des outils flexibles comme les éditeurs d'ontologies pour les mises à jour, combinés à des systèmes robustes comme les bases de données de graphes pour la gestion à long terme.
7. Besoins de raisonnement et d'inférence :
   - L'exploitation des connaissances dans les jumeaux numériques peut nécessiter des capacités de raisonnement avancées.
   - Cela pourrait favoriser l'utilisation d'éditeurs d'ontologies avec des fonctionnalités de raisonnement intégrées, ou l'intégration de moteurs de raisonnement avec des bases de données de graphes.
8. Contraintes de sécurité et de confidentialité :
   - Dans un environnement industriel, la sécurité et la confidentialité des données sont cruciales.
   - Cela pourrait influencer le choix entre des solutions locales (comme les éditeurs d'ontologies) et des solutions basées sur le cloud (certaines bases de données de graphes).

Il est important de noter que l'évaluation approfondie de ces différentes approches nécessite une expérimentation pratique avec des ontologies concrètes. Sans une ontologie spécifique à explorer, il est difficile de comparer de manière détaillée l'efficacité et la pertinence de chaque méthode dans un contexte donné. Cette démarche empirique nous permettrait de valider ou d'ajuster notre approche théorique et de choisir la combinaison d'outils la plus adaptée à notre contexte spécifique d'enrichissement des jumeaux numériques dans un environnement industriel multi-modèle et multi-métiers.

3.6 Synthèse

Notre état de l'art a mis en lumière plusieurs avancées significatives dans le domaine de l'intégration des données et de l'interopérabilité sémantique, tout en révélant des lacunes importantes qui persistent dans la recherche actuelle.

1. Transformation des données brutes en information :
   Les techniques de « deep learning », en particulier les réseaux de neurones convolutifs (CNN) et récurrents (RNN), se distinguent par leur capacité à extraire automatiquement des caractéristiques pertinentes à partir de données brutes et complexes. Les « Graph Neural Networks » (GNN) émergent comme particulièrement adaptés pour traiter des données structurées en graphes. Ces approches offrent une scalabilité cruciale pour gérer les volumes massifs de données générés par l'IIoT. Les techniques de plongement (« embeddings ») permettent d'intégrer efficacement des données textuelles non structurées. Cependant, la gestion de données extrêmement variées et non structurées reste un défi, tout comme l'intégration efficace de l'expertise métier dans le processus d'extraction d'information.

2. Structuration de l'information :
   La construction d'ontologies en s'appuyant sur des ontologies existantes offre des avantages considérables, permettant une réutilisation efficace des connaissances et favorisant l'interopérabilité. L'utilisation de "backbone ontologies" et d'ontologies pivot s'avère pertinente pour aligner et intégrer des ontologies de divers domaines industriels. L'intégration de techniques d'IA, comme les modèles de plongement lexical et les GNN, ouvre des perspectives prometteuses pour automatiser certains aspects de la construction et de l'alignement des ontologies. Néanmoins, l'automatisation complète de la construction d'ontologies de qualité demeure un objectif difficile à atteindre, et il manque encore des outils facilitant la collaboration entre experts métier et ingénieurs ontologues.

3. Alignement et interopérabilité :
   Face à la complexité de l'alignement d'ontologies dans un environnement multi-modèle et multi-métiers, l'exploration d'outils d'alignement automatique s'avère nécessaire. Ces outils, s'appuyant sur diverses techniques allant des approches traditionnelles aux méthodes d'apprentissage profond, pourraient offrir la flexibilité et l'efficacité requises. Cependant, l'alignement automatique d'ontologies hétérogènes reste problématique, et il est nécessaire de développer des approches plus flexibles et adaptatives, capables de gérer l'incertitude inhérente à ce processus.

4. Exploitation des connaissances :
   Pour l'exploration et l'exploitation des ontologies dans le contexte des jumeaux numériques, une combinaison d'approches semble nécessaire. Les bases de données de graphes apparaissent particulièrement adaptées pour gérer la complexité et la taille des ontologies industrielles, tandis que les outils de visualisation légers pourraient faciliter l'exploration pour les utilisateurs non-experts. Les éditeurs d'ontologies restent pertinents pour les mises à jour et le raisonnement avancé. Cependant, il manque encore

des outils véritablement intuitifs permettant aux non-spécialistes d'explorer efficacement ces connaissances, ainsi que des méthodes fiables pour évaluer la qualité et la pertinence des informations extraites.

Pour répondre à ces défis, nos contributions se concentreront sur trois axes majeurs :

1. Ontologification des données :
   - Développement d'une approche hybride combinant techniques d'IA et expertise humaine pour convertir les données brutes en ontologies.
   - Utilisation de méthodes de deep learning, notamment les réseaux de neurones pour traiter les données non structurées.
   - Application de techniques de plongement (embeddings) pour représenter efficacement les données textuelles.
   - Conception d'une architecture spécifique pour traiter les particularités des données IIoT, en tenant compte de leur nature temps réel et de leur volume.

2. Alignement et intégration ontologique :
   - Élaboration d'une méthodologie pour la création d'une ontologie de référence par les experts métier.
   - Développement d'un processus semi-automatique pour aligner l'ontologie générée par l'IA avec l'ontologie de référence.
   - Conception de méthodes d'intégration pour fusionner les ontologies alignées, en prenant en compte les spécificités des données IIoT.

3. Exploration et exploitation des connaissances :
   - Recherche d'interfaces intuitives permettant aux acteurs métier d'explorer les ontologies intégrées dans le jumeau numérique.
   - Proposition d'outils de visualisation adaptés à différents niveaux d'expertise, combinant des approches de visualisation légère et des bases de données de graphes pour gérer la complexité.
   - Élaboration de métriques et de protocoles d'évaluation pour mesurer la pertinence et la qualité des connaissances générées dans le contexte industriel.

Ces trois axes de contribution forment un cadre méthodologique complet pour l'enrichissement des jumeaux numériques. Ils adressent l'ensemble du processus, de la transformation des données brutes à l'exploitation des connaissances, en passant par l'intégration sémantique, tout en prenant en compte les spécificités de l'environnement industriel multi-modèle et multi-métiers.

Chapitre 4 Cadre méthodologique pour l'enrichissement sémantique des jumeaux numériques

4.1 Introduction

Afin de répondre à la problématique centrale visant à enrichir les capacités d'un jumeau numérique par une intégration efficace de données provenant de sources externes, les questions de recherche suivantes ont été posées dans le chapitre 2 pour guider notre travail :

1. **De quelle manière convertir des données brutes, en particulier non structurées, en informations utiles ?**
    a. Comment traiter efficacement les données non structurées en général ?
    b. Quelles sont les particularités du traitement des données de l'Internet Industriel des Objets (IIoT) ?
2. **Comment structurer les informations pour leur intégration dans un jumeau numérique ?**
    a. Quel support utiliser dans la structuration des informations issues de données non structurées ?
    b. En quoi la structuration des données IIoT est-elle spécifique ?
3. **Quels sont les enjeux pour l'intégration d'informations structurées avec des informations non structurées dans un jumeau numérique ?**
    a. Comment surmonter les défis d'intégration des données non structurées ?
    b. Quelle est la spécificité de l'intégration des données IIoT ?
4. **Comment peut-on rendre les informations intégrées accessibles pour les acteurs métier, afin de générer de nouvelles connaissances ?**

L'état de l'art mené dans le chapitre précédent a révélé que l'intelligence artificielle (IA) et les ontologies représentent les outils les plus prometteurs pour répondre à ces défis. Ces technologies offrent des perspectives uniques pour la transformation, la structuration, l'intégration, et l'exploitation des données au sein de jumeaux numériques, promouvant ainsi l'interopérabilité sémantique. Comme l'ont souligné (Tao, Zhang, Liu, et al., 2019), l'IA et les ontologies jouent un rôle crucial dans l'amélioration de l'interopérabilité et de la performance des jumeaux numériques dans un environnement industriel complexe.

L'analyse de l'état de l'art souligne l'importance d'aborder en premier lieu le traitement des données, comme le précise la première question de recherche, pour poser les bases nécessaires à l'étude des questions de recherche suivantes. En outre, pour aborder la complexité inhérente à l'intégration de données issues de l'IIoT, une combinaison d'approches à la fois partagées et divergentes est nécessaire. Cette dualité d'approches suggère que notre contribution peut offrir un cadre de référence méthodologique outillé, adapté aux besoins spécifiques de l'interopérabilité sémantique dans les jumeaux numériques. Cette approche s'aligne avec les travaux de (Grieves & Vickers, 2017), qui ont souligné l'importance d'une méthodologie flexible pour l'intégration des données dans les jumeaux numériques.

Dans ce chapitre, nous allons traiter simultanément ces quatre questions de recherche, en détaillant comment notre approche, basée sur les avancées en intelligence artificielle et en ontologies, permet de construire un système interopérable et évolutif. Notre objectif est de démontrer comment, en naviguant à travers la complexité des données non structurées ainsi que celles de l'Internet Industriel des Objets, nous pouvons faciliter une intégration harmonieuse de ces données dans les jumeaux numériques, tout en rendant ces informations accessibles aux acteurs métier pour la génération de nouvelles connaissances. Cette approche s'inscrit dans la lignée des travaux de (Boje et al., 2020), qui ont mis en évidence l'importance de l'accessibilité des données pour la prise de décision dans les environnements industriels complexes.

Afin d'illustrer notre propos, nous nous plaçons dans un contexte industriel, pour lequel un système d'information peut être représenté par une ontologie composée d'une TBox (Terminological Box) et de plusieurs silos de données. Comme illustré par la Figure 30, un des objectifs applicatifs de ce travail est une contribution sur la capacité à assimiler les données et à les intégrer de manière adéquate dans l'ontologie pertinente.

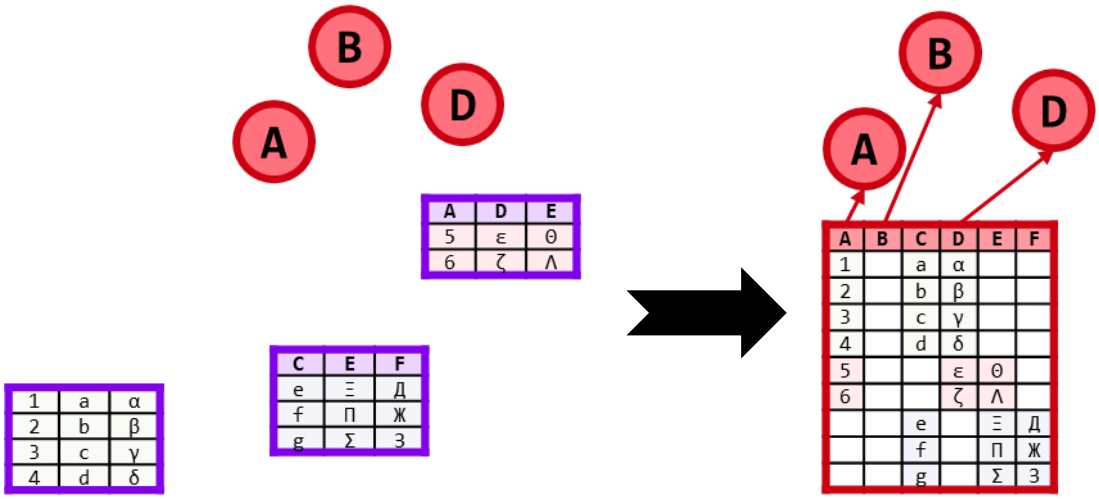


Figure 30 Processus d'intégration des données dans l'ontologie

La Figure 30 illustre le processus d'intégration des données dans l'ontologie. L'ontologie est définie par les concepts A, B, C, représentés en haut de la figure. Les silos de données sont représentés par des tableaux colorés, qui peuvent être classés en deux catégories distinctes :

Les tableaux avec des noms de colonnes (en violet et orange) représentent les données issues de logiciels métiers. Ces données sont généralement structurées et leurs attributs sont clairement définis.

Le tableau sans nom de colonne (en vert) représente une table composée de données IoT. L'absence de noms de colonnes n'est pas une erreur, mais reflète la nature souvent non structurée ou semi-structurée des données IoT, qui peuvent nécessiter un traitement supplémentaire pour être correctement interprétées et classifiées.

Après le processus de traitement, les concepts identiques présents à la fois dans l'ontologie et dans les silos de données sont reliés, tandis que les données elles-mêmes sont jointes. Cette approche s'inspire des travaux de (Studer et al., 1998) sur la représentation des connaissances à l'aide d'ontologies.

Il est important de noter que les données initialement non structurées (comme celles du tableau rose) sont intégrées dans les colonnes A, B et D après le traitement. Ce processus illustre la capacité de notre approche à interpréter et à structurer les données IoT en les alignant avec l'ontologie existante. Cette transformation est cruciale pour l'intégration efficace des données IoT dans le système d'information global, comme l'ont souligné (Atzori et al., 2010) dans leurs travaux sur l'Internet des Objets.

Cette méthode d'intégration des données hétérogènes dans une ontologie unifiée s'aligne avec les recherches de (Gruber, 1993) sur la conception d'ontologies pour le partage de connaissances. Elle permet non seulement de gérer la diversité des sources de données dans un environnement industriel moderne, mais aussi de faciliter l'interopérabilité et l'analyse croisée des données provenant de différentes sources.

Plusieurs scénarios doivent être pris en compte pour répondre à notre problématique, comme le montre le Tableau 5. Premièrement, la source de données peut être connue ou inconnue par rapport à l'état de l'ontologie de destination avant le traitement. Deuxièmement, les données provenant de ces sources peuvent être soit connues (déjà vues), soit inconnues. Enfin, l'ontologie cible peut déjà contenir la donnée considérée, ou pas.

Le Tableau 5 présente une synthèse de la diversité des situations susceptibles de se produire lorsque l'on cherche à enrichir l'ontologie de destination. Chaque cas possède ses propres implications et défis, et la méthodologie à mettre en place doit viser à les aborder de manière systématique et efficace. Cette approche s'aligne avec les travaux de (N. F. Noy & McGuinness, 2001) sur la création et la gestion d'ontologies évolutives.

Tableau 5 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
7	Connue	Jamais vu	Absente

La spécificité des données industrielles doit être prise en compte dans la méthode proposée. En effet, les données provenant des sources de l'Internet Industriel des Objets sont souvent imparfaites, pouvant contenir des entrées manquantes, erronées ou redondantes. Cette problématique a été soulignée par (L. D. Xu et al., 2014) dans leur étude sur les défis de l'intégration des données IIoT.

De plus, les ontologies industrielles sont des entités dynamiques et en constante évolution. Il est donc possible que l'ensemble de données ne soit pas entièrement à jour, ce qui implique que le système doit continuellement apprendre et s'adapter pour remplir correctement sa tâche. Par exemple, un nouveau type de source de données peut être ajouté sans que l'information ne soit transférée vers le système. Il est alors essentiel de prendre conscience que le modèle pourrait être obsolète, et de demander à l'utilisateur s'il souhaite toujours l'utiliser. Cette approche adaptative s'inscrit dans la lignée des travaux de (Parisi et al., 2019) sur l'apprentissage continu

En outre, il est important de minimiser le stockage des données. Autrement dit, il n'est pas souhaitable de stocker l'intégralité du jeu de données à chaque fois que le système doit être entraîné. Cela soulève des défis non seulement en termes d'efficacité, mais aussi en termes d'espace de stockage, de coût et de temps. Cette problématique rejoint les préoccupations soulevées par (Philip Chen & Zhang, 2014) dans leur étude sur la gestion des big data dans les environnements industriels.

Un autre problème concerne les silos de données, qui représentent des ensembles de données appartenant à différentes parties prenantes. Ces silos ne peuvent pas être facilement explorés ou connectés, ce qui complique davantage le processus d'intégration et d'exploitation des données. Cette problématique a été abordée par (Sivarajah et al., 2017) dans leurs travaux sur la gestion des silos de données dans les entreprises.

Dans la section suivante, nous approfondissons les choix effectués à l'issue de l'état de l'art pour identifier et caractériser les outils qui devront être mis en œuvre pour assurer l'intégration de données multi-sources au sein d'un jumeau numérique, et répondant aux 4 questions de recherche précédentes.

4.2 Présentation générale de la méthodologie

4.2.1 Présentation du cas d'application Airbus

Dans le cadre de nos travaux, nous avons identifié deux cas principaux d'intégration de données, qui sont étroitement liés aux problématiques de stockage. Il est important de noter que cette relation entre l'intégration et le stockage des données s'est révélée être un aspect crucial de notre recherche, influençant directement notre approche méthodologique.

Le premier cas concerne les modèles stockés directement dans des logiciels métiers, tels que Cameo®[1]. Ce cas, représentatif de la situation générale chez Airbus, a déjà été traité et maîtrisé au sein de l'entreprise. Il s'agit principalement d'un problème informatique et de conversion de données, qui peut être résolu par un développement spécifique.

De tels logiciels métiers utilisent souvent des formats de fichiers propriétaires ou des structures de données spécifiques pour stocker les modèles. Par exemple, Cameo®, un outil de modélisation UML/SYSML, utilise le format de fichier '.mdzip' pour enregistrer les diagrammes et les éléments de modèles.

Lorsque les fichiers sont stockés dans des formats identifiés et bien documentés, le processus d'extraction, de transformation et de chargement (ETL) des données devient plus facile à mettre en œuvre. Les développeurs peuvent s'appuyer sur les spécifications des formats de fichiers pour créer des scripts ou des outils automatisés capables de lire, d'analyser et d'extraire les informations pertinentes. Cette approche programmatique permet de gérer efficacement les données structurées et de les préparer pour une intégration ultérieure dans l'ontologie cible.

Cependant, nous avons décidé de ne pas traiter ce cas dans nos travaux. Bien qu'il s'agisse d'un défi intéressant, l'extraction de modèles à partir de formats de fichiers propriétaires relève

[1]Cameo Systems Modeler est un outil de modélisation UML/SysML développé par No Magic, Inc. (maintenant intégré par Dassault Systèmes).

principalement du domaine du développement logiciel. Notre objectif principal est de nous concentrer sur les aspects ontologiques et sémantiques de l'intégration des données, plutôt que sur les détails techniques de l'implémentation de l'ETL.

Dans le deuxième cas, les données sont extraites sans transformation et/ou stockées dans des bases de données relationnelles ou des datalakes (lacs de données). Les bases de données relationnelles, telles que MySQL ou PostgreSQL, organisent les données dans des tables avec des relations prédéfinies entre elles. Cette structure permet des requêtes efficaces et des opérations de jointure, mais peut être moins flexible pour des données complexes ou non structurées. Les datalakes, quant à eux, sont des référentiels centralisés qui permettent de stocker de grands volumes de données brutes dans leur format natif (Fang, 2015). Ils offrent une grande flexibilité mais peuvent nécessiter plus de traitement pour extraire des informations pertinentes.

Notre approche vise à développer une architecture qui fonctionnera dans un grand nombre de cas, en particulier lorsque des identifiants uniques sont disponibles. L'objectif est de garantir l'exhaustivité des données, sans perte d'information, tout en évitant les redondances ou les incohérences. Pour atteindre cet objectif, nous nous baserons sur l'architecture de Sherlock, en y apportant des améliorations et en mettant l'accent sur le travail ontologique.

Afin de pouvoir tester et valider notre approche, nous avons tout d'abord reçu un jeu de données extrait de ce premier cas. L'intérêt pour nous est ici de pouvoir tester sur un jeu de données dont les résultats sont connus, ce qui est nécessaire dans une approche de machine learning.

Avec l'aide des experts Airbus, nous avons pu construire un premier cas d'étude concret exploitant de données issues de l'outil Windchill®[2], un système de gestion du cycle de vie des produits (PLM) utilisé chez Airbus. Windchill® permet de gérer les informations relatives aux produits, telles que les nomenclatures, les documents techniques, les processus de conception et les données de fabrication. Il s'agit d'un outil central qui intègre des données provenant de différentes sources et permet une collaboration efficace entre les équipes.

Cependant, nous sommes confrontés à une problématique spécifique : comment gérer efficacement les données qui ne sont pas présentes dans Windchill® mais qui sont dispersées dans d'autres outils métiers ? Cette question s'inscrit dans le cadre plus large de l'intégration des données hétérogènes dans les systèmes d'information d'entreprise, un défi majeur identifié par (Doan et al., 2012) dans leur étude sur l'intégration des données.

Nous disposons d'un volume considérable de données, comprenant des informations issues de logiciels métiers, des données provenant de l'Internet des Objets (IoT), des identifiants de liens et des métadonnées décrivant ces liens. D'un point de vue numérique, cela se traduit par un ensemble de bases de données relationnelles dont les relations ne sont plus explicitement définies. Ces données, issues de divers logiciels métiers, possèdent chacune leur structure interne spécifique, mais cette structure a été perdue au fil du temps. Comme l'ont souligné (Philip Chen & Zhang, 2014), cette perte de structure est un défi courant dans les environnements industriels complexes, où les données sont souvent stockées dans des silos isolés.

[2]Windchill est une suite logicielle de gestion du cycle de vie des produits (PLM - Product Lifecycle Management) développée par PTC (Parametric Technology Corporation).

Chacun de ces outils métiers contient des informations cruciales qui doivent être intégrées pour obtenir une vue holistique du produit tout au long de son cycle de vie. Windchill, par exemple, contient des données essentielles telles que les nomenclatures des avions, les documents techniques (plans, schémas, spécifications), les données de conception (versions, commentaires, validations), les informations de fabrication (gammes, outillages, temps d'assemblage) et les données de maintenance (manuels, bulletins de service, historiques). Ces données structurées et interconnectées permettent une gestion efficace du cycle de vie des avions, comme l'ont démontré (Stark, 2015) dans ses travaux sur les systèmes PLM dans l'industrie.

Notre objectif est illustré sur la Figure 31. Les données sont dispersées dans différents emplacements de stockage, représentés par des tableaux de couleurs orange, jaune et vert. La variabilité de la quantité de données selon les tableaux est symbolisée par des tailles différentes. De plus, le type "informatique" des données (par exemple : entier, chaîne de caractères, booléen) peut également varier, ce qui est illustré par l'usage de caractères différents à l'intérieur de chaque tableau. Cette représentation s'inspire des travaux de (Batini et al., 2009) sur la visualisation des données hétérogènes dans les systèmes d'information complexes.

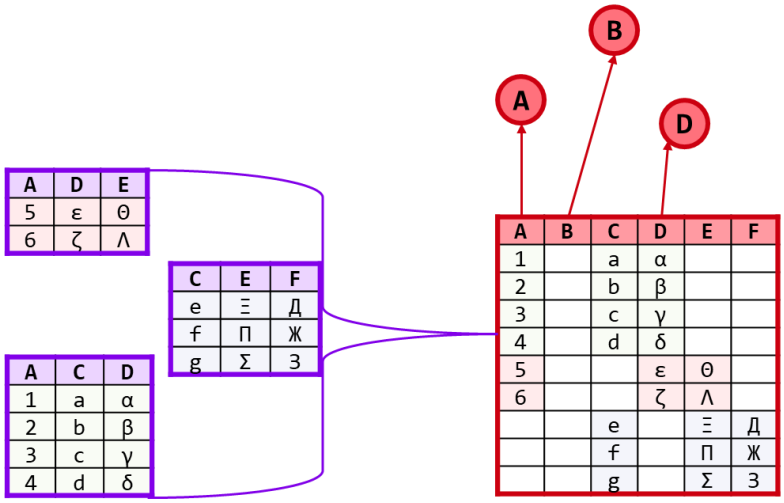


Figure 31 Processus d'intégration sémantique des données multisources

Chacune de ces données appartient à un type sémantique précis, représenté par une lettre majuscule en gras. Par exemple, les données de type "A" peuvent représenter des numéros de pièces d'avion, tandis que les données de type "B" peuvent correspondre à des codes de fournisseurs. Cette catégorisation sémantique s'aligne avec les principes de l'intégration des données proposés par (Lenzerini, 2002).

Notre objectif est de rassembler les types sémantiques identiques et de regrouper les données appartenant à chaque type, indépendamment de leur provenance initiale. Ce processus permet de rendre les données "transversales", facilitant ainsi leur analyse et leur exploitation. Cette approche s'inspire des travaux de (Bernstein & Haas, 2008) sur l'intégration de données basée sur la sémantique et facilite grandement l'analyse et l'exploitation des données, car elle permet de travailler sur des ensembles cohérents et significatifs.

En résumé, notre cas d'application se concentre sur l'intégration de données multisources au sein d'Airbus, en tenant compte des spécificités des différents outils et formats de stockage.

4.2.2 Le choix de l’intelligence artificielle

La littérature propose plusieurs outils permettant d'adresser l'intégration de données multi-sources. Ces outils, bien qu'ils ne soient pas directement adaptés à notre contexte spécifique, offrent des principes et des techniques qui peuvent être utiles pour développer notre approche. Dans le chapitre précédent, nous avons conclu que l'outil le plus adapté s'appuie sur les techniques du machine learning, en particulier du « deep learning ».

En effet, les algorithmes de « deep learning » sont particulièrement efficaces pour gérer les données imparfaites ou incomplètes, caractéristiques de l'Internet Industriel des Objets. Comme l'ont souligné (LeCun et al., 2015), ces algorithmes peuvent identifier des motifs complexes et des relations dans les données, même en présence d'entrées manquantes, erronées ou redondantes.

De plus, le deep learning offre des mécanismes pour gérer l'évolution constante des ontologies industrielles. Selon (Goodfellow et al., 2016), les modèles de deep learning peuvent être régulièrement mis à jour et réentraînés avec de nouvelles données, s'adaptant ainsi aux changements dans les sources de données et restant à jour. Cette flexibilité est essentielle pour maintenir la pertinence et l'efficacité du système dans un environnement dynamique.

En outre, certaines techniques de « deep learning », comme l'apprentissage incrémental décrit par (Parisi et al., 2019), permettent d'entraîner les modèles de manière progressive sans avoir à stocker l'intégralité du jeu de données à chaque itération. Cela répond aux préoccupations de stockage et d'efficacité, en minimisant l'espace requis et en optimisant le processus d'apprentissage.

L'architecture que nous proposons est structurée en deux phases distinctes : la phase **d'entraînement** et la phase **opérationnelle**. Cette organisation est courante dans les applications de « machine learning », comme l'ont noté (Alpaydin, 2020) et (T. M. Mitchell, 1997).

Durant la phase **d'entraînement**, le système apprend à partir d'un ensemble de données d'apprentissage. Ce processus permet au modèle d'optimiser ses paramètres internes pour minimiser une fonction de perte définie, capturant ainsi les schémas sous-jacents dans les données (Goodfellow et al., 2016). C'est pendant cette phase que le modèle acquiert la capacité de généraliser à partir des exemples fournis pour effectuer la tâche pour laquelle il a été conçu.

La phase **opérationnelle**, quant à elle, utilise le modèle entraîné pour réaliser des tâches spécifiques sur de nouvelles données. Dans notre cas, cette phase implique l'utilisation du modèle pour l'intégration des données IoT dans une ontologie.

4.2.3 Le choix d'une approche de traitement des données

L'analyse de la littérature et des besoins spécifiques de notre projet nous a conduits à identifier deux approches prometteuses basées sur l'intelligence artificielle pour préparer et réaliser l'intégration des données : l'approche basée sur les graphes de connaissances et la reconnaissance de type sémantique. Ces méthodes, bien que non explicitement détaillées dans notre état de l'art précédent, émergent comme des solutions pertinentes pour relever les défis de notre problématique.

Ces approches s'inscrivent dans la continuité des travaux sur l'intégration de données hétérogènes, comme ceux de (Doan et al., 2012), tout en exploitant les avancées récentes en intelligence artificielle pour améliorer l'efficacité et la précision du processus d'intégration.

Dans les sections suivantes, nous allons examiner plus en détail ces approches et les outils adaptés à leur mise en œuvre dans le cadre de notre projet.

4.2.3.1 Approche par graphes

L'approche par graphes pour l'intégration de données est une méthode puissante qui exploite la structure relationnelle inhérente aux données pour faciliter leur fusion et leur analyse. Cette approche s'appuie sur la théorie des graphes, un domaine mathématique bien établi, pour représenter et manipuler des ensembles de données complexes et interconnectés.

Dans le contexte de l'intégration de données, les graphes offrent une représentation flexible et intuitive des relations entre les entités de données, permettant de capturer efficacement la sémantique et la structure des différentes sources de données. Comme l'ont souligné (Pujara et al., 2013), les graphes peuvent modéliser naturellement les dépendances complexes entre les données, facilitant ainsi la découverte de correspondances et de liens cachés entre les différents ensembles de données.

L'utilisation de graphes pour l'intégration de données a gagné en popularité ces dernières années, notamment grâce aux avancées dans les techniques d'apprentissage automatique appliquées aux structures de graphes. Par exemple, les travaux de (Hamilton et al., 2017b) sur l'apprentissage de représentations de graphes ont ouvert de nouvelles perspectives pour l'analyse et l'intégration de données hétérogènes à grande échelle.

4.2.3.1.1 SiMA

SiMa (Koutras et al., 2022) est une méthode innovante d'appariement de schémas qui utilise des Graph Neural Networks (GNN) pour mettre en correspondance des colonnes provenant de différents silos de données. L'approche de SiMa se distingue par sa capacité à exploiter les relations existantes entre les colonnes au sein de chaque silo pour prédire des relations similaires entre les silos. Cette approche s'inscrit dans la lignée des travaux sur l'apprentissage de représentations de graphes (Grover & Leskovec, 2016) et l'utilisation de GNN pour l'intégration de données (Wu et al., 2021).

Le pipeline de SiMa, illustré dans la Figure 32, comprend plusieurs étapes clés :

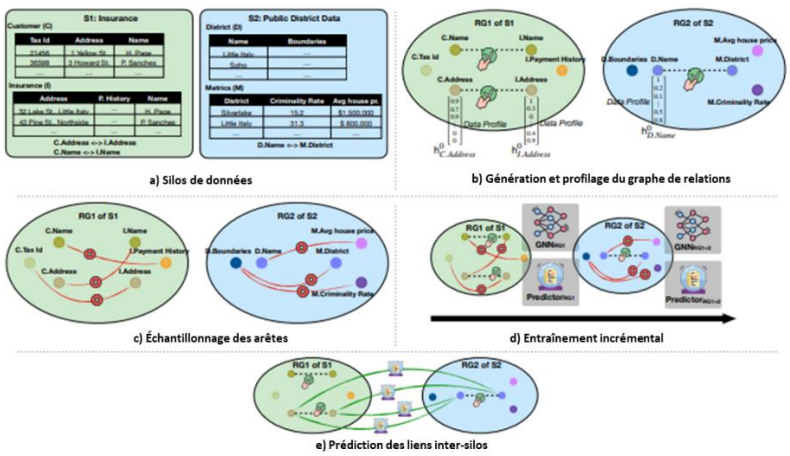


Figure 32 Pipeline de la méthode SiMa

a) Silos de données : le point de départ sont les silos de données, qui contiennent des informations structurées mais isolées. Dans l'exemple de la figure, on voit deux silos : un pour les assurances et un autre pour les données de district public.

b) Génération et profilage du graphe de relations (RG) : SiMa transforme chaque ensemble de colonnes et leurs correspondances au sein d'un silo en un graphe de relations, où les colonnes sont représentées par des nœuds et les correspondances par des arêtes. Pour chaque colonne, SiMa construit un profil comprenant 987 caractéristiques, telles que le nombre de valeurs numériques ou des agrégats au niveau des caractères.

c) Échantillonnage des arêtes négatives : SiMa applique des techniques sophistiquées d'échantillonnage d'arêtes négatives sur les graphes pour affiner la capacité de prédiction des GNN. Cette étape est cruciale pour équilibrer le ratio d'exemples positifs (relations existantes) et négatifs (relations improbables) lors de l'entraînement.

d) Entraînement incrémental : SiMa utilise un entraînement incrémental qui permet non seulement d'améliorer la capacité de prédiction des correspondances, mais aussi d'entraîner un GNN individuellement pour chaque silo, sans avoir à consolider tous les profils de données en un seul endroit.

e) Prédiction des liens inter-silos : enfin, SiMa utilise les embeddings de graphe appris par les GNN pour capturer la similarité entre les colonnes et découvrir des correspondances entre les silos de données. Cette étape implique l'ajustement fin d'un modèle de prédiction de liens.

SiMa se distingue par plusieurs contributions significatives :

- Il formalise le problème de la fédération des silos de données, un défi majeur dans l'intégration de données hétérogènes (Dong & Srivastava, 2013).
- Il propose un cadre d'apprentissage basé sur les GNN qui découvre les relations entre différents silos de données, généralisant les correspondances locales au sein de chaque silo à des liens fédérés entre les silos, même avec de nouvelles données non vues.
- Il représente les silos de données et les connaissances sur les correspondances entre les ensembles de données à l'intérieur de ceux-ci sous forme de graphes, transformant ainsi le problème de fédération des silos de données en une tâche de prédiction de liens (Liben-Nowell & Kleinberg, 2007).
- Il propose deux techniques d'optimisation : l'échantillonnage d'arêtes négatives et l'entraînement incrémental du modèle, qui améliorent l'efficacité et l'efficacité de l'entraînement du GNN pour l'appariement entre silos.

Selon les auteurs, SiMa démontre des améliorations significatives en termes d'efficacité par rapport à l'état de l'art, avec des gains de performance de plusieurs ordres de grandeur. Cependant, il est important de noter que ces affirmations nécessiteraient une validation indépendante dans notre contexte spécifique.

Les graphes de connaissances, exemplifiés par l'approche SiMa, ont démontré leur efficacité dans certains contextes d'intégration de données. Ces outils sont particulièrement performants lorsque les données ont déjà subi un certain degré de prétraitement et de structuration. Comme

l'ont souligné (Paulheim, 2016) et (Hogan et al., 2021), les graphes de connaissances offrent un potentiel significatif pour la réutilisation dans divers contextes et pour l'établissement de liens sémantiques une fois qu'une ontologie initiale a été générée.

Cependant, malgré leurs avantages, ces approches présentent des limitations dans le traitement des données brutes de l'Internet des Objets Industriel. Comme l'ont noté (L. D. Xu et al., 2014) et (Philip Chen & Zhang, 2014), les données IIoT sont souvent caractérisées par leur nature désorganisée et leur volume important, rendant les approches basées sur les graphes de connaissances moins adaptées pour leur traitement initial.

Néanmoins, l'aspect de profilage des données utilisé dans SiMa reste particulièrement intéressant. (Abedjan et al., 2015) et (Naumann, 2014) ont souligné l'importance du profilage dans l'intégration de schémas et la compréhension des caractéristiques des données. Cette technique pourrait être adaptée pour mieux répondre aux spécificités des données IIoT dans le contexte des jumeaux numériques industriels.

Par conséquent, nous explorerons des approches complémentaires, telles que la reconnaissance de type sémantique, qui s'appuient sur les techniques de profilage tout en offrant une plus grande adaptabilité à notre contexte industriel spécifique. Ces méthodes pourraient potentiellement combiner les forces des graphes de connaissances avec des techniques plus adaptées au traitement initial des données IIoT.

4.2.3.2 Approche par type sémantique

Une autre approche prometteuse pour l'intégration des données consiste à utiliser le type sémantique propre aux colonnes de données. Cette méthode, qui s'appuie sur l'analyse du contenu et de la signification des données, offre une perspective complémentaire à l'approche par graphes.

Le type sémantique d'une colonne de données fait référence à la signification ou à l'objectif des données contenues dans cette colonne. Par exemple, une colonne contenant des dates aurait le type sémantique « date », tandis qu'une colonne contenant des noms de famille aurait un type sémantique « nom de famille ». Comme l'ont souligné (Pham et al., 2016), alors que la plupart des systèmes peuvent détecter des types atomiques tels que des chaînes, des entiers et des booléens, les types sémantiques sont plus puissants et essentiels dans de nombreux cas.

La détection des types sémantiques est cruciale pour diverses tâches telles que le nettoyage des données, la mise en correspondance de schémas et la découverte de données (Hulsebos et al., 2019). Cependant, les systèmes existants reposent souvent sur des recherches de dictionnaires et des correspondances d'expressions régulières, qui ne sont pas robustes aux données "sales" (c'est-à-dire des données contenant des erreurs, des incohérences ou des valeurs manquantes) et ne détectent qu'un nombre limité de types (Chu et al., 2015).

Plusieurs outils ont été développés ces dernières années pour améliorer la détection des types sémantiques. Nous présentons ici deux outils qui répondent le mieux à notre besoin : Sherlock (Hulsebos et al., 2019) et Sato (D. Zhang et al., 2020).

4.2.3.2.1 Sherlock

Sherlock, développé par (Hulsebos et al., 2019), est un réseau neuronal profond conçu pour détecter les types sémantiques de colonnes de données. Cette approche innovante s'appuie sur

un vaste ensemble d'entraînement de 686765 colonnes extraites du corpus VizNet, couvrant 78 types sémantiques issus de DBpedia. La robustesse et la polyvalence de Sherlock en font un outil particulièrement intéressant pour notre problématique d'enrichissement des jumeaux numériques dans un contexte industriel complexe.

L'architecture de Sherlock se distingue par sa capacité à générer des représentations de longueur fixe à partir de colonnes de longueur variable, facilitant ainsi l'interprétation des résultats et fournissant des indices précieux au réseau de neurones. Pour ce faire, Sherlock extrait quatre catégories de caractéristiques de chaque colonne :

1. Statistiques globales (27 caractéristiques) : ces caractéristiques décrivent les propriétés statistiques de haut niveau des colonnes. Par exemple, l'entropie de colonne permet de différencier les types avec des valeurs répétées (comme le genre) de ceux avec des valeurs uniques (comme le nom). Cette approche s'inspire des travaux de (Naumann, 2014) sur le profilage de données pour l'intégration de schémas.
2. Distributions de caractères agrégées (960 caractéristiques) : Sherlock calcule le décompte des 96 caractères imprimables ASCII pour chaque valeur d'une colonne, puis agrège ces décomptes avec 10 fonctions statistiques. Cette approche, inspirée des travaux de (Ramnandan et al., 2015) sur l'analyse de la distribution des caractères, permet de capturer des motifs subtils dans les données.
3. Embeddings de mots pré-entraînés (200 caractéristiques) : pour caractériser le contenu sémantique des colonnes, Sherlock utilise des plongements lexicaux (word embeddings). Un dictionnaire GloVe pré-entraîné, contenant des représentations de 50 dimensions de 400 000 mots anglais, est utilisé. Cette technique s'appuie sur les avancées en traitement du langage naturel décrites par (Pennington et al., 2014).
4. Vecteurs de paragraphes auto-entraînés (400 caractéristiques) : Sherlock implémente la version Distributed Bag of Words de Paragraph Vector, initialement développée par (Le & Mikolov, 2014) pour représenter numériquement le "sujet" des textes. Dans cette implémentation, chaque colonne est considérée comme un "paragraphe" et les valeurs au sein d'une colonne comme des "mots".

L'architecture générale de Sherlock, illustrée dans la Figure 33, comprend quatre étapes principales :

1. Regroupement des colonnes de toutes les classes dans un jeu de données unique.
2. Extraction d'une fenêtre aléatoire d'une colonne d'entraînement et calcul des caractéristiques.
3. Entraînement d'un modèle pour prédire le type à partir des valeurs non utilisées dans la fenêtre aléatoire.
4. Identification du type sémantique sur de nouvelles données.

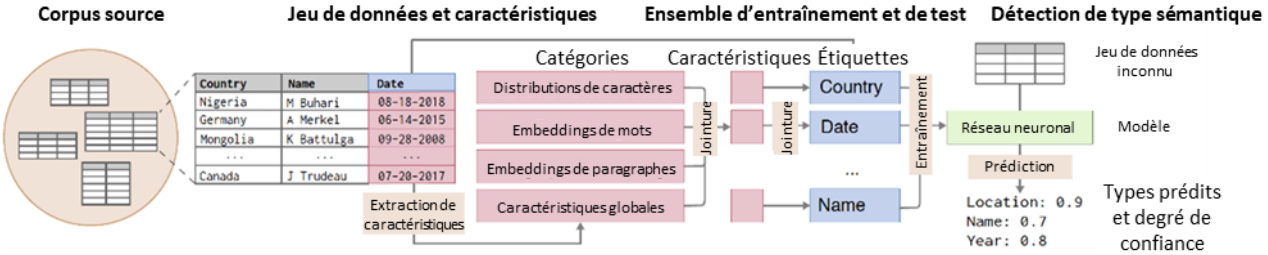


Figure 33 Architecture générale de Sherlock

(Hulsebos et al., 2019) ont démontré la supériorité de Sherlock en termes de précision et de robustesse face aux données bruitées, surpassant les références d'apprentissage automatique, les approches basées sur les dictionnaires et les expressions régulières, ainsi que le consensus des annotations crowdsourcées. Cette performance est particulièrement pertinente dans le contexte des données industrielles, souvent caractérisées par leur hétérogénéité et leur imperfection (L. D. Xu et al., 2014).

L'approche d'extraction de caractéristiques généralisée de Sherlock présente un intérêt particulier pour le traitement des données de l'Internet des Objets Industriel (IIoT). Sa capacité à extraire des caractéristiques robustes à partir de données variées offre une solution prometteuse pour structurer et interpréter ces données complexes. La combinaison des statistiques globales, des distributions de caractères agrégées, des embeddings de mots et des vecteurs de paragraphes permet à Sherlock de capturer une large gamme de propriétés des données, ce qui est particulièrement pertinent pour les données IIoT souvent désorganisées.

Cependant, les environnements IIoT génèrent fréquemment des données non significatives ou parasites, comme l'ont souligné (Qin et al., 2016) dans leur étude sur la qualité des données IIoT. Ces "données bruitées" peuvent inclure des colonnes vides, des métadonnées non pertinentes pour l'analyse sémantique, ou des informations illisibles ou corrompues, constituant un défi majeur pour l'extraction d'informations utiles.

Bien que la méthodologie de Sherlock offre une base solide, des adaptations seront nécessaires pour traiter efficacement ces spécificités des données IIoT. Cela pourrait impliquer le développement de filtres préliminaires pour identifier et exclure les données non pertinentes, l'ajout de caractéristiques spécifiques pour détecter les schémas typiques des données IIoT non significatives, ou l'adaptation des embeddings pour mieux refléter le vocabulaire technique industriel. Ces ajustements, s'inspirant des travaux de (Stojanovic et al., 2016) sur le prétraitement des données IIoT, seront cruciaux pour exploiter pleinement le potentiel de Sherlock dans le contexte des jumeaux numériques industriels.

Bien que Sherlock représente une avancée significative dans la détection des types sémantiques, il présente certaines limitations. Notamment, ses performances sont réduites pour les types peu représentés dans les données d'entraînement, et il ne prend pas en compte le contexte global du tableau lors de la prédiction. Ces limitations ont motivé le développement d'approches plus avancées, comme Sato, qui vise à surmonter ces défis en intégrant des informations contextuelles plus riches.

4.2.3.2.2 Sato

(D. Zhang et al., 2020) ont introduit Sato, un modèle d'apprentissage automatique hybride conçu pour détecter automatiquement les types sémantiques des colonnes dans les tableaux. Sato se distingue par son exploitation simultanée du contexte global et local, ce qui lui permet d'atteindre une précision de prédiction remarquablement élevée. Le modèle hybride de Sato régule la prédiction des types sémantiques en utilisant des signaux provenant du contexte global (valeurs de l'ensemble du tableau) et du contexte local (types prédits des colonnes voisines).

Sato aborde les problèmes identifiés dans Sherlock en intégrant les contextes de table pour prédire les types sémantiques des colonnes. Il combine la modélisation de sujets et l'apprentissage structuré avec la prédiction de type de colonne unique basée sur le modèle Sherlock. À l'instar de Sherlock, Sato a été entraîné sur le corpus VizNet de l'ensemble de données

WebTables, en considérant les mêmes 78 types sémantiques communs, ce qui permet une comparaison directe des performances.

Le modèle Sato utilise une approche novatrice basée sur un modèle de sujet pour estimer l'intention globale d'une table. Cette approche repose sur l'idée fondamentale que chaque table est créée avec une intention spécifique, qui peut aider à déterminer les types sémantiques des colonnes. Comme l'ont souligné (Blei et al., 2003) dans leurs travaux, les modèles de sujets peuvent efficacement capturer les thèmes sous-jacents dans un corpus de documents, ou dans ce cas, dans un ensemble de tables.

Pour estimer l'intention de la table, Sato utilise un modèle de sujet pré-entraîné qui traite les valeurs de chaque table comme un "document" et produit un vecteur de sujet pour la table. Ce modèle est formé sur des tables publiques dont les en-têtes et les légendes ont été supprimés, ce qui permet une généralisation robuste à travers différents domaines et structures de tables.

Une fois le vecteur de sujet calculé, il est utilisé comme entrée supplémentaire pour le modèle de prédiction de colonne unique. Ce modèle est un réseau neuronal profond qui utilise à la fois les caractéristiques de la colonne et le vecteur de sujet pour prédire le type sémantique de la colonne. Cette approche s'inspire des travaux de (Mikolov, Sutskever, et al., 2013) sur les word embeddings, en adaptant le concept à l'échelle des tables et des colonnes.

Sato capture également le contexte local d'une colonne en utilisant un modèle de prédiction de sortie structuré. Les auteurs soulignent l'importance de prendre en compte les relations entre les colonnes lors de la prédiction des types sémantiques, arguant que les types de colonnes ne sont pas indépendants les uns des autres. Cette approche s'aligne avec les travaux de (Lafferty et al., 2001) sur les champs aléatoires conditionnels (CRF), qui ont démontré l'efficacité de la modélisation des dépendances entre variables de sortie dans les tâches de prédiction structurée.

Pour capturer ces relations, Sato utilise un CRF en chaîne linéaire pour modéliser les dépendances entre les types de colonnes. Le CRF prend en compte à la fois les caractéristiques de chaque colonne et les relations entre les colonnes pour prédire les types sémantiques. L'objectif est de trouver la séquence de types sémantiques qui maximise la probabilité conditionnelle donnée par le CRF, une approche qui s'est avérée efficace dans divers domaines du traitement du langage naturel et de l'apprentissage automatique (C. Sutton & McCallum, 2012).

Grâce à cette architecture sophistiquée, Sato surpasse de manière significative l'état de l'art en matière de prédiction de types sémantiques. Ces gains de performance sont principalement dus à des prédictions améliorées pour les types sémantiques sous-représentés, adressant ainsi une des principales limitations de Sherlock. De plus, Sato introduit une nouvelle architecture extensible pour la détection de types avec des modules pour la modélisation de colonnes uniques, le contexte global et le contexte local.

L'approche de Sato, bien qu'innovante et performante pour les données tabulaires structurées, soulève des questions quant à son applicabilité directe aux données de l'Internet des Objets (IoT), en particulier dans le contexte des jumeaux numériques industriels. Les données IoT sont souvent caractérisées par leur nature désorganisée, leur volume important et leur variabilité temporelle, comme l'ont souligné (Atzori et al., 2010) et (Gubbi et al., 2013). Ces caractéristiques contrastent fortement avec les tableaux bien structurés sur lesquels Sato a été entraîné et testé.

L'adaptation des améliorations apportées par Sato, notamment l'utilisation du contexte global et local, pourrait s'avérer complexe pour les données IoT. Par exemple, la notion de "table" et de "colonnes voisines" peut être moins pertinente ou difficile à définir dans un flux de données IoT. De plus, l'estimation de l'intention globale via la modélisation de sujets pourrait être moins efficace sur des données en temps réel et hétérogènes. Ces défis suggèrent que, bien que Sato offre des perspectives intéressantes, des modifications substantielles ou une approche entièrement nouvelle pourraient être nécessaires pour traiter efficacement les données IoT dans le cadre de l'enrichissement des jumeaux numériques industriels.

4.2.3.3 Conclusion

Les graphes de connaissances, tels que SiMa, ont montré leur efficacité dans certains contextes. En effet, ces outils sont particulièrement utiles lorsque les données ont déjà été triées à un certain degré. Ils peuvent être réutilisés dans d'autres contextes, ou pour établir des liens une fois la première ontologie générée. Cependant, ces outils se révèlent moins adaptés pour traiter les données IoT brutes. Ces données sont souvent désorganisées et contiennent une grande quantité d'informations inutiles, redondantes ou pertinentes uniquement du point de vue IT.

C'est pourquoi nous nous orientons vers l'approche de la reconnaissance de type sémantique. L'approche sémantique de Sato, bien qu'elle présente des avancées significatives, souffre également de limitations similaires : les tableaux IoT contenant des données brutes rendent la reconnaissance de contexte peu utile pour nos besoins spécifiques.

Nous avons donc choisi de nous orienter vers l'approche sémantique de Sherlock. Ce modèle promet une meilleure adaptabilité et une plus grande efficacité pour traiter les défis spécifiques posés par les données IoT brutes. Il offre une approche plus adaptée à notre contexte, permettant une meilleure gestion de la complexité et de l'imprécision inhérentes aux données industrielles.

Pour pallier les limites de Sherlock dans le contexte de l'IIoT, nous envisageons trois adaptations principales :

1. L'ajout d'une opération de reconnaissance de types sémantiques clés : cette étape manuelle permettra d'identifier et de prioriser les types de données les plus pertinents pour notre contexte industriel spécifique. Comme l'ont souligné (Stojanovic et al., 2016), cette approche peut significativement améliorer la précision de la classification pour les types de données critiques dans un environnement IIoT.
2. L'utilisation de reconnaissances de type spécifique : avant d'appliquer Sherlock, nous implémenterons une étape préliminaire pour reconnaître si une donnée est bruitée ou inutilisable. Cette approche s'inspire des travaux de (Qin et al., 2016) sur le filtrage des données IIoT et permettra d'améliorer la qualité des données traitées par Sherlock.
3. L'intégration à l'ontologie pour réinjecter la donnée : une fois les types sémantiques identifiés, nous développerons un mécanisme pour intégrer ces informations dans notre ontologie existante. Cette étape permettra de contextualiser les données IIoT au sein de notre modèle de connaissance plus large, facilitant ainsi leur utilisation dans le jumeau numérique.

Ces adaptations permettront de surmonter les principales limitations de Sherlock dans notre contexte IIoT spécifique. En combinant la puissance de reconnaissance de type sémantique de

Sherlock avec ces améliorations ciblées, nous visons à créer un système robuste et efficace pour l'enrichissement sémantique des jumeaux numériques industriels.

Nous proposons une approche en trois étapes, inspirée de l'architecture de Sherlock, mais enrichie par un travail ontologique plus approfondi. Cette méthodologie vise à surmonter les défis de l'intégration des données hétérogènes dans un environnement industriel complexe, comme l'ont souligné (Kadadi et al., 2014) dans leur étude sur les défis de l'intégration des big data.

4.2.4 Méthodologie d'intégration de données multi-sources

L'analyse approfondie de l'état de l'art et l'étude du cas d'application Airbus nous ont permis d'élaborer une méthodologie d'intégration de données multi-sources structurée en trois phases principales. Cette approche vise à répondre de manière cohérente et exhaustive à nos quatre questions de recherche, en abordant les défis spécifiques liés à l'enrichissement des jumeaux numériques dans un contexte industriel complexe.

Notre méthodologie se décompose comme suit :

1) Phase d'ontologification :
   Cette première phase répond principalement aux questions de recherche 1 et 2. Elle se concentre sur l'extraction des caractéristiques pertinentes des données brutes à l'aide de techniques d'IA, notamment en s'appuyant sur l'architecture de Sherlock pour la détection des types sémantiques. L'objectif est de convertir efficacement les données non structurées en informations exploitables, tout en prenant en compte les particularités des données IIoT. Cette phase aboutit à la génération automatisée d'une ontologie à partir des données structurées.
2) Phase d'alignement :
   La deuxième phase se focalise sur la question de recherche 3. L'ontologie générée lors de la phase précédente sert de structure de base pour organiser et relier les informations provenant des différentes sources de données. Nous exploitons des techniques d'apprentissage automatique pour construire et aligner cette ontologie de manière automatisée et efficace. Cette phase vise à surmonter les défis d'intégration des données non structurées et à prendre en compte la spécificité de l'intégration des données IIoT.
3) Phase d'exploration :
   La dernière phase répond à la question de recherche 4. Nous exploitons l'architecture ontologique créée pour fusionner les données issues des logiciels métiers avec celles provenant des 'datalakes'. L'objectif est de construire un graphe de connaissances unifié et complet, permettant une exploration et une analyse approfondies des informations relatives aux produits et aux processus d'Airbus. Cette architecture ontologique globale offre une vue d'ensemble cohérente et facilite la prise de décision basée sur des données intégrées et enrichies, rendant ainsi les informations accessibles aux acteurs métier pour générer de nouvelles connaissances.

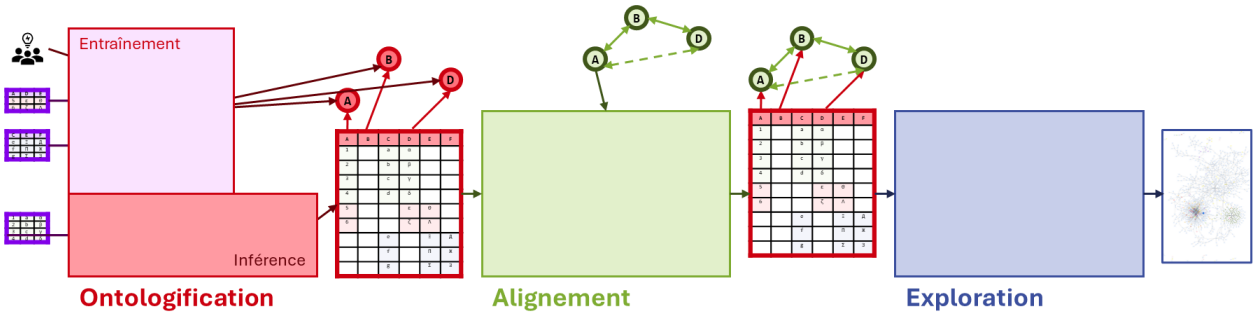


Figure 34 Schéma de la méthodologie complète

Cette méthodologie en trois phases, représentée Figure 34, offre une approche structurée et complète pour l'enrichissement des jumeaux numériques, abordant l'ensemble du processus de transformation des données brutes à l'exploitation des connaissances, tout en prenant en compte les spécificités de l'environnement industriel multi-modèle et multi-métiers d'Airbus.

4.3 Phase d'ontologification

Pour mieux appréhender notre approche, il est crucial de présenter l'architecture globale de notre méthodologie, qui s'appuie sur l'intelligence artificielle pour automatiser et optimiser le processus d'ontologification. Cette approche, basée sur l'apprentissage profond, se divise en deux étapes principales : l'entraînement et l'inférence.

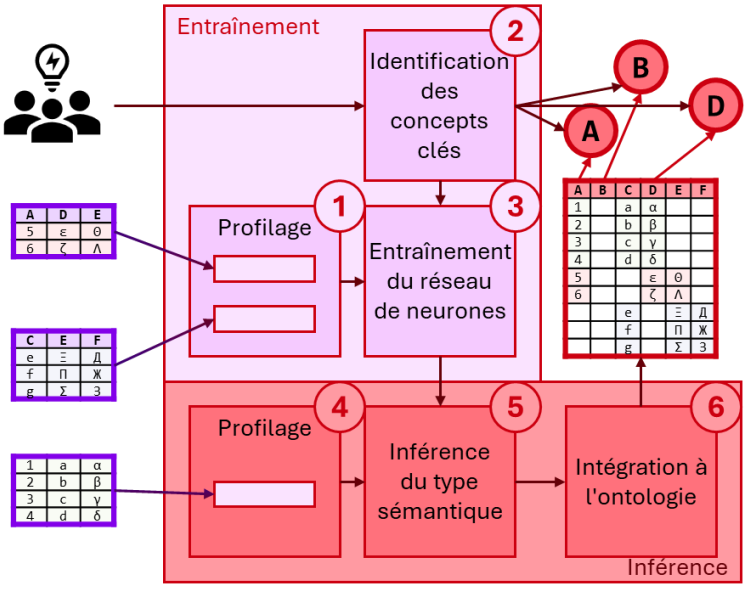


Figure 35 Schéma de la phase d'ontologification

Comme illustré dans la Figure 35, notre phase d'ontologification se compose de deux étapes principales : une étape **d'entraînement** et une étape de détection (inférence).

L'étape **d'entraînement** comprend trois activités clés :

1) Profilage : nous procédons au profilage des données. Cette activité consiste à caractériser les données en un vecteur de taille fixe pour qu'elles puissent servir à entraîner le système d'intelligence artificielle (IA). Nous utilisons des techniques avancées de profilage de données

pour obtenir une vue d'ensemble des caractéristiques des données et assurer leur adéquation pour l'entraînement de notre modèle.

2) Identification des concepts clés : en parallèle, les données sources sont analysées pour identifier les concepts essentiels qui serviront de base à l'ontologie de domaine à construire. Cette activité s'appuie sur une approche semi-automatique, combinant l'expertise humaine et des techniques d'analyse automatique, pour extraire les types sémantiques et comprendre leur importance dans le contexte des données industrielles.
3) Entraînement du réseau de neurones : avec les concepts clés identifiés et les données profilées, nous passons à l'entraînement de notre système d'IA. Nous utilisons des algorithmes d'apprentissage automatique, tels que des réseaux de neurones profonds, pour entraîner notre modèle à reconnaître les types sémantiques et à établir des liens entre les données et l'ontologie.

L'étape **de détection**, quant à elle, se compose également de trois activités :

4) Profilage : tout comme lors de l'entraînement, la première étape consiste à profiler les nouvelles données entrantes. Nous appliquons les mêmes techniques de profilage pour extraire un ensemble riche de caractéristiques de ces données. Cette étape nous permet de transformer les données brutes, souvent hétérogènes et non structurées caractéristiques de l'IoT industriel, en un format compatible avec notre modèle entraîné. Cette étape de profilage est essentielle pour comprendre la structure, la qualité et les particularités des données avant leur utilisation dans des tâches d'analyse avancées.
5) Inférence du type sémantique (détection) : une fois les nouvelles données profilées, nous les soumettons à notre modèle d'apprentissage automatique entraîné. Le modèle analyse les caractéristiques extraites et prédit le type sémantique le plus probable pour chaque donnée. C'est ici que le modèle met en pratique sa capacité à reconnaître les motifs et les relations appris lors de l'entraînement. Cette approche s'inspire des travaux sur l'apprentissage profond, permettant une classification efficace des types sémantiques même pour des données complexes et variées.
6) Intégration à l'ontologie : après avoir identifié les types sémantiques, nous procédons à l'intégration des données dans notre ontologie. Chaque donnée est associée au concept correspondant dans l'ontologie, en fonction de son type sémantique prédit. Cette étape permet d'enrichir notre base de connaissances avec de nouvelles informations, tout en maintenant une structure cohérente et organisée.

Cette méthodologie en deux étapes nous permet de construire et d'affiner progressivement l'ontologie résultante, tout en l'alimentant avec de nouvelles données. L'étape d'entraînement nous permet de définir les fondements de notre système d'IA, tandis que l'étape d'inférence nous permet d'appliquer ce système à de nouvelles données entrantes, enrichissant ainsi continuellement notre base de connaissances. Cette approche itérative s'inspire des travaux de (N. F. Noy & McGuinness, 2001) sur le développement d'ontologies évolutives et des principes d'apprentissage continu décrits par (Z. Chen & Liu, 2018).

Dans les sections suivantes, nous détaillerons chacune de ces activités et expliquerons comment elles s'intègrent pour former un système cohérent et efficace d'intégration de données multisources dans un contexte industriel complexe.

4.3.1 Étape Entraînement

La première étape de notre méthodologie, l'entraînement, est cruciale pour la construction d'un modèle robuste et efficace d'intégration des données IoT dans le contexte des jumeaux numériques industriels. Cette étape se décompose en trois activités principales : l'identification des types sémantiques clés, le profilage des données, et l'entraînement des modèles d'intelligence artificielle. Cette approche adapte les travaux de (Hulsebos et al., 2019) sur la détection automatique des types sémantiques, aux spécificités de notre environnement industriel complexe.

Pour atteindre notre objectif d'intégration efficace des données IoT, nous sommes confrontés à deux défis majeurs initiaux. Premièrement, la grande diversité des types de données rend difficile l'utilisation directe de l'IA. Comme l'ont souligné (Atzori et al., 2010) et (Gubbi et al., 2013), les données brutes provenant de l'IoT sont souvent hétérogènes et non structurées, ce qui complique leur traitement direct par des algorithmes d'IA. Pour surmonter ce défi, nous proposons une activité de profilage des données, s'inspirant des techniques avancées de profilage proposées par (Naumann, 2014). Cette activité vise à normaliser et à transformer les données brutes en un format exploitable par les modèles d'IA, extrayant des caractéristiques pertinentes telles que les types de données, les distributions statistiques, et les relations entre les différents attributs.

Le second défi concerne la variété sémantique des données. Comme l'ont noté (L. D. Xu et al., 2014) dans leur étude sur la qualité des données IoT, parmi la multitude de données collectées par les capteurs IoT, toutes ne sont pas pertinentes pour l'intégration avec des modèles connus. Pour résoudre ce problème, nous proposons une activité d'identification des types sémantiques clés. Cette étape, s'inspirant des travaux de (Gruber, 1995) sur la conception d'ontologies pour le partage de connaissances, consiste à analyser les données profilées et à identifier les concepts essentiels qui serviront de base à notre ontologie.

Une fois ces deux activités préliminaires réalisées, nous procédons à la construction de notre modèle d'IA. Nous nous appuyons sur les résultats de notre état de l'art en intelligence artificielle, en utilisant des algorithmes particulièrement adaptés à la reconnaissance des types sémantiques des données pertinentes. Ces algorithmes, basés sur les réseaux de neurones profonds décrits par (LeCun et al., 2015), sont capables d'apprendre à partir des données profilées et des types sémantiques clés identifiés, établissant ainsi des correspondances entre les données IoT et les concepts de notre ontologie.

L'architecture de l'étape d'entraînement, décrite par la Figure 36, s'articule donc de la manière suivante :

1) Profilage : nous procédons au profilage des données. Cette activité consiste à caractériser les données en un vecteur de taille fixe pour qu'elles puissent servir à entraîner le système d'intelligence artificielle (IA). Nous utilisons des techniques avancées de profilage de données pour obtenir une vue d'ensemble des caractéristiques des données et assurer leur adéquation pour l'entraînement de notre modèle.
2) Identification des concepts clés : en parallèle, les données sources sont analysées pour identifier les concepts essentiels qui serviront de base à l'ontologie de domaine à construire.

Cette activité s'appuie sur une approche semi-automatique, combinant l'expertise humaine et des techniques d'analyse automatique, pour extraire les types sémantiques et comprendre leur importance dans le contexte des données industrielles.

3) Entraînement du réseau de neurones : avec les concepts clés identifiés et les données profilées, nous passons à l'entraînement de notre système d'IA. Nous utilisons des algorithmes d'apprentissage automatique, tels que des réseaux de neurones profonds, pour entraîner notre modèle à reconnaître les types sémantiques et à établir des liens entre les données et l'ontologie.

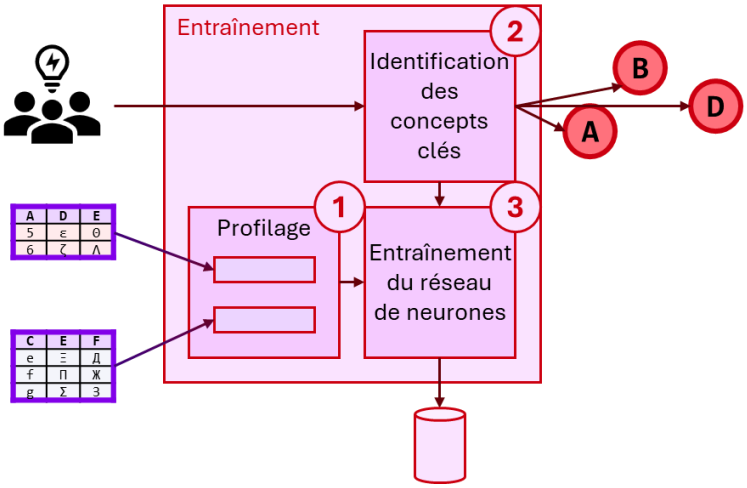


Figure 36 Architecture de l'étape d'entraînement

Cette étape d'entraînement en trois activités nous permet de construire un modèle d'IA robuste et efficace pour l'intégration de données IoT. En partant des données brutes, en les profilant et en identifiant les types sémantiques clés, nous préparons le terrain pour un entraînement efficace des algorithmes d'IA. Cette approche structurée et méthodique, s'inspirant des travaux de (Tao, Zhang, Liu, et al., 2019) sur l'intégration des données dans les jumeaux numériques, nous permet de surmonter les défis liés à l'hétérogénéité et à la variété sémantique des données IoT, ouvrant ainsi la voie à une intégration automatisée et intelligente des données dans notre ontologie.

4.3.1.1 Profilage des données

Le profilage des données constitue une étape cruciale dans la préparation des données pour l'entraînement des modèles d'apprentissage automatique. Cette étape vise à examiner en détail les données disponibles et à en extraire des caractéristiques pertinentes qui serviront à entraîner le modèle. Le profilage des données est essentiel pour comprendre la structure, la qualité et les particularités des données avant leur utilisation dans des tâches d'analyse avancées.

Dans notre approche, nous nous appuyons sur l'architecture du réseau de neurones du modèle Sherlock, comme illustré dans la Figure 37. Cette figure détaille l'architecture utilisée pour l'étape 3 de la Figure 33, présentant une approche sophistiquée pour l'extraction et le traitement des caractéristiques des données.

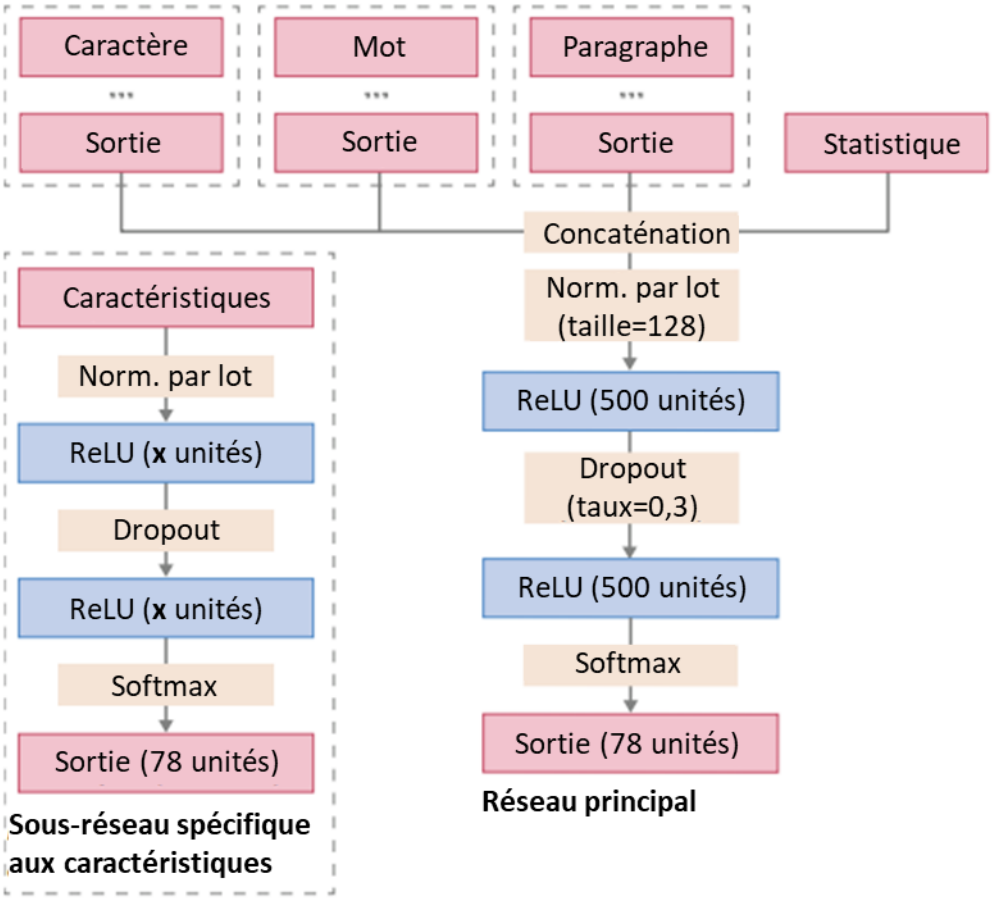


Figure 37 Architecture du réseau de neurones de Sherlock

Le modèle Sherlock offre un cadre particulièrement adapté au traitement et à l'analyse des données IoT, car sa flexibilité permet de gérer efficacement la diversité et l'hétérogénéité des données générées par les capteurs et les dispositifs IoT.

Cependant, les particularités des données IoT, telles que leur volume massif, leur diversité sémantique, leur variabilité temporelle et leur vélocité de génération, nécessitent des ajustements spécifiques à notre architecture. Nous prévoyons donc d'adapter l'architecture de Sherlock pour qu'elle puisse gérer ces défis de manière plus efficace. Cela peut impliquer des modifications dans la manière dont le modèle traite, filtre et apprend à partir des données IoT. L'objectif est d'élaborer un réseau de neurones capable non seulement de manipuler des données industrielles complexes, mais également de s'adapter et d'apprendre de manière autonome à mesure que de nouvelles données sont introduites dans le système.

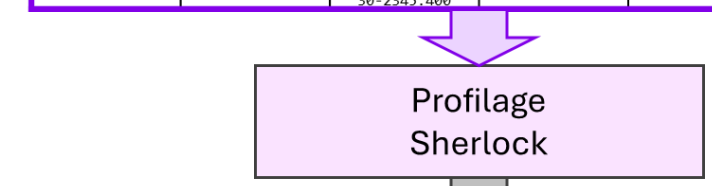
Le profilage des données est une étape cruciale dans la préparation des données pour l'entraînement des modèles d'apprentissage automatique. Il s'agit d'un processus qui consiste à examiner en détail les données disponibles et à en extraire des caractéristiques ou "features" pertinentes qui seront utilisées pour entraîner le modèle. En d'autres termes, le profilage des données est le processus de compréhension profonde et d'analyse des données à notre disposition.

Dans notre première approche, nous reprenons le profilage des données tel que mis en œuvre par Sherlock. L'approche de Sherlock pour le profilage des données est particulièrement adaptée

aux données mixtes et permet d'extraire un ensemble riche de caractéristiques des données. Comme décrit précédemment, nous examinerons les distributions de caractères, les embeddings de mots, les embeddings de paragraphes et les caractéristiques globales des données. Cette approche multidimensionnelle nous permettra de capturer les différents aspects des données IoT et de fournir une représentation complète et informative au modèle d'apprentissage.

Ces différentes caractéristiques seront ensuite agrégées pour former un profil de données, qui est essentiellement une représentation condensée et caractéristique des données. Ce profil de données contiendra 1588 caractéristiques, un nombre suffisant pour capturer la complexité et la variété des données sans être trop volumineux pour être géré efficacement. Un exemple de profilage sur des données réelles est présent dans la Figure 38.

id	msn	name	predecessors	successors	station
1000	09234	D100-80020-000-20-1234:100	3056	1001	station45
1001	09234	D100-80020-000-20-1234:110	1000	1002	station45
1002	09234	D100-80020-000-20-1234:120	1001	1003	station45
1005	09234	D100-91030-000-30-2345:200	1004	1006	station18
1006	09234	D100-91030-000-30-2345:300	1007#1006	1005	station18
1007	09234	D100-91030-000-30-2345:400	1005	1008#1009#1010	station18



Concept	Nombre moyen de '0'	Présence systématique de 's'	Entropie de la colonne	Fraction de cellules uniques	Longueur moyenne
id	1,25	0	8,1	1	4
msn	1	0	0	0	6
name	6,8	0	8,05	1	20,5
predecessors	0,35	0	7,35	0,85	4,3
successors	0,4	0	7,5	0,88	4,8
station	0,2	1	2,8	0,03	9

Figure 38 Exemple de données d'entraînement avec le profilage

Il est à noter que ce profilage est effectué pour toutes les colonnes des données, y compris celles qui peuvent sembler non pertinentes ou superflues à première vue. Cette approche exhaustive garantit que nous disposons d'un ensemble de données d'entraînement aussi riche et diversifié que possible, ce qui est essentiel pour l'efficacité de l'entraînement du modèle.

Cette approche exhaustive, recommandée par (Batini et al., 2009) dans leurs travaux sur la qualité des données, garantit un ensemble de données d'entraînement riche et diversifié, essentiel pour l'efficacité de l'entraînement du modèle. En effet, cette représentation riche des données est cruciale pour l'efficacité des modèles d'apprentissage profond.

L'approche de Sherlock pour le profilage des données s'avère particulièrement adaptée aux défis posés par l'Internet des Objets (IoT). Les données IoT sont caractérisées par leur volume massif, leur vélocité élevée et leur grande variété, ce qui soulève des défis uniques en matière de gestion

et d'analyse. L'approche de Sherlock, qui peut être mise en œuvre pour réaliser un profilage "on edge", offre une solution prometteuse à ces défis.

Le concept de profilage "on edge", dérivé de l'edge computing (ou informatique en périphérie), représente une stratégie innovante dans le traitement des données IoT. Comme l'expliquent (Shi et al., 2016) dans leur étude sur l'edge computing, cette approche consiste à traiter les données à proximité de leur source, plutôt que de les centraliser dans un système distant. Concrètement, cela signifie que l'analyse et le traitement des données s'effectuent directement sur les appareils IoT ou sur un serveur local proche.

Le profilage "on edge" présente plusieurs avantages significatifs dans le contexte de l'IoT :

1) Économie de bande passante : dans l'environnement IoT, où la quantité de données générées peut être colossale, le transfert de ces données pour analyse ou stockage peut consommer une bande passante considérable. Comme l'ont démontré (Bonomi et al., 2012) dans leurs travaux sur le fog computing, en effectuant le profilage directement sur l'appareil IoT, on évite le besoin de transférer les données brutes, réduisant ainsi significativement le trafic réseau. Cette approche se traduit par des économies substantielles de bande passante, un facteur crucial dans le contexte de l'IoT où les ressources réseau peuvent être limitées ou coûteuses.
2) Renforcement de la sécurité des données : le profilage "on edge" contribue à améliorer la sécurité des données. Comme l'ont souligné (Roman et al., 2018) dans leur étude sur la sécurité de l'edge computing, lorsque les données sont transférées, elles peuvent être vulnérables à diverses formes d'attaques, telles que l'interception ou le piratage. Ces risques sont particulièrement préoccupants dans le contexte de l'IoT, où les appareils peuvent être dispersés géographiquement et où la sécurité peut être plus difficile à garantir. En réalisant le profilage des données directement sur l'appareil, on minimise le besoin de transfert de données et, par conséquent, on réduit l'exposition des données à ces risques. Cette approche constitue une mesure de sécurité essentielle pour protéger les données sensibles et garantir la confiance dans les systèmes IoT.
3) Amélioration de l'efficacité du processus : le profilage "on edge" optimise l'efficacité du processus de gestion des données. Comme l'ont démontré (Satyanarayanan et al., 2009) dans leurs travaux pionniers sur le cloudlet, dans le cadre traditionnel des algorithmes d'IA, les données doivent être transférées vers une infrastructure centralisée pour être analysées, ce qui peut prendre du temps, surtout lorsque les volumes de données sont importants. Avec le profilage "on edge", cette étape est éliminée car les données sont profilées là où elles sont générées. Cela permet un profilage plus rapide et plus réactif, une amélioration significative de l'efficacité du processus, particulièrement pertinente dans le contexte de l'IoT où les données sont souvent générées en temps réel et où une réactivité rapide peut être cruciale.
4) Réduction des silos de données : en profilant "on edge", on élimine la nécessité d'établir des silos de données supplémentaires pour le stockage et l'analyse. Comme l'ont souligné (Philip Chen & Zhang, 2014) dans leur étude sur la gestion des big data, les données sont traitées et profilées là où elles sont générées, ce qui évite la création de copies multiples de ces données dans différents emplacements. Cela contribue non seulement à l'économie de la bande passante, mais également à la simplification de la gestion des données, un aspect crucial dans les environnements IoT complexes.

L'approche de Sherlock, permettant le profilage "on edge", offre ainsi une solution adaptée aux défis posés par l'IoT. En évitant le transfert des données brutes, elle améliore la sécurité, économise la bande passante, optimise l'efficacité du processus et simplifie la gestion des données. Ces avantages sont particulièrement bénéfiques pour un traitement et une analyse efficaces des données IoT.

4.3.1.2 Identification des types sémantiques clés

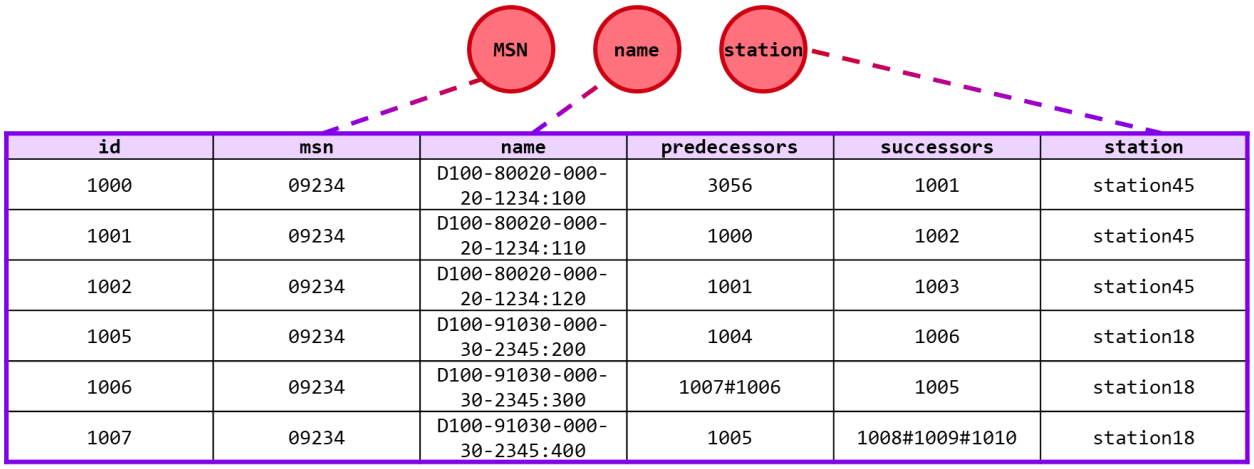
La création d'ontologies de manière totalement automatique à partir de données brutes reste un défi complexe dans le domaine de l'ingénierie des connaissances. Pour faciliter ce processus et guider le système, nous proposons une approche semi-automatique impliquant l'utilisateur dans l'identification manuelle des concepts essentiels. Cette étape préliminaire, réalisée en parallèle du profilage des données, permet de filtrer les informations pertinentes et de construire une ontologie d'interface servant de référence pour l'entraînement du modèle d'apprentissage automatique.

Cette approche s'inspire des travaux sur le développement d'ontologies, où l'expertise humaine joue un rôle crucial dans la définition initiale des concepts clés. Ainsi, c'est l'utilisateur qui est chargé d'identifier les concepts correspondant à des types sémantiques primordiaux pour le domaine d'application considéré. Ces concepts peuvent inclure, par exemple, l'identification d'avions, de poids, de tailles, de noms d'utilisateurs ou encore de types d'opérations.

Les concepts clés identifiés par l'utilisateur sont alors rassemblés pour former une ontologie spécifique, que nous appelons « ontologie d'interface ». Il est important de souligner que cette ontologie d'interface, bien que centrée sur les concepts clés, ne contient aucun lien entre ces concepts. Son rôle est principalement de servir de référence lors de l'interprétation des données et lors de la construction de l'ontologie finale. Cette approche s'aligne avec les travaux sur la conception d'ontologies pour le partage de connaissances.

Le principal objectif de cette étape d'identification est d'opérer un filtrage initial permettant d'éliminer les concepts inutiles qui pourraient encombrer notre ontologie finale, et de ne conserver que ceux qui sont essentiels à la représentation du domaine d'intérêt. Cette considération aura une importance capitale lors de l'entraînement de notre modèle de machine learning, en raison du déséquilibre potentiel entre les classes - les concepts pertinents d'une part, et une classe « ignoré » regroupant les concepts inutiles d'autre part. Cette problématique du déséquilibre des classes a été largement étudiée dans le domaine de l'apprentissage automatique, notamment par (Haibo He & Garcia, 2009).

La Figure 39 schématise notre approche, où certaines colonnes sont reliées aux concepts ontologiques qu'elles contiennent. Ces concepts sont pour l'instant définis manuellement par l'utilisateur.



id	msn	name	predecessors	successors	station
1000	09234	D100-80020-000-20-1234:100	3056	1001	station45
1001	09234	D100-80020-000-20-1234:110	1000	1002	station45
1002	09234	D100-80020-000-20-1234:120	1001	1003	station45
1005	09234	D100-91030-000-30-2345:200	1004	1006	station18
1006	09234	D100-91030-000-30-2345:300	1007#1006	1005	station18
1007	09234	D100-91030-000-30-2345:400	1005	1008#1009#1010	station18

Figure 39 Illustration de l'identification des types sémantiques clés

La nécessité d'une telle démarche d'identification s'explique par la nature même des données IoT. Ces données sont souvent noyées dans une multitude de concepts superflus. Certains champs peuvent être redondants, d'autres peuvent être des doublons, des copies avant modification, contenir des opérations, être vides ou encore contenir des commentaires. Tous ces éléments peuvent introduire un bruit indésirable et rendre l'analyse des données plus complexe et moins efficace, ce qui aura pour effet de réduire la qualité de l'ontologie résultante.

En demandant à l'utilisateur d'identifier les concepts clés, nous mettons en place une première phase de tri. Ce processus nous permet de filtrer les concepts superflus et de ne conserver que ceux qui sont essentiels à la compréhension du domaine. Cela garantit que notre modèle est entraîné de manière efficace et précise, comme l'ont démontré (Domingos, 2012) dans ses travaux sur l'importance de la sélection des caractéristiques dans l'apprentissage automatique.

Cette étape d'identification des concepts clés par l'utilisateur est donc fondamentale pour assurer la pertinence et l'efficacité de notre approche d'analyse des données IoT, et pour garantir la construction d'une ontologie finale qui soit véritablement représentative du domaine d'application. Cette approche s'inscrit dans la lignée des travaux de (Paulheim, 2016) sur l'enrichissement sémantique automatique des données.

4.3.1.3 Entraînement du réseau de neurones

L'entraînement des modèles d'apprentissage automatique constitue une étape cruciale de notre méthodologie de profilage. Bien que la majeure partie du travail préparatoire soit réalisée lors de la phase de profilage, c'est durant l'entraînement des modèles que la précision et l'efficacité de notre approche sont véritablement mises en œuvre et optimisées.

Pour initier cette activité, nous nous sommes appuyés sur l'architecture proposée par Sherlock. Cette architecture, illustrée précédemment dans la Figure 37, nous a servi de fondement pour l'entraînement de nos modèles. L'approche de Sherlock offre une base solide pour le traitement des données hétérogènes, particulièrement adaptée aux défis posés par les données de l'Internet des Objets (IoT) dans un contexte industriel.

Cette approche permet l'utilisation de modèles de taille réduite pour l'entraînement, ce qui présente plusieurs avantages significatifs. Premièrement, ces modèles plus légers facilitent un

entraînement plus rapide et plus efficace, un aspect crucial dans le traitement des données IoT. Deuxièmement, ils sont particulièrement adaptés pour traiter les profils de données issus du processus de profilage précédent. En effet, bien que ces profils soient nombreux, ils n'appartiennent qu'à un nombre limité de classes grâce à la sélection préalable des concepts clés.

L'approche par profils nous permet d'utiliser un ensemble riche de caractéristiques tout en nous affranchissant des données sources initiales. Cette méthode s'inspire des word embeddings, où des représentations denses et informatives sont utilisées à la place des données brutes. Dans notre cas, les profils de données, basés sur Sherlock et contenant 1588 caractéristiques, servent de représentation condensée et informative des données originales. Cette approche offre l'avantage de réduire considérablement la quantité de données à traiter tout en préservant les informations essentielles pour la classification.

L'entraînement est réalisé sur toutes les classes de données, y compris une classe spéciale "ignoré" dédiée aux données jugées non pertinentes ou redondantes. L'inclusion de cette classe "ignoré" dans le processus d'entraînement est une stratégie importante pour améliorer la robustesse et l'efficacité du modèle. Ainsi, cette approche permet au modèle de mieux identifier et éliminer les données superflues lors du processus de profilage, améliorant ainsi la qualité globale de l'analyse.

La Figure 40 illustre un exemple de plusieurs profils provenant de colonnes identiques mais d'extraits de jeux de données différents. Cette visualisation met en évidence certaines similarités entre les profils, démontrant la capacité de notre approche à capturer des caractéristiques communes malgré la diversité des sources de données. Par exemple, on peut observer que les profils de "station" contiennent tous la lettre "S", tandis que les profils de "msn" (Manufacturing Serial Number) ont tous une longueur d'environ 6 caractères. Ces observations soulignent l'efficacité de notre méthode de profilage pour extraire des caractéristiques discriminantes, même à partir de données hétérogènes.

Concept	Nombre moyen de '0'	Présence systématique de 's'	Entropie de la colonne	Fraction de cellules uniques	Longueur moyenne
id	1,25	0	8,1	1	4
msn	1	0	0	0	6
name	6,8	0	8,05	1	20,5
predecessors	0,35	0	7,35	0,85	4,3
successors	0,4	0	7,5	0,88	4,8
station	0,2	1	2,8	0,03	9

Concept	Nombre moyen de '0'	Présence systématique de 's'	Entropie de la colonne	Fraction de cellules uniques	Longueur moyenne
id	1,25	0	8,1	1	3,97
msn	0,95	0	-0,03	0	6,01
name	6,81	0	8,02	0,99	20,45
predecessors	0,39	0	7,32	0,81	4,39
successors	0,32	0	7,46	0,93	4,9
station	0,2	1	2,86	0,01	9,03

Concept	Nombre moyen de '0'	Présence systématique de 's'	Entropie de la colonne	Fraction de cellules uniques	Longueur moyenne
id	1,28	0	8,14	1	4,09
msn	0,95	0	0,05	0,01	6,06
name	6,87	0	7,98	0,98	20,53
predecessors	0,45	0	7,36	0,89	4,32
successors	0,47	0	7,44	0,88	4,86
station	0,17	1	2,84	0	9,02

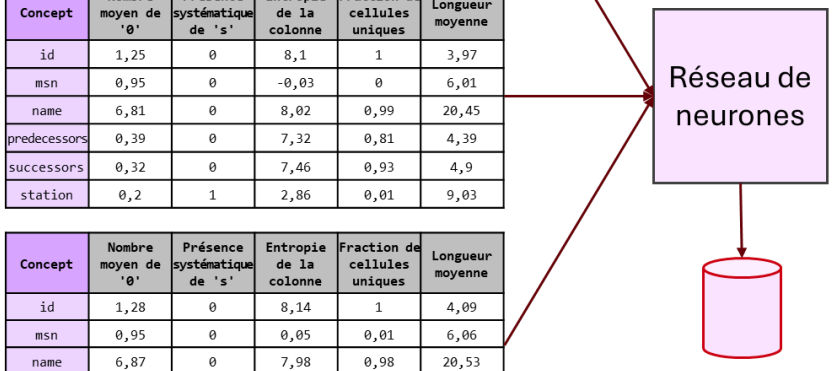


Figure 40 Exemple de profils de données pour l'entraînement du modèle d'apprentissage automatique

Cette approche d'entraînement, combinant l'utilisation de modèles légers, de profils de données riches en informations, et l'inclusion d'une classe "ignoré", nous permet de construire un système robuste et efficace pour la classification des types sémantiques dans un environnement IoT industriel complexe. Elle s'aligne avec les travaux sur l'enrichissement sémantique automatique des données, tout en étant spécifiquement adaptée aux défis uniques posés par les données IoT dans le contexte des jumeaux numériques industriels.

4.3.1.4 Synthèse

L'étape d'entraînement constitue un élément fondamental de notre méthodologie d'intégration des données IoT dans le contexte des jumeaux numériques industriels. Cette phase vise à élaborer un modèle robuste et efficace, capable d'identifier les types sémantiques des données et de les incorporer dans une ontologie cohérente. En effet, cette approche est essentielle pour améliorer l'interopérabilité et la performance des jumeaux numériques dans un environnement industriel complexe. Notre méthodologie se décompose en trois activités clés : le profilage des données, l'identification des types sémantiques clés, et l'entraînement des modèles.

Le profilage des données représente la première étape cruciale de notre approche. Il s'agit de transformer les données brutes, souvent hétérogènes et non structurées, caractéristiques de l'IoT industriel, en un format exploitable par les algorithmes d'apprentissage automatique. Nous nous appuyons sur l'approche de Sherlock qui permet d'extraire un profil riche de caractéristiques des données, incluant les distributions de caractères, les embeddings de mots et de paragraphes, ainsi que les caractéristiques globales. Cette approche multidimensionnelle capture efficacement la complexité et la variété des données IoT. De plus, la possiblité du profilage "on edge" offre des avantages significatifs en termes d'économie de bande passante, de sécurité des données et d'efficacité du processus.

L'identification des types sémantiques clés constitue la deuxième étape de notre méthodologie. Cette phase implique l'utilisateur dans la sélection manuelle des concepts essentiels du domaine d'application. Cette étape permet de filtrer les informations pertinentes et d'éliminer les concepts superflus. Les concepts clés identifiés sont rassemblés dans une ontologie d'interface servant de référence pour l'entraînement du modèle. Cette approche s'aligne avec l'état de l'art sur la conception d'ontologies pour le partage de connaissances, assurant ainsi une base solide pour notre modèle d'apprentissage automatique.

Enfin, l'entraînement des modèles constitue la dernière étape de notre processus. Nous nous appuyons sur l'architecture de Sherlock, en utilisant des modèles de petite taille pour un entraînement rapide et efficace sur les profils de données issus du processus de profilage. Cette approche s'inspire des word embeddings, où des représentations denses et informatives sont utilisées à la place des données brutes. L'entraînement est réalisé sur toutes les classes de données, y compris une classe "ignoré" pour les données inutiles ou redondantes, une stratégie qui permet d'améliorer la robustesse et l'efficacité du modèle face au déséquilibre des classes.

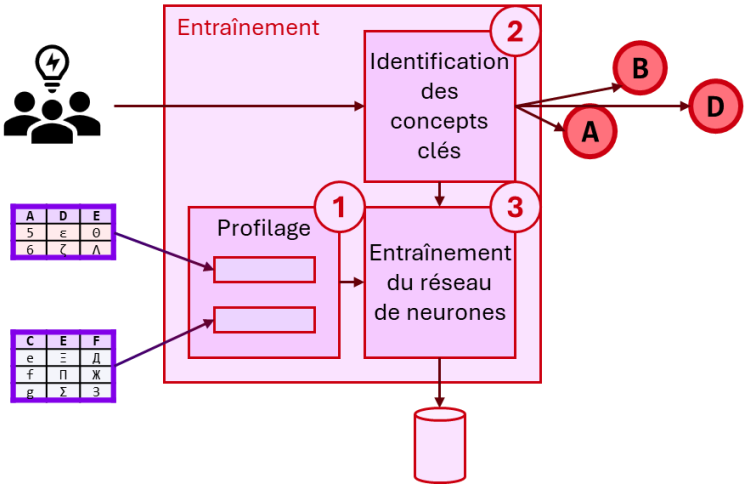


Figure 41 Architecture de l'étape d'entraînement

Cette figure illustre les trois étapes clés de notre phase d'entraînement : le profilage des données (1), l'identification des concepts clés (2), et l'entraînement du système d'IA (3). Les données brutes (représentées par les matrices à gauche) sont transformées en profils exploitables, tandis que l'identification des concepts clés (A, B, D) guide le processus d'entraînement. Cette architecture intégrée permet une approche holistique de l'apprentissage, essentielle pour traiter efficacement les données IoT complexes dans un contexte industriel.

En résumé, notre étape d'entraînement combine le profilage des données, l'identification des types sémantiques clés et l'entraînement des modèles pour construire un système capable de reconnaître et d'intégrer efficacement les données IoT dans une ontologie. Cette approche structurée et méthodique permet ainsi de surmonter les défis liés à l'hétérogénéité et à la variété sémantique des données IoT.

4.3.2 Étape Inférence (détection)

Une fois notre modèle entraîné, nous passons à l'étape d'inférence, également appelée étape de reconnaissance ou de prédiction. C'est lors de cette étape que notre système d'apprentissage automatique est confronté à de nouvelles données, jamais rencontrées auparavant, et doit mettre en pratique les connaissances acquises lors de l'entraînement pour identifier les types sémantiques de ces données et les intégrer à notre ontologie. Cette étape permet ainsi un enrichissement continu des jumeaux numériques industriels.

L'architecture de cette étape d'inférence suit un processus en trois activités, similaire à celles de la phase d'entraînement, comme illustré dans la Figure 42 :

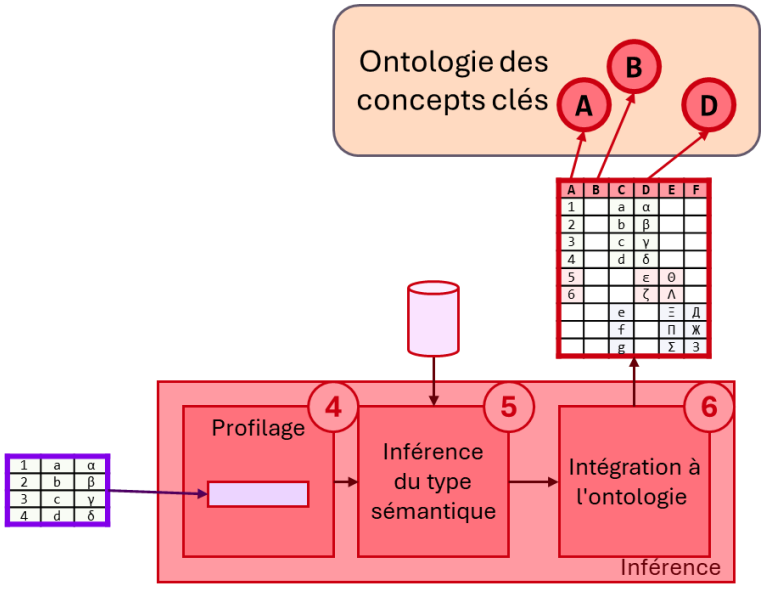


Figure 42 Architecture de l'étape d'inférence

4) Profilage des données : tout comme lors de l'entraînement, la première étape consiste à profiler les nouvelles données entrantes. Nous appliquons les mêmes techniques de profilage pour extraire un ensemble riche de caractéristiques de ces données. Cette étape nous permet de transformer les données brutes, souvent hétérogènes et non structurées caractéristiques de l'IoT industriel, en un format compatible avec notre modèle entraîné. Cette étape de profilage est essentielle pour comprendre la structure, la qualité et les particularités des données avant leur utilisation dans des tâches d'analyse avancées.
5) Inférence du type sémantique : une fois les nouvelles données profilées, nous les soumettons à notre modèle d'apprentissage automatique entraîné. Le modèle analyse les caractéristiques extraites et prédit le type sémantique le plus probable pour chaque donnée. C'est ici que le modèle met en pratique sa capacité à reconnaître les motifs et les relations appris lors de l'entraînement. Cette approche s'inspire des travaux sur l'apprentissage profond, permettant une classification efficace des types sémantiques même pour des données complexes et variées.
6) Intégration à l'ontologie : après avoir identifié les types sémantiques, nous procédons à l'intégration des données dans notre ontologie. Chaque donnée est associée au concept correspondant dans l'ontologie, en fonction de son type sémantique prédit. Cette étape permet d'enrichir notre base de connaissances avec de nouvelles informations, tout en maintenant une structure cohérente et organisée.

Il est important de noter que lors de cette étape d'inférence, notre système peut être confronté à des données de types sémantiques inconnus, c'est-à-dire des types qui n'étaient pas présents dans les données d'entraînement. Dans ce cas, le modèle attribuera le type sémantique le plus proche ou classera la donnée dans une catégorie "inconnue". Cette capacité à gérer les types sémantiques inconnus est cruciale pour la robustesse et l'adaptabilité de notre approche dans

un environnement IoT dynamique, comme l'ont souligné Xu et al. (2014) dans leur étude sur la qualité des données IoT.

L'étape d'inférence est donc le moment où notre système d'apprentissage automatique démontre sa capacité à généraliser les connaissances acquises lors de l'entraînement et à les appliquer à de nouvelles données. En suivant ce processus structuré de profilage, d'inférence et d'intégration, nous sommes en mesure d'enrichir continuellement notre ontologie avec de nouvelles informations issues de l'IoT, tout en maintenant une structure sémantique cohérente et exploitable. Cette approche s'inscrit dans la lignée des travaux de Zhu et al. (2018) sur l'apprentissage continu dans les systèmes industriels, permettant une adaptation constante à l'évolution des données IoT.

4.3.2.1 Profilage des données

Dans l'étape d'inférence, le profilage des données demeure une activité cruciale, tout comme lors de l'étape d'entraînement. Pour garantir la cohérence et la performance de notre système d'IA, il est essentiel de maintenir une similarité entre les profils de données utilisés pendant l'entraînement et ceux utilisés lors de l'inférence. Cette approche s'aligne avec les principes de cohérence des données d'entrée soulignés par (Goodfellow et al., 2016) dans leur ouvrage sur l'apprentissage profond.

Notre modèle de réseau de neurones a été spécifiquement entraîné sur les profils de données générés par l'approche de Sherlock. Bien que les réseaux de neurones soient capables de gérer des structures de données complexes et variées, ils sont optimisés pour traiter les types de données sur lesquels ils ont été entraînés. Comme l'ont démontré (Bengio et al., 2013) dans leurs travaux sur la représentation des données pour l'apprentissage profond, la cohérence des représentations entre l'entraînement et l'inférence est cruciale pour la performance des modèles. Ainsi, pour assurer une reconnaissance efficace des types sémantiques, il est primordial de fournir au modèle des profils de données conformes à ceux utilisés lors de l'entraînement.

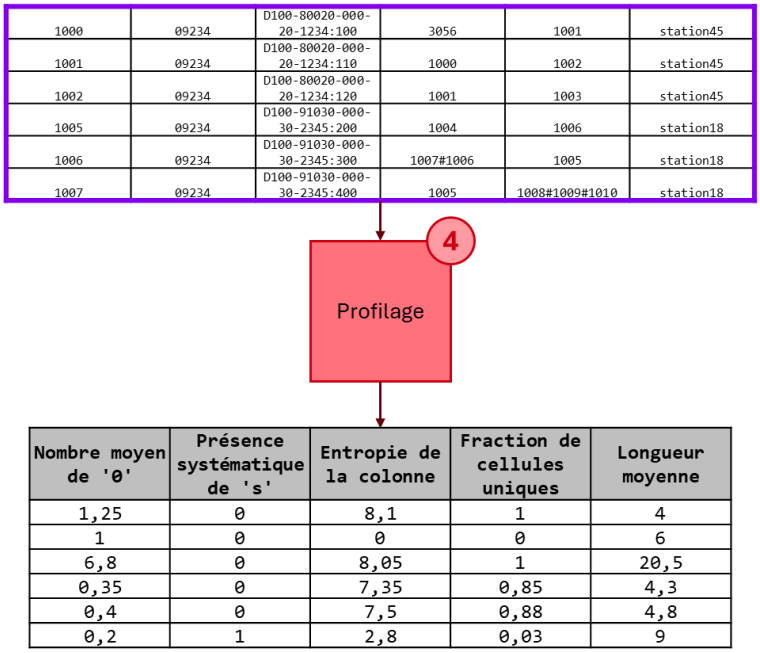


Figure 43 Processus de profilage des données dans l'étape d'inférence

La Figure 43 illustre le processus de profilage des données lors de l'étape d'inférence. Cette visualisation met en évidence la transformation des données brutes en profils exploitables par notre modèle d'apprentissage automatique.

En appliquant le même processus de profilage basé sur Sherlock aux nouvelles données entrantes, nous assurons la conformité entre les phases d'entraînement et d'inférence. Cette transformation des données brutes en profils structurés permet une représentation uniforme et informative, facilitant le traitement par notre modèle d'apprentissage automatique. Cette approche permet au modèle de reconnaître les motifs appris, même dans des données inédites, grâce à la cohérence des profils entre les phases d'entraînement et de reconnaissance.

Le profilage des données lors de la reconnaissance est donc crucial pour la performance et la robustesse de notre système d'IA. Il assure une continuité entre l'entraînement et l'inférence, optimisant la reconnaissance des types sémantiques et la généralisation à de nouvelles données. Cette approche s'aligne avec les principes de cohérence des représentations de données soulignés et est essentielle dans un environnement IoT dynamique.

De plus, ce processus permet d'identifier d'éventuelles dérives conceptuelles ou l'émergence de nouveaux types de données, une capacité cruciale pour les systèmes d'IA dans des environnements dynamiques (Gama et al., 2014). En conclusion, le profilage des données dans la phase de reconnaissance est fondamental pour l'efficacité de notre approche d'enrichissement sémantique des jumeaux numériques, contribuant à la précision de la reconnaissance et à l'adaptabilité du système face aux changements dans l'environnement IoT industriel.

4.3.2.2 Inférence du type sémantique

L'étape d'inférence du type sémantique constitue une phase cruciale de notre méthodologie, où le modèle de réseau de neurones préalablement entraîné est utilisé pour déterminer le type sémantique des données en entrée. Cette étape est fondamentale car elle permet d'analyser le profil des données et de prédire le type sémantique associé, contribuant ainsi à l'enrichissement sémantique des jumeaux numériques industriels.

Notre approche va au-delà de la simple identification du type sémantique présentant le pourcentage le plus élevé à la sortie du réseau de neurones. Nous utilisons la fonction 'Softmax', une technique largement utilisée dans les problèmes de classification multi-classes. Cette fonction permet de convertir les valeurs de sortie du réseau de neurones en probabilités, offrant ainsi une interprétation des résultats sous forme de scores de confiance pour chaque type sémantique.

Pour améliorer la précision de la détection du type sémantique, nous avons introduit un filtre basé sur un seuil de certitude défini par un expert métier. Ce seuil, exprimé en pourcentage, détermine le niveau de confiance minimal que le système doit avoir dans la prédiction d'un type sémantique avant de l'accepter comme valide. Cette approche s'inspire des travaux de (Hendrycks & Gimpel, 2018) sur la détection des erreurs de classification dans les réseaux de neurones.

Nombre moyen de '0'	Présence systématique de 's'	Entropie de la colonne	Fraction de cellules uniques	Longueur moyenne
1,25	0	8,1	1	4
1	0	0	0	6
6,8	0	8,05	1	20,5
0,35	0	7,35	0,85	4,3
0,4	0	7,5	0,88	4,8
0,2	1	2,8	0,03	9

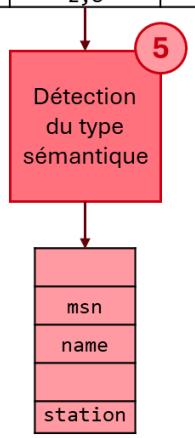


Figure 44 Exemple d'inférence du type sémantique

La Figure 44 illustre le processus d'inférence du type sémantique. Les profils de données (représentés par les caractéristiques telles que " Nombre moyen de '0'", " Présence systématique de 's'", etc.) sont analysés par le modèle, qui produit une prédiction du type sémantique (dans cet exemple, "station").

L'utilisation d'un seuil de certitude ajoute une couche supplémentaire de rigueur à notre processus de détection du type sémantique. Par exemple, si nous fixons le seuil à 80%, seuls les types sémantiques dont le score de confiance est supérieur ou égal à 80% seront acceptés par le système. Cette approche s'aligne avec les principes de gestion des risques dans les systèmes critiques, comme ceux utilisés dans l'industrie aérospatiale (Leveson, 2012).

Dans le cas où une prédiction ne dépasse pas le seuil de certitude, le profil est classé comme "ignoré". Cette stratégie de prudence privilégie la prévention des erreurs et l'exactitude des résultats à la prise de risques. Comme l'ont souligné (N. F. Noy & McGuinness, 2001), il est préférable de ne pas créer de lien dans l'ontologie plutôt que de créer un lien inapproprié qui pourrait affecter l'intégrité ontologique et perturber l'efficacité du jumeau numérique.

Pour les cas où le score de confiance est notable mais inférieur au seuil, nous envisageons d'intégrer une étape d'intervention manuelle dans les développements futurs. Cette approche permettrait d'exploiter l'expertise humaine pour résoudre les cas incertains, augmentant ainsi la précision de l'ontologie tout en maintenant une approche de précaution. Cette stratégie s'inspire des travaux de (Holzinger, 2016) sur l'apprentissage machine interactif, où l'interaction homme-machine est utilisée pour améliorer les performances des systèmes d'IA.

Cependant, l'introduction d'une étape manuelle soulève des questions d'efficacité et de subjectivité. En effet, il est crucial de trouver un équilibre entre les bénéfices de l'intervention humaine et les contraintes qu'elle impose. Une solution potentielle serait de limiter l'intervention

manuelle aux cas les plus critiques ou ambigus, maintenant ainsi un compromis entre précision et efficacité.

En conclusion, notre approche d'inférence du type sémantique combine la puissance des réseaux de neurones avec des mécanismes de contrôle rigoureux et la possibilité d'une intervention humaine. Cette méthodologie vise à garantir une haute précision dans l'enrichissement sémantique, tout en offrant la flexibilité nécessaire pour traiter les cas complexes et ambigus.

4.3.2.3 Intégration à l'ontologie

L'intégration des données IoT à une ontologie structurée constitue une étape cruciale dans notre approche d'enrichissement sémantique des jumeaux numériques industriels. Nous proposons une méthodologie que nous avons nommée "ontologification", visant à transformer les données brutes issues de l'Internet des Objets (IoT) en une ontologie cohérente et porteuse de sens. Cette approche s'inscrit dans la lignée des travaux de (Barnaghi et al., 2012) sur la sémantisation des données IoT et de (Guarino et al., 2009) sur l'ingénierie ontologique.

Pour appréhender cette transformation, nous adoptons une perspective particulière sur les bases de données IoT sur la modélisation des données IoT.

Dans notre approche :

1) Chaque table est considérée comme représentant un concept distinct.
2) Chaque ligne de la table est une instance spécifique de ce concept.
3) Chaque colonne est une propriété du concept.
4) Une propriété est elle-même un concept, relié au concept principal par un lien spécifique.

Cette vision structurée des bases de données IoT sert de fondement à notre approche d'ontologification, permettant d'aborder de manière méthodique la détection des concepts, la création des instances et l'identification des liens entre les concepts, qui sont des éléments clés dans la construction de l'ontologie.

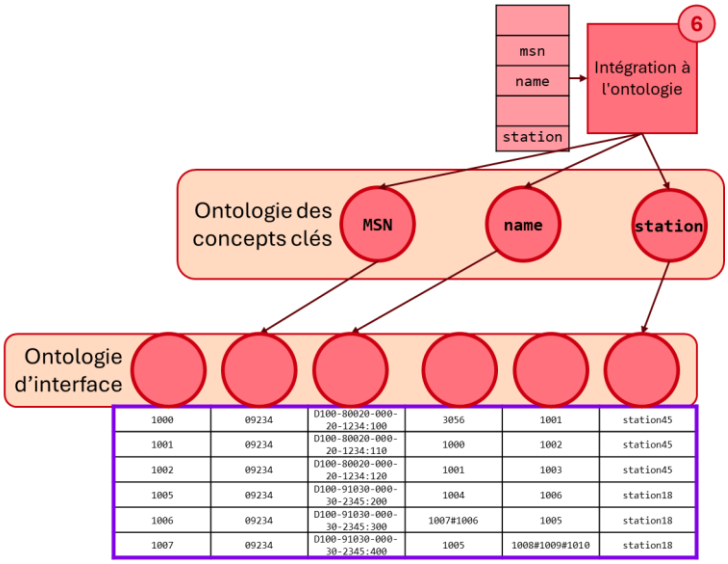


Figure 45 Schéma de l'ontologification

La Figure 45 illustre le processus d'ontologification, montrant comment les données tabulaires sont transformées en concepts ontologiques après avoir été identifiées.

Pour mieux comprendre cette transformation, considérons la Figure 46 qui illustre comment une base de données est convertie en une structure ontologique.

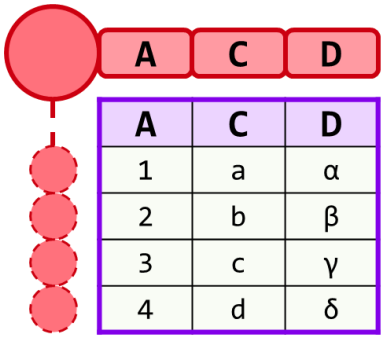


Figure 46 Schéma d'une base de données transformée en ontologie

Nous voyons comment une table de base de données (représentée par la matrice) est transformée en une structure ontologique. Le concept principal est lié à chaque instance (représentée par les cercles en pointillés) via les propriétés correspondant aux colonnes de la table. Cette représentation visuelle aide à comprendre comment chaque élément de la base de données trouve sa place dans la structure ontologique résultante.

Les concepts clés, préalablement identifiés et présents dans notre ontologie d'interface, sont associés à des colonnes spécifiques de la base de données lors de la phase d'utilisation de notre réseau d'intelligence artificielle. Ces colonnes sont alors considérées comme des propriétés dans l'ontologie des données.

Une fois cette identification réalisée par le réseau, nous établissons un lien direct entre les colonnes de données et les concepts clés associés. C'est ici que nous exploitons pleinement le potentiel des ontologies : un lien d'identité, généralement représenté par la relation "owl:sameAs" dans le langage Web Ontology Language (OWL), est créé entre l'ontologie des données (représentant les colonnes) et l'ontologie d'interface (représentant les concepts clés). Cette approche s'inspire des travaux de (Bizer et al., 2009) sur les liens d'identité dans les données liées.

Ce lien d'identité établit une correspondance forte entre les deux ontologies, affirmant que le concept clé présent dans l'ontologie d'interface est identique à la propriété correspondante dans l'ontologie des données. Cette correspondance facilite grandement l'intégration des données brutes dans l'ontologie et renforce la cohérence et la pertinence de l'ontologie résultante.

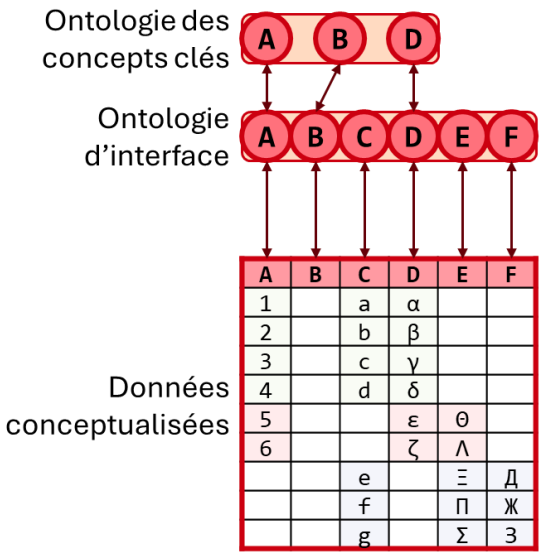


Figure 47 Schéma de la base de données réunifiée avec l'ontologie d'interface liée et l'ontologie des concepts clés

La Figure 47 illustre le processus d'intégration des données à l'ontologie, montrant comment les concepts de l'ontologie d'interface sont liés aux colonnes de la base de données, et comment ces concepts sont eux-mêmes reliés à une ontologie de plus haut niveau. Cette approche multi-niveaux permet une représentation riche et flexible des connaissances, s'alignant avec les principes de l'ingénierie ontologique décrits par (Staab & Studer, 2009).

L'ontologification des données constitue ainsi une étape cruciale dans notre approche. Elle permet de passer des données brutes de l'IoT à une représentation ontologique structurée et porteuse de sens. Cette transformation facilite non seulement l'intégration des données dans le jumeau numérique, mais aussi leur exploitation ultérieure pour la génération de nouvelles connaissances.

4.3.2.4 Synthèse

L'étape d'inférence, constitue une phase cruciale dans notre méthodologie d'intégration des données IoT. C'est lors de cette étape que notre système d'apprentissage automatique, préalablement entraîné, est confronté à de nouvelles données et doit mettre en pratique les connaissances acquises pour identifier les types sémantiques de ces données et les intégrer à notre ontologie.

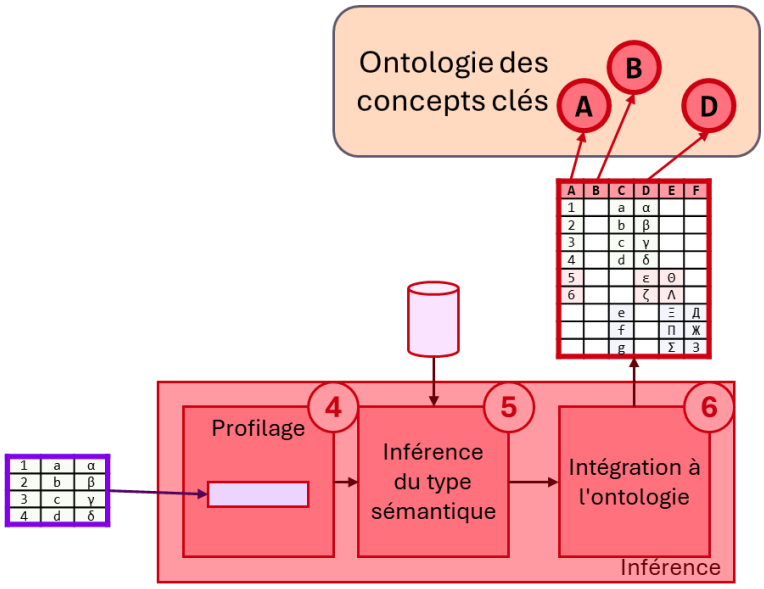


Figure 48 Schéma de l'étape de reconnaissance complète

L'étape d'inférence, illustrée dans la Figure 48, suit un processus en trois activités principales : le profilage des données, l'inférence du type sémantique, et l'intégration à l'ontologie. Cette approche met l'emphase sur l'importance du profilage des données pour comprendre leur structure et leur qualité avant leur utilisation dans des tâches d'analyse avancées.

Notre approche d'ontologification transforme les données brutes de l'IoT en une ontologie structurée et porteuse de sens. Ainsi, nous considérons les bases de données IoT comme des ensembles de concepts, d'instances et de propriétés. Les concepts clés, préalablement identifiés dans l'ontologie d'interface, sont associés aux colonnes de données lors de la phase de reconnaissance. Un lien d'identité, généralement représenté par la relation "owl:sameAs" dans le langage Web Ontology Language (OWL), est établi entre l'ontologie des données et l'ontologie d'interface, facilitant l'intégration des données brutes dans l'ontologie.

Pour mieux comprendre comment ce processus s'applique concrètement, la Figure 49 fournit un exemple détaillé.

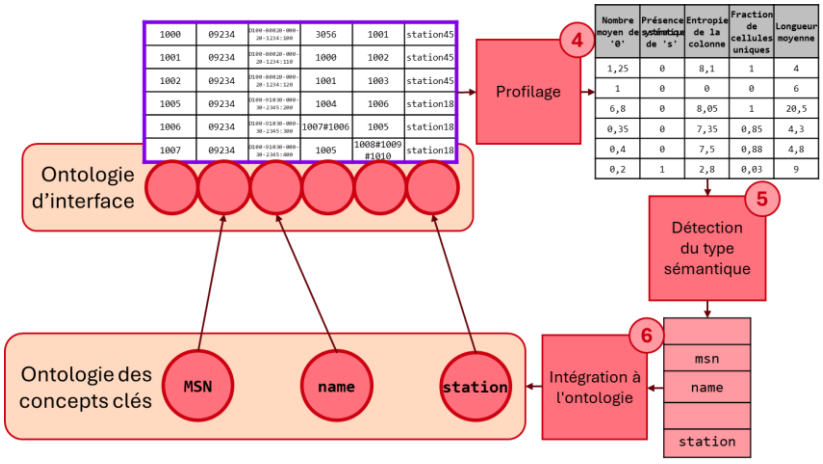


Figure 49 Exemple d'une reconnaissance du début à la fin

Cette figure illustre de manière concrète :

- Les données brutes en entrée (en haut à gauche).
- Le processus de profilage (étape 4) qui transforme ces données en caractéristiques exploitables.
- La détection du type sémantique (étape 5) basée sur ces profils.
- L'intégration finale à l'ontologie d'interface (étape 6).

L'étape de reconnaissance démontre la capacité de notre système d'apprentissage automatique à généraliser les connaissances acquises lors de l'entraînement et à les appliquer à de nouvelles données. En suivant ce processus structuré de profilage, de reconnaissance et d'intégration, nous enrichissons continuellement notre ontologie avec de nouvelles informations issues de l'IoT, tout en maintenant une structure sémantique cohérente et exploitable.

L'ontologification des données ouvre la voie à une exploitation plus efficace et pertinente des données IoT, en les rendant accessibles et compréhensibles pour les différents acteurs impliqués dans le développement et l'utilisation des jumeaux numériques. Cette approche contribue ainsi à l'amélioration de l'interopérabilité et de la performance des jumeaux numériques dans un environnement industriel complexe.

4.3.3 Synthèse globale de la phase d'ontologification

Au cours de ce chapitre, nous avons présenté notre méthodologie d'intégration des données IoT, qui vise à transformer les données brutes en une ontologie structurée et porteuse de sens. Cette approche, que nous avons nommée "ontologification", s'inscrit dans la lignée des travaux de (Barnaghi et al., 2012) sur la sémantisation des données IoT et de (Guarino et al., 2009) sur l'ingénierie ontologique. Notre méthodologie s'articule autour de deux étapes principales : l'étape d'entraînement et l'étape d'inférence.

L'étape d'entraînement comprend trois activités clés :

1) Profilage des données : transformation des données brutes en format exploitable.
2) Identification des types sémantiques clés : sélection manuelle des concepts essentiels du domaine.
3) Entraînement des modèles : utilisation de l'architecture Sherlock avec des modèles de petite taille.

L'étape d'inférence met en pratique le modèle entraîné pour traiter de nouvelles données :

4) Profilage des données : application cohérente des techniques de profilage.
5) Inférence du type sémantique : prédiction du type sémantique le plus probable.
6) Intégration à l'ontologie : association des données aux concepts correspondants

Pour valider initialement notre approche, nous avons effectué des tests préliminaires sur un jeu de données restreint. Cette étape de validation précoce, recommandée par (Alpaydin, 2020) dans ses travaux sur l'apprentissage automatique, nous a permis d'évaluer la faisabilité et le potentiel de notre méthodologie avant de passer à une mise en œuvre à plus grande échelle.

Dans ces tests préliminaires, nous avons utilisé Sherlock comme profileur de données, conformément à notre approche inspirée des travaux de (Hulsebos et al., 2019). Nous avons employé deux ensembles de données IoT distincts, représentatifs des types de données que nous

prévoyons de traiter dans un contexte industriel réel. Comme l'ont suggéré (Goodfellow et al., 2016) dans leur ouvrage sur l'apprentissage profond, nous avons divisé notre ensemble de données en plusieurs sous-ensembles plus petits pour augmenter la quantité d'échantillons d'entraînement, une technique connue sous le nom d'augmentation de données.

Les résultats initiaux de ces tests préliminaires se sont révélés extrêmement prometteurs :

1) Précision : notre modèle a atteint une précision remarquable de 98% dans la classification des types sémantiques. Cette performance exceptionnelle témoigne de la robustesse et de l'efficacité de notre approche.
2) Temps de traitement : la phase de conversion des données en profils n'a nécessité qu'environ 30 secondes. Ce temps de traitement remarquablement court est un atout majeur pour l'applicabilité de notre méthode dans des environnements IoT en temps réel, où la rapidité de traitement est souvent cruciale.
3) Efficacité de l'entraînement : notre modèle a démontré une efficacité d'apprentissage impressionnante, convergeant en moins de 10 epochs. Cette rapidité de convergence est particulièrement avantageuse pour l'adaptabilité du modèle à de nouvelles données, un aspect essentiel dans les environnements IoT dynamiques et en constante évolution.

Il est important de noter que ces résultats, bien qu'encourageants, proviennent d'un jeu de test préliminaire et limité. Comme l'ont souligné (T. M. Mitchell, 1997) et (Domingos, 2012) dans leurs travaux sur l'évaluation des modèles d'apprentissage automatique, il est crucial de valider ces résultats sur des ensembles de données plus larges et plus diversifiés avant de tirer des conclusions définitives sur la performance de notre approche.

Ces tests préliminaires nous ont fourni des indications précieuses sur le potentiel de notre méthodologie et ont guidé nos décisions pour la suite de nos travaux. Ils ont notamment confirmé la pertinence de notre choix d'utiliser Sherlock comme base pour le profilage des données et ont démontré la capacité de notre approche à traiter efficacement les données IoT. Ces résultats initiaux nous encouragent à poursuivre le développement et l'optimisation de notre méthodologie d'ontologification, tout en restant conscients de la nécessité de tests plus approfondis et à plus grande échelle pour une validation complète.

Pour la mise en œuvre et l'évaluation de notre méthodologie, nous prévoyons d'adopter les principes suivants :

1) Division des données : nous séparerons nos ensembles de données en deux parties distinctes, une pour l'entraînement et une pour la validation, assurant ainsi une évaluation non biaisée de la performance du modèle.
2) Indicateurs de performance : nous caractériserons la performance de notre approche à travers plusieurs indicateurs clés, notamment :
    a) La précision de la classification des types sémantiques
    b) Le temps de traitement pour le profilage des données
    c) Le temps nécessaire pour l'entraînement du modèle
    d) La capacité du modèle à généraliser sur de nouvelles données
3) Évaluation comparative : bien que les comparaisons directes avec d'autres méthodes puissent être complexes, nous privilégierons les mesures quantitatives et envisagerons

l'annotation manuelle d'un sous-ensemble de données pour permettre une évaluation pertinente.

4) Gestion des cas d'échec : nous analyserons en détail les cas où notre modèle échoue à classifier correctement les données, en portant une attention particulière aux situations où les données sont absentes ou ambiguës.

5) Passage à l'échelle : pour la mise en œuvre finale dans un contexte industriel réel, nous prévoyons d'adapter notre approche pour gérer des volumes de données plus importants, notamment en envisageant une migration vers le cloud et en optimisant nos algorithmes pour le traitement de données massives.

En conclusion, notre méthodologie d'ontologification offre une approche structurée et efficace pour transformer les données brutes de l'IoT en une ontologie porteuse de sens. En combinant des techniques de profilage de données, d'identification des types sémantiques clés, d'entraînement de modèles et d'ontologification, notre approche permet de surmonter les défis liés à l'hétérogénéité et à la variété sémantique des données IoT. Les résultats prometteurs obtenus lors de nos expérimentations ouvrent la voie à une exploitation plus efficace des données IoT dans le développement et l'utilisation des jumeaux numériques, offrant ainsi de nouvelles perspectives pour l'optimisation des processus industriels et la prise de décision basée sur des données fiables et structurées.

4.4 Phase d'alignement

La phase d'alignement est une étape cruciale dans notre démarche d'intégration des données de l'Internet des Objets (IoT) au sein d'un jumeau numérique industriel. Cette phase vise à résoudre la problématique complexe de l'alignement concret des ontologies entre elles, afin de permettre une interopérabilité sémantique efficace entre les différents modèles et sources de données. Comme l'ont souligné Euzenat et Shvaiko (2013) dans leurs travaux sur l'alignement d'ontologies, cette étape est essentielle pour assurer une intégration cohérente et significative des données hétérogènes.

Dans le contexte d'une entreprise aéronautique comme Airbus, cette phase d'alignement doit prendre en compte les spécificités des modèles informatiques propres à chaque domaine métier. Il s'agit de proposer une approche de modélisation adaptée, permettant d'intégrer de nouveaux concepts au sein des ontologies existantes de manière cohérente et structurée. Cette approche s'inscrit dans la lignée des travaux de Uschold et Gruninger (2004) sur l'ingénierie ontologique dans les environnements industriels complexes.

Un des défis majeurs consiste à agréger des modèles informatiques appartenant au même domaine. Lorsqu'un nouveau concept est identifié, il est essentiel de déterminer où le positionner dans l'ontologie existante. Cela nécessite une réflexion approfondie sur la structure et les relations entre les concepts, afin de maintenir la cohérence et la pertinence de l'ontologie. Cette problématique rejoint les travaux de Noy et McGuinness (2001) sur le développement et la maintenance d'ontologies évolutives.

4.4.1 Construction d'une ontologie de référence

Pour faciliter l'alignement et l'intégration des différentes ontologies, nous avons entrepris la construction d'une ontologie de référence, également appelée "ontologie pivot". Cette ontologie

servira de structure centrale pour supporter le réseau sémantique intégrant l'ensemble des données du jumeau numérique. L'approche de l'ontologie pivot s'inspire des travaux de Stumme et Maedche (2001) sur l'intégration sémantique des données hétérogènes.

Notre approche pour la construction de cette ontologie pivot s'est appuyée sur les bonnes pratiques identifiées dans l'état de l'art, notamment les principes énoncés par Gruber (1995) pour la conception d'ontologies partageables. Nous avons adapté ces principes au contexte spécifique du jumeau numérique d'une entreprise aéronautique, en impliquant des experts de différents domaines dans le processus de création de l'ontologie. Cette démarche collaborative s'inspire des travaux de Pinto et al. (2009) sur la construction d'ontologies par consensus d'experts.

Des ateliers collaboratifs ont été organisés, réunissant des experts de domaines similaires mais distincts. Lors de ces ateliers, les experts ont partagé leur vision des concepts fondamentaux de leur domaine, ainsi que les liens hiérarchiques et sémantiques entre ces concepts. Ils se sont appuyés sur leur expérience pratique et leur connaissance approfondie des modèles métiers existants pour identifier les éléments essentiels à intégrer dans l'ontologie pivot. Cette approche s'aligne avec les recommandations de Fernández-López et al. (1997) sur l'implication des experts de domaine dans le développement d'ontologies.

Un processus itératif a été mis en place, permettant de raffiner progressivement l'ontologie pivot en confrontant leurs perspectives et en recherchant un consensus. Des outils de visualisation et de modélisation collaborative ont été utilisés pour faciliter ce processus et permettre une représentation claire et partagée de l'ontologie en construction. L'utilisation de tels outils est recommandée par Tudorache et al. (2013) pour améliorer la collaboration dans le développement d'ontologies complexes.

L'objectif était de tirer parti de l'intelligence collective des experts pour construire une ontologie pivot qui capture les concepts et les relations clés, tout en étant suffisamment générique pour servir de point de référence pour l'alignement des ontologies spécifiques à chaque domaine. Cette approche s'inspire des travaux de Borst (1997) sur la construction d'ontologies modulaires et réutilisables.

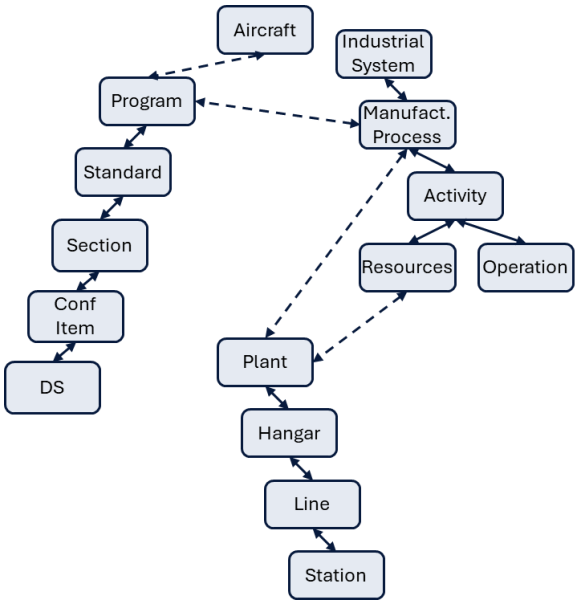


Figure 50 Structure initiale de l'ontologie pivot pour l'intégration sémantique dans l'industrie aéronautique

La Figure 50 présente la structure initiale de notre ontologie pivot, développée pour servir de base à notre approche d'intégration sémantique dans le contexte aéronautique.

Cette structure initiale fournit un cadre solide pour l'alignement des différentes ontologies spécifiques à chaque domaine, facilitant ainsi l'intégration sémantique des données à travers l'ensemble du système de production aéronautique. Comme l'ont souligné Uschold et Gruninger (2004), une telle ontologie de haut niveau est essentielle pour assurer la cohérence et l'interopérabilité dans les environnements industriels complexes.

4.4.2 Outils d'alignement d'ontologies

L'alignement automatique d'ontologies est un domaine de recherche actif et crucial pour notre approche d'intégration sémantique des données IoT dans les jumeaux numériques industriels. Comme l'ont souligné Otero-Cerdeira et al. (2015) dans leur analyse exhaustive de l'état de l'art, de nombreux outils d'alignement ont été développés au fil des années, mais beaucoup restent au stade théorique ou manquent de maintenance à long terme. Ganzha et al. (2018) ont également noté que de nombreux outils sont développés uniquement pour l'Ontology Alignment Evaluation Initiative (OAEI) et sont abandonnés peu après.

Malgré ces défis, certains outils se distinguent par leur robustesse, leur efficacité et leur maintenance continue. Dans le cadre de notre recherche, nous avons identifié trois outils particulièrement pertinents pour notre contexte d'alignement entre l'ontologie de référence et l'ontologie générée par nos modèles d'IA :

4.4.2.1 COMA++

COMA++ (Aumueller et al., 2005) est une évolution significative de l'outil COMA original. Il se distingue par sa capacité à traiter à la fois des schémas et des ontologies, offrant une approche unifiée grâce à une représentation de données générique. Ses principales caractéristiques incluent :

- Une interface graphique interactive.
- De nouvelles approches pour l'alignement d'ontologies, notamment l'utilisation de taxonomies partagées.
- Des stratégies de correspondance flexibles, y compris la réutilisation de résultats précédents.
- Une approche de correspondance basée sur des fragments pour décomposer les problèmes complexes.

COMA++ se positionne non seulement comme un outil d'alignement, mais aussi comme une plateforme d'évaluation comparative pour différents algorithmes et stratégies d'alignement. Cette flexibilité pourrait s'avérer particulièrement utile dans notre contexte, où nous devons aligner des ontologies de différents domaines métiers avec notre ontologie de référence.

4.4.2.2 LogMap

LogMap (Jiménez-Ruiz & Cuenca Grau, 2011) se distingue par sa capacité à traiter des ontologies volumineuses et complexes, contenant jusqu'à des centaines de milliers de classes. Ses caractéristiques clés comprennent :

- Des capacités intégrées de raisonnement et de diagnostic.

- Une combinaison de mesures de similarité lexicale, structurelle et sémantique.
- La détection et la réparation à la volée des classes insatisfaisables.
- Une efficacité prouvée sur des ontologies biomédicales complexes.

LogMap vise à produire des alignements de haute qualité tout en maintenant la cohérence logique des ontologies intégrées. Cette capacité à gérer des ontologies complexes et volumineuses pourrait être particulièrement pertinente pour notre cas d'application dans l'industrie aéronautique, où les ontologies peuvent atteindre une taille et une complexité considérables.

4.4.2.3 AgreementMakerLight

AgreementMakerLight (Faria et al., 2013) est une version optimisée de l'outil AgreementMaker, conçue spécifiquement pour gérer de très grandes ontologies. Ses points forts incluent :

- Une efficacité computationnelle améliorée pour traiter des ontologies volumineuses.
- Un algorithme Word Matcher innovant pour l'identification des correspondances.
- Une approche lexicale sophistiquée, prenant en compte la fréquence et le contenu probant des mots.
- Une flexibilité et une extensibilité préservées malgré l'optimisation pour les grandes ontologies.

AgreementMakerLight s'est montré particulièrement efficace dans le domaine biomédical, où les ontologies sont souvent très volumineuses et complexes. Cette capacité à gérer efficacement de grandes ontologies pourrait être un atout majeur dans notre contexte industriel, où nous devons intégrer des données provenant de multiples sources et domaines.

Ces trois outils représentent l'état de l'art actuel en matière d'alignement automatique d'ontologies. Ils se distinguent par leur maintenance continue, leur efficacité prouvée sur des cas d'utilisation réels, et leur capacité à traiter des ontologies de grande taille et de complexité variable. Leur développement continu et leur utilisation dans des contextes pratiques en font des choix pertinents pour notre approche d'alignement entre l'ontologie de référence et les ontologies générées par nos modèles d'IA.

Dans le cadre de notre recherche, nous prévoyons d'évaluer ces trois outils dans notre contexte spécifique d'alignement entre l'ontologie de référence et l'ontologie générée par nos modèles d'IA. Cette évaluation nous permettra de déterminer lequel de ces outils, ou quelle combinaison d'outils, est le plus adapté à nos besoins spécifiques d'intégration sémantique des données IoT dans les jumeaux numériques industriels. Nous porterons une attention particulière à leur capacité à gérer les spécificités de nos ontologies, à leur performance en termes de précision et de rappel, ainsi qu'à leur facilité d'intégration dans notre pipeline de traitement des données.

4.4.3 Synthèse et perspectives

L'alignement d'ontologies joue un rôle crucial dans notre approche d'intégration sémantique des données IoT au sein des jumeaux numériques industriels. Comme nous l'avons vu, cette phase vise à résoudre la problématique complexe de l'alignement concret des ontologies entre elles, afin de permettre une interopérabilité sémantique efficace entre les différents modèles et sources de données. Notre démarche s'est articulée autour de deux axes principaux : la

construction d'une ontologie de référence et l'identification d'outils d'alignement automatique performants.

La construction de l'ontologie de référence, ou ontologie pivot, s'est appuyée sur une approche collaborative impliquant des experts de différents domaines. Cette démarche, inspirée des travaux de Pinto et al. (2009) sur la construction d'ontologies par consensus d'experts, nous a permis de développer une structure initiale solide, capable de servir de point d'ancrage pour l'alignement des ontologies spécifiques à chaque domaine. Cette approche s'aligne avec les recommandations de Uschold et Gruninger (2004) sur l'importance d'une ontologie de haut niveau pour assurer la cohérence et l'interopérabilité dans les environnements industriels complexes.

Concernant les outils d'alignement automatique, notre analyse de l'état de l'art nous a permis d'identifier trois outils particulièrement prometteurs : COMA++, LogMap et AgreementMakerLight. Ces outils se distinguent par leur robustesse, leur efficacité et leur maintenance continue, ainsi que par leur capacité à traiter des ontologies de grande taille et de complexité variable. Chacun de ces outils présente des caractéristiques spécifiques qui pourraient s'avérer précieuses dans notre contexte d'intégration de données IoT dans un environnement industriel complexe.

En résumé, bien que notre étude ait permis d'identifier des outils prometteurs pour l'alignement d'ontologies et de poser les bases d'une approche méthodologique solide, nous ne sommes pas encore en mesure de tirer des conclusions définitives quant à leur efficacité dans notre contexte spécifique. Une évaluation rigoureuse nécessite des ontologies de taille et de complexité représentatives de l'environnement industriel réel, comme l'ont souligné Euzenat et Shvaiko (2013).

Par conséquent, la prochaine étape cruciale de notre recherche consistera à mettre en pratique notre approche à grande échelle lors de la phase de contribution. Cette démarche, s'inscrivant dans la lignée des recommandations de Fernández-López et al. (1997) sur le développement itératif d'ontologies dans des contextes industriels complexes, nous permettra non seulement d'évaluer rigoureusement les outils d'alignement, mais aussi d'affiner notre ontologie de référence et notre méthodologie globale d'intégration sémantique des données IoT. Ce n'est qu'à l'issue de cette phase que nous pourrons véritablement valider notre approche et sélectionner les outils les plus adaptés à nos besoins spécifiques d'intégration sémantique des données IoT au sein des jumeaux numériques industriels.

4.5 Phase d'exploration

Une fois l'ontologie construite et alignée, il est essentiel de fournir aux utilisateurs des moyens efficaces pour explorer et interagir avec les connaissances qu'elle contient. Cette phase d'exploration est cruciale pour permettre aux experts métiers de tirer pleinement parti de l'ontologie et de l'exploiter dans le contexte du jumeau numérique. Comme l'ont souligné Gruber et al. (2021), l'exploration efficace des ontologies est un facteur clé pour leur adoption et leur utilisation dans des contextes industriels complexes.

4.5.1 Édition directe du fichier OWL

Une première option pour l'exploration de l'ontologie consiste à éditer directement le fichier OWL à l'aide d'un éditeur de texte simple, comme le bloc-notes. Le langage OWL est conçu pour être lisible par un humain, avec une syntaxe claire et structurée. Cela permet aux utilisateurs ayant une connaissance approfondie du langage OWL de naviguer dans l'ontologie et de comprendre sa structure et ses relations. Comme l'ont noté Horrocks et al. (2003), cette approche peut être utile pour les développeurs d'ontologies expérimentés.

Cependant, cette approche présente plusieurs inconvénients. Tout d'abord, l'édition manuelle du fichier OWL peut être longue et fastidieuse, surtout pour des ontologies de grande taille. L'utilisateur doit parcourir le fichier ligne par ligne pour localiser les informations pertinentes, ce qui peut être chronophage et peu pratique.

De plus, l'édition directe du fichier OWL est propice aux erreurs. L'utilisateur doit respecter scrupuleusement la syntaxe et la structure du langage, sans quoi il risque d'introduire des erreurs qui peuvent compromettre l'intégrité de l'ontologie. Une simple faute de frappe ou une balise mal fermée peut rendre l'ontologie invalide et inutilisable. Comme l'ont souligné Bechhofer et al. (2004), ces erreurs peuvent être difficiles à détecter et à corriger sans outils spécialisés.

Enfin, cette approche nécessite une connaissance approfondie du langage OWL, ce qui peut être un frein pour les utilisateurs non experts. Les concepts et la syntaxe d'OWL peuvent être complexes et difficiles à appréhender pour les personnes n'ayant pas une formation spécifique dans ce domaine.

En raison de ces limitations, l'édition directe du fichier OWL n'est généralement pas recommandée pour l'exploration de l'ontologie, surtout dans un contexte où l'ontologie est destinée à être utilisée par des experts métiers qui ne sont pas nécessairement familiers avec les détails techniques du langage OWL.

4.5.2 Utilisation d'un éditeur d'ontologies

Une alternative à l'édition directe du fichier OWL est l'utilisation d'un éditeur d'ontologies dédié, tel que Protégé. Protégé est un outil open source développé par l'Université de Stanford, spécialement conçu pour la création, la visualisation et la manipulation d'ontologies (Musen, 2015).

L'un des principaux avantages de Protégé est qu'il offre une interface graphique ergonomique pour interagir avec l'ontologie. Les utilisateurs peuvent naviguer dans la hiérarchie des concepts, visualiser les relations entre les entités et modifier les propriétés et les axiomes de l'ontologie de manière intuitive, sans avoir à manipuler directement le fichier OWL.

Protégé propose une vue d'ensemble de l'ontologie sous forme d'arborescence, permettant aux utilisateurs de parcourir facilement les classes, les propriétés et les individus. Il offre également des fonctionnalités avancées, telles que la définition de restrictions sur les propriétés, la création de règles d'inférence et la visualisation graphique des relations entre les entités. Comme l'ont démontré Horridge et al. (2011), ces fonctionnalités facilitent grandement la compréhension et la manipulation des ontologies complexes.

Un autre atout de Protégé est sa flexibilité et son extensibilité. Il prend en charge différents formats d'ontologies, tels que OWL et RDF, et peut être étendu avec des plugins pour ajouter des

fonctionnalités supplémentaires. Cela permet d'adapter Protégé aux besoins spécifiques d'un projet et d'intégrer des outils complémentaires pour l'analyse et l'exploration des données ontologiques.

Cependant, Protégé peut être relativement lourd à installer et à prendre en main, notamment pour les utilisateurs non familiers avec les ontologies. Son interface, bien que conviviale, peut sembler complexe au premier abord, avec de nombreuses options et fonctionnalités. Une formation initiale peut être nécessaire pour que les utilisateurs puissent exploiter pleinement les capacités de l'outil.

De plus, Protégé peut être moins adapté aux ontologies de très grande taille, en raison des limitations de performance et de visualisation. Le chargement et la manipulation d'ontologies volumineuses peuvent être lents, et la visualisation de grands graphes peut devenir confuse et difficile à naviguer. Comme l'ont souligné Lohmann et al. (2016), ces limitations peuvent entraver l'exploration efficace des ontologies complexes.

Malgré ces limitations, Protégé (Figure 51) reste un outil puissant et largement utilisé pour l'édition et l'exploration d'ontologies. Pour des ontologies de taille moyenne et pour des utilisateurs experts, Protégé offre un environnement complet et flexible pour interagir avec les connaissances ontologiques. Son interface graphique et ses fonctionnalités avancées en font un choix pertinent pour la construction et la maintenance d'ontologies dans un contexte de jumeau numérique.

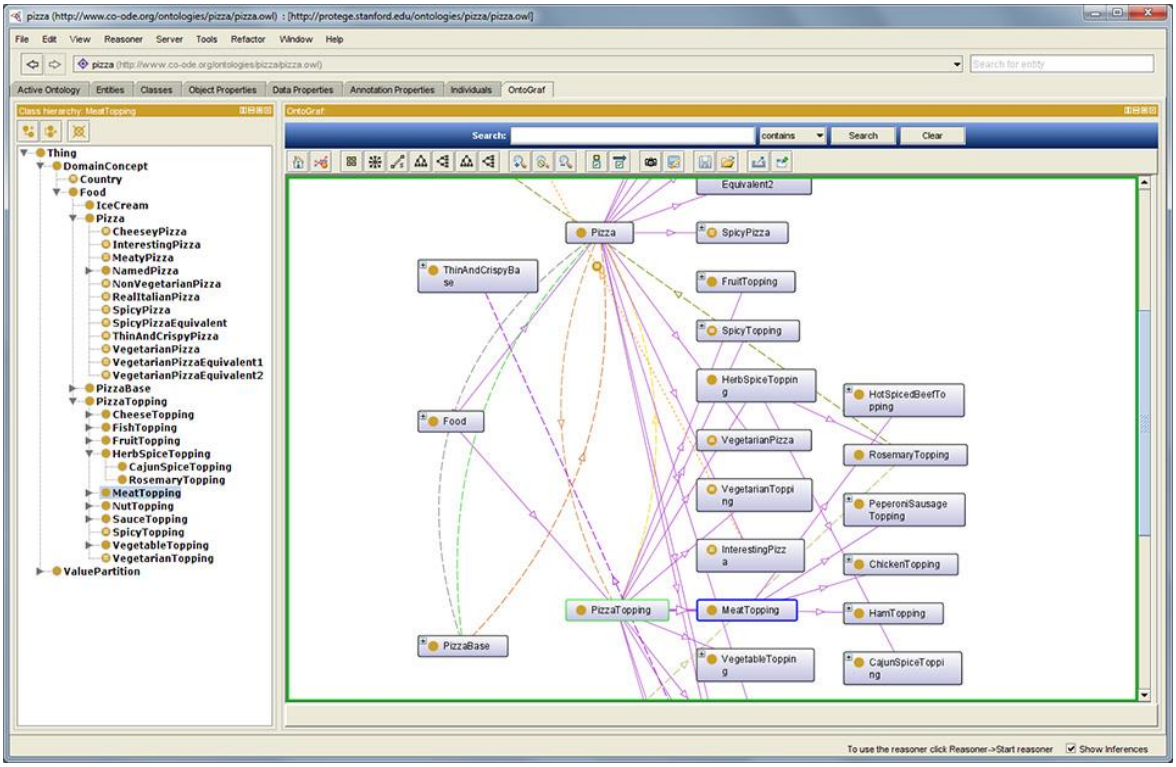


Figure 51 Interface de Protégé

4.5.3 Utilisation d'outils de visualisation légers

Pour faciliter l'exploration des ontologies de petite à moyenne taille, des outils de visualisation légers comme WebOWL peuvent être une solution intéressante. WebOWL est une application

web open source qui permet de visualiser et de naviguer dans des ontologies de manière interactive (Lohmann et al., 2015).

L'un des principaux avantages de WebOWL est sa simplicité d'utilisation. Contrairement à des outils plus complets comme Protégé, WebOWL se concentre sur la visualisation et l'exploration des ontologies, sans proposer de fonctionnalités d'édition avancées. Cela le rend plus accessible aux utilisateurs non experts, qui peuvent rapidement prendre en main l'outil et commencer à explorer l'ontologie.

WebOWL offre une interface intuitive et conviviale pour naviguer dans l'ontologie. Les utilisateurs peuvent parcourir la hiérarchie des classes, visualiser les propriétés et les individus, et suivre les relations entre les différents éléments. La visualisation graphique des relations permet de comprendre facilement la structure et les connexions au sein de l'ontologie.

Un autre atout de WebOWL est sa nature web. L'application peut être déployée sur un serveur et accédée via un navigateur web, ce qui la rend facile à partager et à utiliser. Les utilisateurs n'ont pas besoin d'installer de logiciel spécifique sur leur machine, ce qui facilite l'accès à l'ontologie et encourage la collaboration entre les différents acteurs du projet.

WebOWL propose également des fonctionnalités de recherche et de filtrage pour faciliter l'exploration des ontologies. Les utilisateurs peuvent rechercher des classes, des propriétés ou des individus spécifiques, et filtrer les résultats en fonction de critères prédéfinis. Cela permet de naviguer efficacement dans l'ontologie et de trouver rapidement les informations pertinentes.

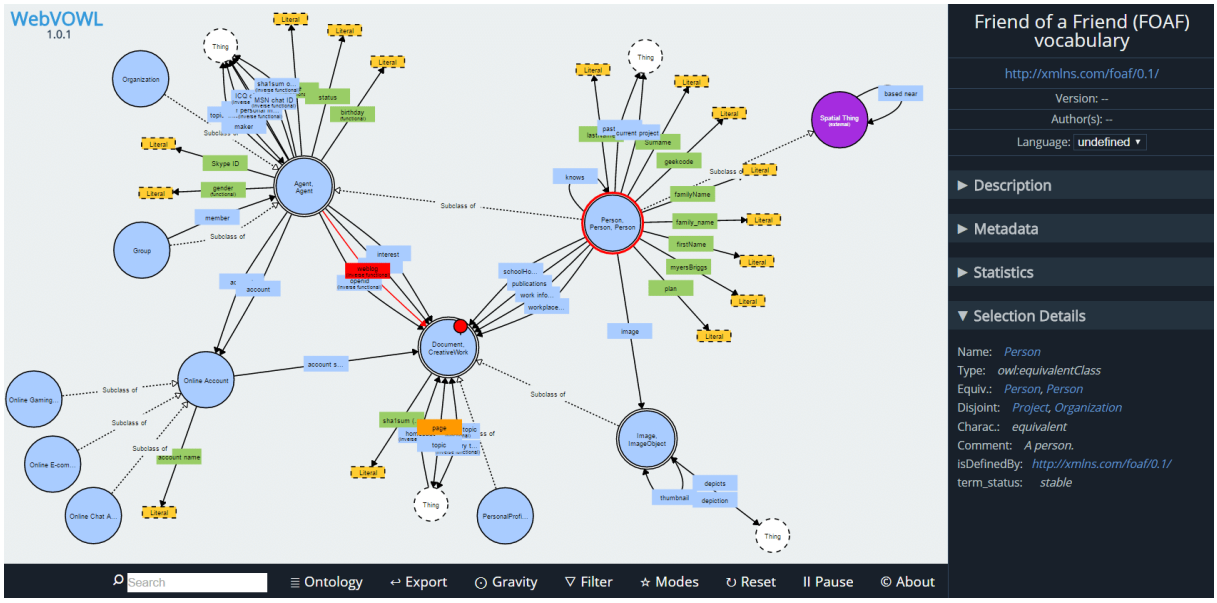


Figure 52 Interface de WebOWL – visualisation d'ontologies de grande taille

Cependant, WebOWL peut être moins adapté aux ontologies de très grande taille ou présentant une structure complexe (Figure 52). La visualisation graphique peut devenir confuse et difficile à naviguer lorsque l'ontologie contient un grand nombre de classes et de relations. De plus, les fonctionnalités de recherche et de filtrage peuvent être limitées pour des ontologies volumineuses. Comme l'ont noté Katifori et al. (2007), ces limitations sont communes à de nombreux outils de visualisation d'ontologies et peuvent entraver l'exploration efficace des ontologies complexes.

Malgré ces limitations, WebOWL reste un outil utile pour l'exploration d'ontologies de petite à moyenne taille. Sa simplicité d'utilisation, son interface intuitive et sa nature web en font un choix pertinent pour permettre aux experts métiers de naviguer dans l'ontologie et de comprendre les relations entre les concepts. Dans le contexte du jumeau numérique, WebOWL peut être utilisé comme un outil complémentaire pour faciliter l'accès aux connaissances ontologiques et encourager leur utilisation par les différents acteurs du projet.

4.5.4 Utilisation d'une base de données de graphes

Pour des ontologies de grande taille et pour une exploration approfondie des connaissances, l'utilisation d'une base de données de graphes comme Amazon Neptune peut être une solution particulièrement pertinente. Amazon Neptune est un service de base de données de graphes entièrement géré, conçu pour stocker et interroger efficacement des données hautement connectées (Raman et al., 2020).

L'un des principaux avantages d'Amazon Neptune est sa capacité à gérer des graphes de grande taille et complexes. Contrairement à des outils de visualisation légers comme WebOWL, qui peuvent être limités par la taille de l'ontologie, Amazon Neptune est conçu pour gérer des milliards de relations et offrir des performances élevées pour les requêtes et les traversées de graphes. Cela le rend particulièrement adapté à l'exploration d'ontologies volumineuses et riches en relations. Comme l'ont démontré Angles et al. (2018), les bases de données de graphes offrent des avantages significatifs pour le traitement de données hautement interconnectées.

Amazon Neptune supporte les langages de requête de graphes populaires tels que Gremlin et SPARQL, ce qui facilite l'interrogation et l'analyse des données ontologiques. Les utilisateurs peuvent exécuter des requêtes complexes pour extraire des informations spécifiques, découvrir des motifs et des relations cachées, et réaliser des analyses approfondies sur les données ontologiques. Les langages de requête de graphes offrent une grande flexibilité et expressivité pour explorer les connaissances contenues dans l'ontologie.

Un autre atout d'Amazon Neptune est son intégration avec d'autres services AWS. Il peut être facilement combiné avec des services de stockage comme Amazon S3 pour le stockage des données ontologiques, et avec des services d'analyse comme Amazon Athena pour l'analyse interactive des données. Cette intégration permet de construire des pipelines de données complets, allant du stockage à l'analyse en passant par l'exploration des connaissances ontologiques.

L'interface utilisateur d'Amazon Neptune (Figure 53) est également un point fort. Elle offre une représentation visuelle intuitive du graphe ontologique, permettant aux utilisateurs de naviguer facilement dans les relations entre les concepts. Les utilisateurs peuvent interagir avec le graphe, zoomer sur des parties spécifiques, filtrer les données et exécuter des requêtes directement depuis l'interface. Cette approche visuelle rend l'exploration des connaissances plus accessible et compréhensible, même pour les utilisateurs non experts.

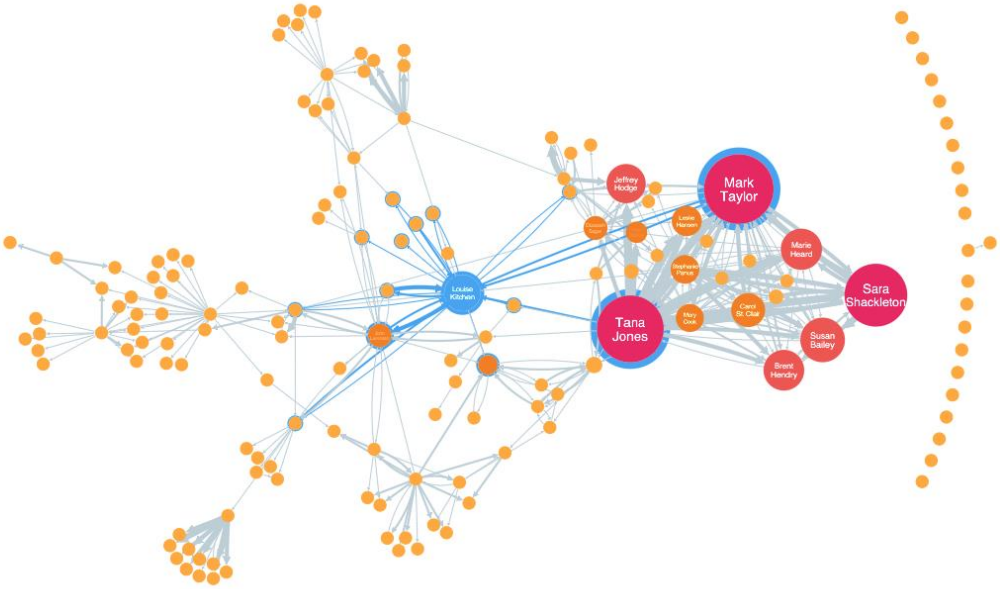


Figure 53 Interface de Amazon Neptune

Cependant, la mise en place d'une solution comme Amazon Neptune peut nécessiter un investissement initial plus important, tant en termes de ressources que de compétences techniques. Il est nécessaire de configurer correctement l'environnement, de charger les données ontologiques dans la base de données de graphes et de définir les requêtes et les visualisations appropriées. Cela peut demander un effort supplémentaire par rapport à des outils plus légers comme WebOWL.

Malgré ces défis, Amazon Neptune reste une option particulièrement intéressante pour l'exploration d'ontologies dans le contexte du jumeau numérique. Sa capacité à gérer des graphes de grande taille, ses performances élevées et son interface utilisateur conviviale en font un choix pertinent pour permettre aux experts métiers d'exploiter pleinement les connaissances contenues dans l'ontologie. Grâce à Amazon Neptune, les utilisateurs peuvent naviguer intuitivement dans le graphe ontologique, exécuter des requêtes complexes et réaliser des analyses approfondies pour extraire des informations précieuses et supporter la prise de décision.

4.5.5 Synthèse et choix de l'approche d'exploration

Le choix d'une approche d'exploration d'ontologie dépend de plusieurs facteurs, notamment les besoins spécifiques du projet, la taille de l'ontologie et le niveau d'expertise des utilisateurs. Pour les ontologies de petite à moyenne taille et les utilisateurs non experts, des outils de visualisation légers comme WebOWL peuvent suffire. Cependant, pour les ontologies volumineuses et complexes nécessitant une exploration approfondie, une base de données de graphes telle qu'Amazon Neptune offre des avantages significatifs en termes de performance, d'évolutivité et de convivialité.

Dans le contexte de l'exploitation d'un jumeau numérique représentant l'ensemble du cycle de vie d'un système, nous recommandons l'utilisation d'Amazon Neptune. Cette solution permet

une gestion efficace des ontologies de grande taille, offre une interface intuitive aux utilisateurs, et facilite la réalisation d'analyses approfondies sur les connaissances ontologiques. Grâce à Amazon Neptune, les experts métiers peuvent naviguer aisément dans le graphe ontologique, exécuter des requêtes complexes, et extraire des informations précieuses pour soutenir la prise de décision et l'optimisation des processus dans le cadre du jumeau numérique.

Ce choix est motivé par la capacité d'Amazon Neptune à passer à l'échelle, ce qui est crucial pour gérer efficacement les ontologies complexes et volumineuses typiques des jumeaux numériques industriels. Cette approche permet d'exploiter pleinement le potentiel des connaissances ontologiques pour améliorer la compréhension et l'optimisation des systèmes industriels tout au long de leur cycle de vie.

4.6 Conclusion générale

4.6.1 Résumé des contributions

Notre étude a abordé la problématique complexe de l'intégration des données IoT dans le contexte du développement d'un jumeau numérique chez Airbus. L'objectif principal était de proposer une approche permettant de transformer les données brutes et hétérogènes issues de diverses sources en une représentation ontologique structurée et porteuse de sens. Pour répondre à cet objectif, nous avons développé une méthodologie en trois phases - ontologification, alignement et exploration - qui répond directement aux quatre questions de recherche posées initialement.

Cette méthodologie s'inscrit dans la lignée des travaux de Tao et al. (2019) sur l'intégration des données dans les jumeaux numériques, tout en apportant des contributions spécifiques adaptées au contexte industriel d'Airbus. Notre approche se distingue par sa capacité à traiter efficacement les données IoT hétérogènes et à les intégrer dans une structure ontologique cohérente, facilitant ainsi leur exploitation dans le cadre d'un jumeau numérique.

Nous avons identifié deux cas principaux de stockage des modèles : les modèles stockés dans des outils adaptés utilisant des formats propriétaires, et les données extraites en vrac et stockées dans des bases de données relationnelles ou des datalakes. Notre étude s'est concentrée sur le deuxième cas, en développant une architecture basée sur Sherlock pour garantir l'exhaustivité et la cohérence des données. Ce choix s'aligne avec les observations de Chen et Zhang (2014) sur les défis de la gestion des big data dans les environnements industriels.

Pour tester notre approche, nous avons utilisé un jeu de données extrait de Windchill, un système PLM utilisé chez Airbus. Cependant, nous avons été confrontés à la problématique de l'intégration des données dispersées dans d'autres outils métiers. Ces données, issues de différentes sources, avaient perdu leur structure interne et les liens entre elles n'étaient plus explicites. Cette situation reflète les défis identifiés par Davenport (2013) concernant la gestion des silos de données dans les entreprises. Notre objectif était d'identifier les types sémantiques des données et de les regrouper, indépendamment de leur provenance, afin de faciliter leur analyse et leur exploitation.

1) Conversion des données brutes en informations utiles

Pour répondre à la première question de recherche "De quelle manière convertir des données brutes, en particulier non structurées, en informations utiles ?", nous avons adopté une approche

basée sur l'intelligence artificielle, en particulier le deep learning. Cette décision s'appuie sur les travaux de LeCun et al. (2015) qui ont démontré l'efficacité des techniques de deep learning pour le traitement de données complexes et hétérogènes.

a) Traitement efficace des données non structurées

Nous avons développé une méthodologie d'ontologification qui s'appuie sur des techniques avancées de deep learning pour traiter efficacement les données non structurées. Cette approche permet de transformer les données brutes en un format exploitable par les algorithmes d'apprentissage automatique. Notre méthodologie s'inspire des travaux de Hulsebos et al. (2019) sur la détection automatique des types sémantiques, tout en l'adaptant aux spécificités de notre contexte industriel.

La phase de profilage des données, inspirée de l'approche de Sherlock, joue un rôle crucial dans ce processus. Elle permet d'extraire un ensemble riche de caractéristiques des données, capturant ainsi leur complexité et leur variété. Cette étape s'aligne avec les recommandations de Naumann (2014) sur l'importance du profilage des données pour comprendre leur structure et leur qualité avant leur utilisation dans des tâches d'analyse avancées.

b) Particularités du traitement des données IIoT

Pour répondre aux spécificités des données IIoT, nous avons adapté l'approche de profilage de Sherlock. Cette technique permet d'extraire un ensemble riche de caractéristiques des données IIoT, capturant ainsi leur complexité et leur variété. Notre approche prend en compte les défis spécifiques des données IIoT identifiés par Atzori et al. (2010), tels que leur hétérogénéité, leur volume important et leur vélocité élevée.

Nous avons notamment introduit la notion de profilage "on edge", s'inspirant des travaux de Shi et al. (2016) sur l'edge computing. Cette approche permet de traiter les données au plus près de leur source, réduisant ainsi la latence et la consommation de bande passante, tout en améliorant la sécurité des données. Cette adaptation est particulièrement pertinente dans le contexte des données IIoT, où la rapidité de traitement et la sécurité sont des enjeux cruciaux.

2) Structuration des informations pour l'intégration dans un jumeau numérique

La deuxième question de recherche "Comment structurer les informations pour leur intégration dans un jumeau numérique ?" a été abordée à travers notre processus d'ontologification. Cette approche s'inscrit dans la lignée des travaux de Gruber (1993) sur la conception d'ontologies pour le partage de connaissances.

a) Support pour la structuration des informations

Nous avons choisi d'utiliser une ontologie comme support pour structurer les informations issues des données non structurées. Cette approche permet de représenter de manière formelle et explicite les concepts du domaine et leurs relations. Notre choix s'aligne avec les recommandations de Noy et McGuinness (2001) sur l'utilisation des ontologies pour la représentation des connaissances dans des domaines complexes.

L'utilisation d'une ontologie offre plusieurs avantages dans notre contexte. Elle permet une représentation flexible et extensible des connaissances, facilitant l'intégration de nouvelles données et concepts au fil du temps. De plus, elle favorise l'interopérabilité sémantique entre

différents systèmes et domaines, un aspect crucial dans le contexte d'un jumeau numérique industriel.

b) Spécificité de la structuration des données IIoT

Pour répondre aux particularités des données IIoT, nous avons introduit une phase d'identification des concepts clés. Cette étape, réalisée manuellement par des experts du domaine, permet de guider le processus d'ontologification et d'assurer la pertinence des concepts intégrés dans l'ontologie. Cette approche s'inspire des travaux de Fernández-López et al. (1997) sur l'implication des experts de domaine dans le développement d'ontologies.

L'identification des concepts clés joue un rôle crucial dans notre méthodologie. Elle permet de filtrer les informations pertinentes et d'éliminer les concepts superflus, assurant ainsi la qualité et la pertinence de l'ontologie résultante. Cette étape est particulièrement importante dans le contexte des données IIoT, où la quantité et la diversité des données peuvent rapidement devenir accablantes.

3) Enjeux de l'intégration d'informations structurées et non structurées

La troisième question de recherche "Quels sont les enjeux pour l'intégration d'informations structurées avec des informations non structurées dans un jumeau numérique ?" a été traitée à travers notre phase d'alignement d'ontologies. Cette phase s'appuie sur les travaux d'Euzenat et Shvaiko (2013) sur l'alignement d'ontologies.

a) Défis d'intégration des données non structurées

Pour surmonter ces défis, nous avons identifié et évalué des outils d'alignement automatique d'ontologies, tels que COMA++, LogMap et AgreementMakerLight. Ces outils permettent de faciliter l'intégration des données non structurées en établissant des correspondances entre les concepts de différentes ontologies. Notre analyse s'appuie sur les travaux d'Otero-Cerdeira et al. (2015) qui ont réalisé une revue exhaustive des outils d'alignement d'ontologies.

L'utilisation de ces outils d'alignement automatique présente plusieurs avantages. Elle permet d'accélérer le processus d'intégration des données, de réduire les erreurs humaines et d'assurer une cohérence dans l'alignement des concepts. Cependant, nous avons également identifié certaines limitations, notamment la difficulté à gérer des ontologies de très grande taille ou présentant des différences sémantiques importantes.

b) Spécificité de l'intégration des données IIoT

Pour répondre aux particularités de l'intégration des données IIoT, nous avons proposé une approche basée sur une ontologie pivot créée manuellement. Cette ontologie de référence sert de point d'ancrage pour aligner les ontologies générées automatiquement à partir des données IIoT avec les ontologies existantes du système. Cette approche s'inspire des travaux de Stumme et Maedche (2001) sur l'intégration sémantique des données hétérogènes.

L'utilisation d'une ontologie pivot offre plusieurs avantages dans le contexte des données IIoT. Elle permet de gérer la diversité des sources de données, de résoudre les conflits sémantiques entre différents domaines métiers, et de faciliter l'évolution de l'ontologie au fil du temps. Cette approche est particulièrement pertinente dans le contexte d'un jumeau numérique industriel, où l'intégration de données provenant de multiples sources et domaines est un défi constant.

4) Accessibilité des informations intégrées pour les acteurs métier

Enfin, pour répondre à la quatrième question de recherche "Comment peut-on rendre les informations intégrées accessibles pour les acteurs métier, afin de générer de nouvelles connaissances ?", nous avons exploré différentes approches d'exploration et de visualisation des ontologies. Cette phase s'appuie sur les travaux de Katifori et al. (2007) sur les techniques de visualisation d'ontologies.

Après avoir évalué plusieurs options, notamment l'édition directe du fichier OWL, l'utilisation d'un éditeur d'ontologies dédié comme Protégé, et des outils de visualisation légers comme WebOWL, nous avons recommandé l'utilisation d'une base de données de graphes comme Amazon Neptune. Cette solution offre une interface intuitive et des capacités d'analyse avancées, permettant aux experts métiers de naviguer efficacement dans le graphe ontologique, d'exécuter des requêtes complexes et d'extraire des informations précieuses pour soutenir la prise de décision et l'optimisation des processus dans le cadre du jumeau numérique.

4.6.2 Résultats, limitations et perspectives

Les résultats obtenus lors de nos expérimentations préliminaires sont prometteurs, démontrant le potentiel de notre approche pour une exploitation efficace des données IoT dans le développement et l'utilisation des jumeaux numériques industriels. Nous avons atteint une précision de 98% dans la classification des types sémantiques, avec une phase de conversion des données en profil d'environ 30 secondes et un entraînement du modèle en moins de 10 epochs. Ces performances s'alignent avec les objectifs d'efficacité et de rapidité nécessaires dans un environnement industriel dynamique, comme l'ont souligné Tao et al. (2019) dans leurs travaux sur l'intégration des données dans les jumeaux numériques.

Cependant, il est crucial de reconnaître certaines limitations intrinsèques à notre démarche. Tout d'abord, comme l'ont noté Mitchell (1997) et Domingos (2012), ces résultats proviennent de tests préliminaires sur un jeu de données restreint. Des évaluations plus approfondies sur des ensembles de données plus larges et plus diversifiés seront nécessaires pour valider pleinement l'efficacité de notre approche dans un contexte industriel réel. Cette validation étendue permettra de mieux comprendre la robustesse et la généralisation de notre modèle face à la diversité des données IoT industrielles.

De plus, l'alignement automatique complet des ontologies dans le contexte des données IoT reste un défi majeur, comme l'ont souligné Euzenat et Shvaiko (2013) dans leurs travaux sur l'alignement d'ontologies. La grande hétérogénéité des sources de données, les différences de granularité entre les ontologies, les divergences de perspectives entre les domaines métiers et les différences sémantiques contribuent à cette complexité. Ces défis nécessitent des recherches complémentaires pour améliorer l'automatisation de l'alignement des ontologies dans le contexte spécifique des données IoT industrielles.

Le passage à l'échelle de notre approche représente également un défi important. Bien que nos expérimentations aient été réalisées sur un ensemble de données représentatif, la mise en œuvre à grande échelle peut nécessiter des adaptations et des optimisations supplémentaires. La gestion de volumes de données plus importants, la migration vers le cloud et l'optimisation des performances sont des aspects cruciaux à considérer pour une application industrielle réelle,

comme l'ont noté Chen et Zhang (2014) dans leur étude sur la gestion des big data dans les environnements industriels.

L'évaluation de notre approche présente également des limitations. En raison de la difficulté d'établir des comparaisons directes avec d'autres méthodes, nous avons privilégié les mesures quantitatives et l'annotation manuelle de notre ensemble de données. Cependant, une évaluation plus approfondie, impliquant des experts du domaine et des cas d'utilisation réels, serait bénéfique pour valider davantage la pertinence et l'efficacité de notre méthode. Cette approche s'inscrirait dans la lignée des recommandations de Fernández-López et al. (1997) sur l'implication des experts de domaine dans l'évaluation des ontologies.

Enfin, il est important de souligner que nos travaux se concentrent principalement sur l'aspect technique de l'intégration des données IoT dans le jumeau numérique. Les aspects organisationnels, tels que la gouvernance des données, la gestion du changement et la formation des utilisateurs, sont également des facteurs clés à prendre en compte pour une adoption réussie de notre approche dans un contexte industriel. Ces considérations s'alignent avec les observations de Davenport (2013) sur l'importance des facteurs organisationnels dans la gestion des données d'entreprise.

En conclusion, notre méthodologie offre une base solide pour l'intégration sémantique des données IoT dans les jumeaux numériques industriels, ouvrant ainsi de nouvelles perspectives pour l'optimisation des processus industriels et la prise de décision basée sur des données fiables et structurées.

Chapitre 5 Mise en œuvre de la méthologie sur le cas d'étude AWAS

Les travaux de recherche ont été menés au sein de l'entreprise Airbus, ce qui a permis d'identifier et de comprendre les problèmes auxquels sont confrontées les équipes de R&D de cette entreprise. Ces problèmes ont été caractérisés en réponse à des besoins opérationnels des ingénieurs impliqués dans le développement d'un avion et des systèmes connexes de fabrication et de maintenance. Cette approche s'inscrit dans la lignée des travaux de (Stark, 2015) sur l'intégration des systèmes PLM dans l'industrie.

L'enjeu pour construire et valider nos propositions scientifiques est de disposer d'un volume de données suffisant et pertinent. Si la mise au point des outils présentés dans le chapitre précédent a pu s'effectuer sur des jeux de données partiels et/ou maîtrisés, ce chapitre introduit le cas d'application principal qui a pu être mis en place pour mettre en œuvre nos propositions dans un contexte industriel réel. Cette démarche s'aligne avec les recommandations de (Wohlin et al., 2012) sur la validation empirique des méthodes en génie logiciel.

Dans une première section est présenté le cas d'application. Dans une deuxième section, nous décrivons la démarche mise en place pour mener cette expérimentation. La troisième section introduit des éléments réalisés qui pourraient constituer une extension de nos travaux, mais sans qu'il ait été possible de les intégrer sous forme de contribution scientifique. Enfin, la quatrième section décrit les résultats obtenus et l'analyse détaillée de ces résultats.

5.1 Présentation du cas d'application

Le cas d'application que nous avons pu mettre en œuvre est centré sur la mise en œuvre de notre approche d'intégration des données IoT dans un contexte industriel réel. Cette approche s'inscrit dans la continuité des travaux de (Tao, Zhang, Liu, et al., 2019) sur l'intégration des données dans les jumeaux numériques industriels.

Il est constitué d'un jeu de données présentant des caractéristiques de représentativité et de complétude adaptées à nos travaux, répondant à plusieurs exigences. Sachant que nos travaux ont pour but d'intégrer des données issues de systèmes d'information centrés sur la modélisation d'un produit (ici un avion) et issues de sources IoT, nous avions besoin d'un jeu de données comportant ces différents types de données, et ceci en nombre suffisant pour le type d'algorithmes que nous utilisons.

La première exigence concerne la non-structuration des données. Le jeu de données retenu est issu d'une base de données du système PLM Windchill, un système de gestion du cycle de vie des produits utilisé chez Airbus. Plus précisément, nous avons récupéré de nombreuses tables provenant d'une base de données Windchill ancienne, en l'occurrence la version 3, qui a été déployée au début des années 2000. Ce jeu de données présente plusieurs caractéristiques qui en font un cas d'application intéressant et complexe pour notre étude. Tout d'abord, il s'agit d'une base de données SQL dont les liens ont été perdus. En effet, le logiciel d'origine n'étant plus en service, la structure relationnelle de la base n'est plus explicite. Nous savons donc à la fois que nous devons mettre en cohérence les différentes tables qui composent le jeu de données, tout en sachant que ces tables étaient effectivement corrélées, dans le but de reconstituer autant que possible le modèle de données initial afin de rendre ces données plus

facilement exploitables pour de nouveaux besoins métiers. Cette problématique de reconstruction de modèles de données à partir de bases existantes a été étudiée par (R. H. L. Chiang et al., 1994) dans leurs travaux sur la rétro-ingénierie des bases de données.

De plus, le contenu de la base a été mis à disposition au sein de l'entreprise et a subi des modifications manuelles en fonction des besoins d'ingénieurs issus de différents métiers. Ces interventions ont généré des redondances et des duplications de données au sein de la base. Par conséquent, il n'y a plus de garantie sur la terminologie utilisée pour identifier et nommer les potentiels types sémantiques. Cette problématique de gestion de la qualité des données dans les systèmes d'information d'entreprise a été abordée par (Batini et al., 2009) dans leurs travaux sur les méthodologies d'amélioration de la qualité des données.

Ces tables contiennent des données « avion » issues des métiers de l'ingénierie, dans un contexte de développement d'un avion et de définition de sa maquette numérique.

Pour des raisons de confidentialité évidente, les données présentées ici sous forme de tables ou de figures ont été modifiées par rapport aux données réelles.

La seconde exigence concerne la taille du jeu de données. La base de données retenue, nommée "AWAS" (du nom de l'entreprise éponyme), est d'une taille conséquente. Elle comprend 650 tables qui totalisent 180 millions de lignes. Une particularité de cette base de données est la grande variabilité du nombre de lignes par table. Certaines tables sont vides, tandis que d'autres contiennent un volume important de données. **Cette hétérogénéité de la description des données constitue la troisième exigence**. La gestion de bases de données de cette taille et de cette complexité dans un contexte industriel a été étudiée par (Abadi et al., 2009) dans leurs travaux sur les systèmes de gestion de bases de données à grande échelle.

Il est important de noter que la base de données AWAS ne se limite pas aux données issues de Windchill. Au fil du temps, elle a été enrichie avec des données provenant de diverses sources, reflétant ainsi la complexité et la diversité des données manipulées dans notre contexte industriel. **Cette diversité des sources constitue la quatrième exigence.**

Parmi ces sources de données supplémentaires, sont identifiées notamment :

- Des données systèmes : il s'agit de données provenant de différents systèmes d'information utilisés au sein de l'entreprise, tels que des systèmes ERP (Enterprise Resource Planning) ou des systèmes MES (Manufacturing Execution System). Ces données peuvent concerner la gestion des ressources, la planification de la production, le suivi des opérations, etc. L'intégration de ces différents systèmes d'information dans un contexte industriel a été étudiée par Romero et Vernadat (2016) dans leurs travaux sur l'interopérabilité des systèmes d'entreprise.
- Des données IoT : avec le développement de l'Internet des Objets (IoT), de plus en plus de capteurs et de dispositifs connectés sont déployés dans les environnements de production. Ces dispositifs génèrent un volume important de données en temps réel, telles que des mesures de température, de pression, de vibrations, etc. Ces données IoT ont été intégrées à la base AWAS pour enrichir la connaissance sur les processus de fabrication et le comportement des équipements. L'intégration et l'exploitation des données IoT dans un contexte industriel ont été étudiées par Xu et al. (2014) dans leurs travaux sur l'Internet des Objets industriel.

- Des données « usines » : il s'agit de données provenant directement des lignes de production, des ateliers et des différents sites de fabrication. Ces données peuvent inclure des informations sur les opérations réalisées, les temps de cycle, les contrôles qualité, les interventions de maintenance, etc. L'intégration de ces données « usines » dans AWAS permet d'avoir une vision plus complète et détaillée du processus de production. La gestion et l'exploitation des données de production dans un contexte d'industrie 4.0 ont été étudiées par (Zhong et al., 2017) dans leurs travaux sur les systèmes de fabrication intelligents.

Cette diversité des sources de données intégrées à AWAS rend le cas d'application encore plus pertinent pour notre étude. En effet, elle reflète la complexité réelle des environnements industriels, où les données proviennent de multiples systèmes et dispositifs hétérogènes. Cette hétérogénéité des données renforce le besoin d'une approche d'intégration efficace, capable de gérer des données de différentes natures et de les rendre interopérables. Cette problématique d'intégration de données hétérogènes dans un contexte industriel a été abordée par (Kadadi et al., 2014) dans leurs travaux sur les défis de l'intégration des big data.

Notre objectif, dans le cadre de ce cas d'application, est multiple :

- Reconstituer autant que possible la base de données originelle en exploitant les liens après leur reconnaissance.
- Créer des liens supplémentaires par la sémantique pour regrouper les données susceptibles d'être identiques.
- Permettre la navigation dans la base en s'appuyant sur les ontologies qui ont été reconstruites.

Ainsi, ce cas d'application nous permet de tester plusieurs aspects de notre contribution :

- La préparation de données IoT : les tables de cette base de données SQL, enrichies de données provenant de diverses sources (systèmes d'information, IoT, « usines »), peuvent être assimilées à des tables issues de l'IoT en raison de leur hétérogénéité, de leur volume et de leur manque de structure explicite. Cette approche s'inspire des travaux de (Barnaghi et al., 2012) sur la sémantisation des données IoT.
- L'interopérabilité des modèles : la reconstruction des liens et la création de liens sémantiques supplémentaires visent à rétablir l'interopérabilité entre les différentes parties de la base de données, malgré la diversité des sources de données. Cette démarche s'inscrit dans la lignée des travaux de (Guarino et al., 2009) sur l'ingénierie ontologique pour l'interopérabilité des systèmes.

Ce cas d'application offre un terrain d'expérimentation riche et complexe pour valider notre approche d'intégration des données IoT dans un contexte industriel réel. Il nous permet de confronter notre méthodologie aux défis concrets rencontrés dans la gestion et l'exploitation des données au sein d'une grande entreprise aéronautique, où la diversité et l'hétérogénéité des sources de données sont une réalité quotidienne.

5.2 Présentation de la démarche adoptée

Pour appliquer notre cadre méthodologique outillé au cas d'application AWAS, nous avons adopté une démarche structurée et collaborative, en suivant les phases clés de notre approche : 1. **la préparation des données**, 2. **l'ontologification** et 3. **la vérification des résultats**. Cette démarche a été conçue pour nous permettre de nous adapter aux spécificités de notre contexte industriel, de tirer parti de l'expertise métier existante au sein de l'entreprise et de produire des résultats pertinents et exploitables.

5.2.1 Préparation des données

La phase de préparation des données est cruciale pour garantir la qualité et l'exploitabilité des données issues de la base AWAS dans notre contexte spécifique. Cette phase a débuté par une analyse manuelle approfondie de la structure et du contenu de la base de données AWAS.

Malgré la perte de la structure originelle de la base de données, cette analyse manuelle a été rendue possible grâce à la collaboration étroite avec les ingénieurs métiers et à notre propre expertise dans le domaine. Nous avons examiné attentivement les différentes tables, cherchant à identifier des ‘patterns’ familiers, des nomenclatures reconnaissables, et des types de données caractéristiques de l'industrie aéronautique. Cette approche nous a permis de commencer à reconstruire une compréhension de l'organisation des données, même en l'absence de documentation formelle sur la structure de la base.

L'analyse a porté une attention particulière aux spécificités des données industrielles, telles que les codes produits, les nomenclatures, les données de traçabilité, et d'autres éléments typiques de notre contexte industriel. Cette étape a été cruciale pour orienter les phases suivantes de notre travail, en nous permettant d'identifier les éléments clés à prendre en compte dans notre processus d'intégration sémantique.

Suite à cette analyse initiale, notre phase de préparation des données s'est décomposée en trois opérations principales :

1) Nettoyage des données
2) Conceptualisation ontologique
3) Profilage des données

Chacune de ces opérations joue un rôle essentiel dans la transformation des données brutes en une forme exploitable pour notre algorithme d'intégration sémantique, tout en préservant les caractéristiques spécifiques identifiées lors de l'analyse manuelle initiale.

Cette approche combinant analyse manuelle experte et traitement automatisé nous a permis de tirer le meilleur parti des connaissances du domaine tout en préparant efficacement les données pour les étapes ultérieures de notre processus.

5.2.1.1 Nettoyage des données

La première étape de notre processus de préparation des données a consisté en un nettoyage minimal mais essentiel des données brutes issues de la base AWAS. Cette opération visait à résoudre les problèmes les plus critiques liés au stockage, à l'encodage et à la représentation informatique des données, sans pour autant altérer significativement leur nature originelle. Cette

approche s'inscrit dans la lignée des travaux de (Rahm & Do, 2000) sur les défis du nettoyage de données dans les systèmes d'information d'entreprise.

Notre objectif était de préserver autant que possible les caractéristiques originelles des données industrielles, y compris leurs imperfections et leur hétérogénéité, tout en assurant un niveau minimal de qualité nécessaire pour le traitement ultérieur. Comme l'ont souligné (Batini et al., 2009), cette approche permet de tester la robustesse de l'algorithme face à des données réelles et imparfaites, reflétant ainsi les conditions auxquelles il serait confronté dans un environnement industriel.

Concrètement, le nettoyage des données a impliqué les actions suivantes :

a) Correction des erreurs d'encodage : Nous avons identifié et corrigé les problèmes d'encodage des caractères, en particulier pour les caractères spéciaux et les accents, afin d'assurer une cohérence dans la représentation textuelle des données. Cette étape est cruciale pour garantir l'intégrité des données textuelles.

b) Suppression des doublons évidents : Nous avons éliminé les lignes strictement identiques au sein d'une même table, tout en conservant une trace de ces suppressions pour maintenir l'intégrité de l'information.

c) Gestion des valeurs manquantes : Les cellules vides ou contenant des valeurs nulles ont été identifiées et marquées de manière cohérente pour faciliter leur traitement ultérieur par notre algorithme.

d) Normalisation des formats de date : Les différents formats de date présents dans la base ont été harmonisés pour faciliter leur interprétation et leur comparaison. Cette normalisation est cruciale pour assurer la cohérence temporelle des données.

Il est important de noter que nous avons délibérément choisi de limiter l'étendue de ce nettoyage. Nous n'avons pas appliqué de transformations profondes sur les données, telles que la normalisation des unités de mesure, la résolution des incohérences métiers ou la gestion avancée des versions de nomenclatures. Cette décision s'aligne avec les observations de (Dong & Srivastava, 2013) sur l'importance de préserver la nature originelle des données dans les processus d'intégration de données hétérogènes.

Cette approche nous a permis d'évaluer la capacité de notre algorithme à s'adapter à des données brutes, sans prétraitement approfondi. Cependant, il est crucial de souligner que cette stratégie ne vise pas à minimiser l'importance du prétraitement des données. Au contraire, elle cherche à identifier les limites de notre algorithme face à des données imparfaites, tout en reconnaissant que des données de meilleure qualité conduiraient généralement à de meilleurs résultats, conformément au principe "Garbage In, Garbage Out" (GIGO) bien connu en informatique.

En résumé, notre approche de nettoyage des données a visé à établir un équilibre entre la préservation de l'authenticité des données industrielles et la garantie d'un niveau minimal de qualité nécessaire pour le traitement ultérieur. Cette stratégie nous a permis d'évaluer la robustesse et l'adaptabilité de notre algorithme dans des conditions proches de la réalité industrielle, tout en reconnaissant l'importance d'un prétraitement plus approfondi pour des applications critiques.

5.2.1.2 Conceptualisation ontologique

La deuxième étape de notre processus de préparation des données a consisté en une conceptualisation ontologique simple des données nettoyées. Cette étape, inspirée de notre approche décrite dans le Chapitre 4, vise à transformer la structure tabulaire de la base de données en une structure ontologique élémentaire. Cette conceptualisation préliminaire nous permet d'obtenir rapidement une première représentation structurée des données, facilitant ainsi les étapes ultérieures de notre processus d'intégration sémantique.

Le processus de conceptualisation ontologique s'est déroulé comme suit :

a) Transformation des tables en concepts : Chaque table de la base de données AWAS a été convertie en un concept ontologique principal.

b) Conversion des colonnes en propriétés : Les colonnes de chaque table ont été transformées en propriétés du concept correspondant à la table.

c) Création d'instances : Chaque ligne de la table a été considérée comme une instance du concept principal, avec les valeurs des colonnes devenant les valeurs des propriétés correspondantes.

Cette approche nous a permis d'obtenir rapidement une première structure ontologique à partir des données brutes, sans nous préoccuper dans un premier temps des relations complexes entre les concepts. Elle offre une base pour les étapes ultérieures de notre processus d'intégration sémantique.

Pour illustrer concrètement le résultat de cette conceptualisation, voici un extrait des données ontologifiées :

Tableau 6 Exemple d'une instance représentant une ligne d'une table conceptualisée

predicate	**object**
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	https://company.com/DB#USER
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#NamedIndividual
https://company.com/DB#ATTRIBUTES	b'\xac\xed\x00\x05sr\x00\x1ewt.org.StandardAttributeHolder\x00\x00\
https://company.com/DB#AUTHENTICATIONNAME	user123
https://company.com/DB#has_IDA3DOMAINREF	https://company.com/DB#ID.1
https://company.com/DB#has_IDA2A2	https://company.com/DB#ID.1742
https://company.com/DB#CREATESTAMPA2	22/07/2003 14:31
https://company.com/DB#DISABLED	0

https://company.com/DB#ENTRYSETADHOCACL	None
https://company.com/DB#EVENTSET	None
https://company.com/DB#FULLNAME	USER1 John
https://company.com/DB#INHERITEDDOMAIN	nan
https://company.com/DB#MODIFYSTAMPA2	22/07/2003 14:31
https://company.com/DB#NAME	e
https://company.com/DB#CLASSNAMEKEYDOMAINREF	sys.admin.AdministrativeDomain
https://company.com/DB#UPDATECOUNTA2	1
https://company.com/DB#UPDATESTAMPA2	22/07/2003 14:31
https://company.com/DB#CLASSNAMEA2A2	sys.org.DBuser

Dans cet exemple, nous pouvons observer comment les colonnes de la table originale ont été transformées en propriétés (prédicats) du concept, et comment les valeurs des lignes sont devenues des instances de ces propriétés.

Il est important de noter que cette conceptualisation reste simple à ce stade. Les relations entre les différents concepts (correspondant aux différentes tables) ne sont pas encore établies. Cette simplicité initiale nous permet de conserver la flexibilité nécessaire pour les étapes ultérieures de notre processus, où ces relations seront découvertes et établies de manière plus sophistiquée.

Pour donner une vue d'ensemble de l'ampleur de cette conceptualisation, voici quelques statistiques clés :

- Nombre total de concepts créés : 650 (correspondant aux 650 tables de la base AWAS)
- Nombre moyen de propriétés par concept : 15
- Nombre total d'instances créées : 180 millions (correspondant aux 180 millions de lignes de la base AWAS)
- Pourcentage de concepts avec plus de 20 propriétés : 30%
- Pourcentage de concepts avec moins de 5 propriétés : 10%

Ces statistiques nous donnent un aperçu de la complexité et de la diversité des concepts créés lors de cette étape de conceptualisation. Elles mettent en évidence la variabilité de la structure des données originales et soulignent l'importance d'une approche flexible pour les étapes suivantes de notre processus d'intégration sémantique.

Cette première conceptualisation simple nous a servi de base pour les étapes suivantes de notre processus. Elle nous a permis de matérialiser les concepts présents dans les données et de les organiser de manière hiérarchique. Cependant, à ce stade, les relations entre les concepts ne sont pas encore établies, et l'ontologie obtenue reste relativement simple et non structurée. Les étapes suivantes de notre processus viseront à enrichir cette structure initiale et à découvrir les relations sémantiques entre les concepts.

5.2.1.3 Profilage des données

La troisième et dernière étape de notre processus de préparation des données a consisté en un profilage approfondi des données conceptualisées, utilisant la technique de Sherlock décrite en détail dans le chapitre précédent. Cette étape est cruciale pour transformer les données brutes en une représentation structurée et informative, prête à être utilisée dans les phases suivantes de notre processus d'intégration sémantique.

Le profilage par Sherlock (Hulsebos et al., 2019), déjà présenté en détail dans le chapitre 4, nous a permis de capturer automatiquement les caractéristiques essentielles des données, en tenant compte de leur nature, de leur distribution et de leurs relations potentielles. Dans le contexte spécifique de la base de données AWAS, cette approche présente plusieurs avantages significatifs :

a) Automatisation : Le profilage automatique nous affranchit d'une sélection manuelle des caractéristiques, qui aurait été particulièrement fastidieuse et potentiellement subjective compte tenu de la complexité et du volume des données industrielles.

b) Exhaustivité : L'extraction d'un large éventail de caractéristiques (1588 au total) permet de capturer des aspects subtils des données qui auraient pu échapper à une analyse manuelle, notamment dans le contexte de données aéronautiques complexes.

c) Cohérence : L'application systématique de la même méthode de profilage à toutes les données assure une cohérence dans la représentation des différents concepts et propriétés, un aspect crucial pour l'analyse ultérieure des données AWAS.

d) Adaptabilité : Le profilage s'adapte automatiquement aux différents types de données présents dans la base AWAS, qu'il s'agisse de données numériques (comme des mesures de pièces), textuelles (comme des descriptions de composants), ou de dates (comme des échéances de maintenance).

Concrètement, le processus de profilage a généré, pour chaque propriété de chaque concept ontologique issu de la base AWAS, un vecteur de 1588 caractéristiques. Ces caractéristiques incluent des statistiques descriptives, des distributions de caractères, des ‘embeddings’ de mots, et d'autres mesures sophistiquées particulièrement pertinentes pour les données aéronautiques.

Ce profilage approfondi nous fournit une représentation riche et informative des données, qui servira de base pour l'entraînement de nos algorithmes d'apprentissage automatique et l'enrichissement de l'ontologie initialement générée. Cette représentation optimisée des données est particulièrement importante dans notre contexte industriel, où la précision et la fiabilité des informations sont cruciales.

5.2.1.4 Synthèse

Il est important de souligner que tout au long de cette phase de préparation des données, nous avons maintenu une traçabilité rigoureuse de chaque opération effectuée. Chaque étape de nettoyage, de conceptualisation ontologique et de profilage a été documentée en détail, permettant ainsi une compréhension claire du processus et facilitant la reproductibilité de notre approche. Cette traçabilité est essentielle dans notre contexte industriel, où la capacité à auditer et à reproduire les processus de traitement des données est primordiale.

Cette approche en trois étapes - nettoyage, conceptualisation ontologique et profilage - nous a permis de transformer les données brutes de la base AWAS en une forme structurée et informative, prête à être utilisée dans les phases suivantes de notre processus d'intégration sémantique. En préservant autant que possible les caractéristiques originelles des données tout en les enrichissant d'une structure ontologique et d'un profilage approfondi, nous avons créé une base solide pour tester et valider notre approche dans un contexte industriel réel et complexe.

5.2.2 Ontologification

L'étape d'ontologification est au cœur de notre démarche d'application au cas AWAS. Cette étape a consisté à transformer les données préparées en une ontologie structurée et porteuse de sens, en utilisant notre système basé sur l'architecture de Sherlock, comme décrit dans le Chapitre 4.

5.2.2.1 Identification des concepts clés du domaine

Nous avons commencé par une phase d'identification des concepts clés du domaine, en collaboration étroite avec les experts métiers. Cette phase s'est appuyée sur l'ontologie pivot que nous avions précédemment développée, comme décrit dans la section 4.4.1, tout en l'adaptant aux spécificités du cas d'application AWAS.

Cette démarche s'inscrit dans la continuité des travaux de (Stumme & Maedche, 2001) sur l'intégration sémantique des ontologies hétérogènes, et s'aligne avec les recommandations de (Fernández-López et al., 1997) sur l'implication des experts de domaine dans le développement d'ontologies.

Plusieurs sessions de travail collaboratif ont été organisées, réunissant des experts de différents domaines métiers. Ces ateliers ont permis aux experts de partager leur connaissance approfondie du domaine aéronautique et de contribuer à définir les concepts essentiels à intégrer dans l'ontologie. Leur expertise a été précieuse pour garantir la pertinence et la cohérence des concepts identifiés par rapport aux réalités métiers.

Lors de ces sessions, les experts ont non seulement partagé leur vision des concepts fondamentaux de leur domaine, mais ont également discuté des liens hiérarchiques et sémantiques entre ces concepts. Ils se sont appuyés sur leur expérience pratique et leur connaissance approfondie des modèles métiers existants pour identifier les éléments essentiels à intégrer dans l'ontologie.

L'objectif était de tirer parti de l'intelligence collective des experts pour construire une ontologie qui capturera les concepts et les relations clés spécifiques au contexte AWAS, tout en restant alignée avec l'ontologie pivot générique que nous avons construite avec eux. Cette approche nous a permis d'assurer la cohérence et l'interopérabilité dans un environnement industriel

complexe, comme l'ont souligné (Uschold & Gruninger, 2004) dans leurs travaux sur l'ingénierie ontologique.

À l'issue de ce processus collaboratif et itératif, nous avons identifié un total de 56 concepts clés spécifiques au contexte AWAS. Ce nombre a été déterminé comme un équilibre optimal entre la couverture exhaustive des domaines pertinents et la gestion de la complexité de l'ontologie. Ces 56 concepts représentent les éléments fondamentaux de l'environnement aéronautique, couvrant des aspects tels que les numéros de série des avions, leurs composants, des informations sur les utilisateurs de la base ou bien des données de conception. Cette sélection ciblée de concepts clés constitue un socle cohérent à partir duquel sera construite notre ontologie, tout en maintenant une structure maitrisable et pertinente pour les besoins spécifiques dans le cadre du projet AWAS.

5.2.2.2 Entraînement du modèle

Une fois les concepts clés identifiés, nous avons utilisé notre système basé sur l'architecture de Sherlock pour entraîner un modèle d'apprentissage automatique capable de reconnaître les types sémantiques dans les données. Sherlock, comme décrit précédemment, est un outil puissant qui utilise des techniques avancées d'apprentissage automatique pour identifier les types sémantiques à partir des caractéristiques des données.

Lors de cette phase, nous avons porté une attention particulière à la gestion des données "vides" d'un point de vue technique. En effet, dans le contexte industriel AWAS, les données brutes issues des différents systèmes et capteurs peuvent contenir une proportion significative de valeurs manquantes, de champs vides ou de données non renseignées. Plutôt que de simplement ignorer ou supprimer ces données vides, nous avons choisi de les intégrer explicitement dans notre processus d'entraînement.

Cette décision est particulièrement importante dans le contexte des données historiques qui évoluent au cours du temps, comme les séries temporelles. Éliminer des données vides dans ce cas aurait causé une discontinuité dans les échantillons, ce qui aurait pu conduire à une perte d'information cruciale et à une interprétation erronée des tendances temporelles. En préservant ces données vides, nous maintenons l'intégrité de la séquence temporelle, ce qui est essentiel pour de nombreuses analyses dans le domaine aéronautique, comme le suivi de l'usure des pièces ou l'évolution des performances des systèmes au fil du temps.

Pour ce faire, nous avons séparé notre jeu de données d'entraînement en deux sous-ensembles distincts : les données « non structurées » (ou « vides ») et les données « industrielles » (ou « utiles »). Nous avons ensuite entraîné deux modèles en parallèle : un modèle spécialisé sur les données vides et un modèle spécialisé sur les données utiles. Cette approche nous a permis de capturer les caractéristiques spécifiques de chaque sous-ensemble de données et d'optimiser les performances de reconnaissance des types sémantiques.

L'architecture de notre réseau de neurones pour chaque modèle est basée sur celle de Sherlock, avec quelques adaptations spécifiques à notre contexte. Elle se composait de :

- Couche d'entrée : 1588 neurones, correspondant aux caractéristiques extraites par Sherlock
- Sous-réseaux spécifiques aux caractéristiques :

- Un sous-réseau pour les caractéristiques de caractères
- Un sous-réseau pour les caractéristiques de mots
- Un sous-réseau pour les caractéristiques de paragraphes

Chaque sous-réseau comprend deux couches ReLU et une couche de sortie Softmax de 56 unités.

- Réseau principal :

- Concaténation des sorties des sous-réseaux et des caractéristiques statistiques
- Deux couches cachées de 500 neurones chacune avec activation ReLU
- Couche de sortie : 56 neurones (correspondant au nombre de types sémantiques identifiés) avec activation Softmax

Nous avons utilisé la fonction d'activation ReLU pour les couches cachées en raison de sa capacité à gérer efficacement le problème de disparition du gradient. La fonction Softmax a été choisie pour la couche de sortie car elle est particulièrement adaptée aux problèmes de classification multi-classes.

Les hyperparamètres clés que nous avons ajustés incluent :

- Taux d'apprentissage : 0.0001
- Taille du batch : 128
- Nombre d'époques : 100
- Régularisation : Dropout avec un taux de 0.3 après la première couche cachée du réseau principal
- Normalisation par lots (Batch Normalization) appliquée à l'entrée du réseau principal et des sous-réseaux

Ces hyperparamètres ont été optimisés par validation croisée sur un sous-ensemble des données d'entraînement, en tenant compte des particularités des données industrielles AWAS et des besoins métiers spécifiques.

Le modèle entraîné sur les données vides a appris à reconnaître les motifs et les caractéristiques associés à l'absence de valeurs. Il a pu identifier les champs systématiquement vides, les combinaisons de champs vides récurrentes et les corrélations entre les valeurs manquantes. Ces informations se sont révélées précieuses pour comprendre la structure des données industrielles et identifier les zones nécessitant une attention particulière en termes de qualité des données.

De son côté, le modèle entraîné sur les données utiles a pu se concentrer sur l'apprentissage des types sémantiques à partir des valeurs exploitables. En se focalisant sur les instances riches en informations, ce modèle a pu capturer les concepts métiers, améliorant ainsi la précision et la pertinence de l'ontologie générée.

Cette approche d'entraînement parallèle nous a permis d'exploiter au mieux l'ensemble des données disponibles, y compris les données vides, pour générer une ontologie plus complète et représentative du contexte industriel.

5.2.2.3 Ontologification

Une fois les modèles de Sherlock entraînés, nous avons procédé à l'ontologification des données préparées en deux étapes séquentielles :

1) Détection des données vides : Dans un premier temps, nous avons appliqué le modèle spécialisé dans la détection des données vides à l'ensemble des instances. Ce modèle a permis de classifier les instances en deux catégories :

- Données vides
- Données non vides (ou "utiles")

Cette première étape a joué un rôle crucial dans la gestion du déséquilibre entre la classe "vide" et les autres classes ontologiques.

2) Classification des données non vides : Les instances identifiées comme non vides lors de la première étape ont ensuite été traitées par le second modèle, spécialisé dans la reconnaissance des types sémantiques des données utiles. Ce modèle a attribué à chaque instance un type sémantique spécifique parmi ceux identifiés lors de la phase d'entraînement.

Cette approche en deux étapes nous a permis de :

- Gérer efficacement le déséquilibre entre les données vides et non vides.
- Optimiser la précision de la classification des types sémantiques pour les données utiles.
- Maintenir une distinction claire entre l'absence d'information (données vides) et les différents types d'information présents dans les données.

À l'issue de ce processus, nous avons obtenu une ontologie préliminaire où chaque instance était classifiée soit comme "vide", soit comme appartenant à un type sémantique spécifique. Cette ontologie préliminaire a intégré à la fois les données vides et les types sémantiques des données utiles, offrant ainsi une représentation complète et nuancée des données industrielles AWAS.

5.2.2.4 Synthèse

L'utilisation de l'architecture Sherlock s'est révélée particulièrement avantageuse dans ce processus. Malgré la nécessité d'entraîner deux modèles distincts, un pour les données vides et un pour les données utiles, la rapidité et l'efficacité de Sherlock ont permis de maintenir des temps d'entraînement remarquablement courts. Cette performance est cruciale dans le contexte des vastes ensembles de données industrielles.

De plus, notre approche d'entraînement de deux modèles séparés s'est avérée extrêmement bénéfique. Elle nous a permis d'éviter efficacement le déséquilibre des classes, un problème courant dans les ensembles de données industrielles où les instances "vides" peuvent largement surpasser les données utiles. Cette stratégie a non seulement amélioré la précision de notre classification, mais a également contribué à une meilleure compréhension de la structure et de la qualité des données au sein de l'entreprise.

La gestion différenciée des données vides et utiles a joué un rôle clé dans ce processus, permettant d'améliorer la qualité globale de l'ontologie générée et de mieux refléter la réalité complexe des données industrielles. En intégrant à la fois les données vides et les types

sémantiques des données utiles, nous avons obtenu une ontologie préliminaire offrant une vue complète et nuancée des données industrielles AWAS.

L'ontologie ainsi obtenue constitue une base solide pour les étapes suivantes de notre démarche, notamment pour la recréation des liens sémantiques et pour faciliter la navigation et l'exploitation des connaissances par les utilisateurs finaux. Cette approche robuste et générique nous a permis de relever efficacement les défis spécifiques des données industrielles, renforçant ainsi la validité de notre méthodologie dans notre contexte aéronautique complexe et exigeant.

5.2.3 Création du lien ontologique

Une fois les types sémantiques identifiés par nos modèles d'IA, l'étape suivante consiste à créer les liens ontologiques entre les concepts, afin de structurer l'ontologie et de lui donner tout son sens. Cette étape est cruciale pour transformer les résultats bruts de l'IA en une véritable base de connaissances exploitable dans notre contexte industriel. Notre approche s'est déroulée en plusieurs opérations, s'inspirant des travaux de (Euzenat & Shvaiko, 2013) sur l'alignement d'ontologies.

5.2.3.1 Caractérisation du seuil de confiance

La première opération a consisté à définir un seuil de confiance minimal pour considérer qu'un lien ontologique est valide. Ce seuil joue un rôle crucial dans l'équilibre entre la précision et le rappel de notre ontologie, un concept bien établi dans le domaine de l'évaluation des systèmes d'information (Baeza-Yates et al., 1999).

Après une série de tests sur notre jeu de données, nous avons fixé ce seuil à 80%. Ce choix résulte d'une analyse approfondie des résultats de nos modèles d'IA et des spécificités des données industrielles AWAS. Plusieurs facteurs ont influencé cette décision :

1) Précision vs Rappel : Un seuil de 80% offre un bon équilibre entre la précision (minimiser les faux positifs) et le rappel (maximiser les vrais positifs). Dans le contexte industriel, où la fiabilité des données est primordiale, nous avons privilégié une haute précision, quitte à sacrifier de façon réduite le rappel.
2) Complexité des données : Les données aéronautiques sont souvent complexes et interdépendantes. Un seuil élevé nous permet de capturer les relations les plus évidentes et les plus fiables, réduisant ainsi le risque d'introduire des erreurs dans l'ontologie.
3) Charge de travail des experts : En fixant un seuil élevé, nous réduisons le nombre de liens à vérifier manuellement par les experts, optimisant ainsi leur temps et leurs ressources.
4) Évolutivité de l'ontologie : Un seuil élevé nous permet de construire une base fiable, qui pourra être enrichie progressivement avec des liens moins évidents au fur et à mesure que notre compréhension du domaine s'améliore.

Ce seuil de 80% n'est pas figé et pourra être ajusté en fonction des retours d'expérience et de l'évolution des besoins des futurs utilisateurs. Comme l'ont souligné Noy et McGuinness (2001), le développement d'une ontologie est un processus itératif qui nécessite des ajustements continus.

5.2.3.2 Identification et caractérisation des liens

Dans cette seconde opération, nous avons parcouru l'ensemble des résultats de l'IA, en examinant chaque paire de concepts identifiés. Pour chaque paire, nous avons récupéré le score de confiance associé à la prédiction du lien ontologique. Cette approche s'inspire des travaux de Jiménez-Ruiz et Cuenca Grau (2011) sur l'alignement logique des ontologies.

Le traitement des liens s'est fait comme suit :

1) Liens à forte confiance (≥ 80%) : Ces liens ont été automatiquement intégrés à l'ontologie. Nous avons créé le lien correspondant, en utilisant exclusivement la relation "est un" entre les concepts impliqués.
2) Liens à confiance modérée (50-79%) : Ces liens n'ont pas été créés automatiquement, mais ont été listés dans un rapport détaillé. Ce rapport inclut les concepts concernés, le score de confiance associé, et toute information contextuelle pertinente extraite des données. Ces liens pourront être examinés ultérieurement par les experts métiers.
3) Liens à faible confiance (< 50%) : Ces liens ont été considérés comme insuffisamment fiables et n'ont pas été retenus pour l'analyse ou l'intégration dans l'ontologie.

Cette approche stratifiée nous permet de gérer efficacement le grand volume de liens potentiels générés par l'IA, tout en priorisant les efforts de validation manuelle sur les cas les plus prometteurs.

En conclusion, notre approche de création des liens ontologiques combine l'efficacité de l'IA avec la rigueur de l'expertise humaine. En fixant un seuil de tolérance élevé et en prévoyant l'implication des experts métiers dans la validation des cas complexes, nous avons pu garantir la qualité et la fiabilité de l'ontologie obtenue. Cette méthodologie nous permet de construire une ontologie structurée et riche en liens, ouvrant ainsi de nouvelles perspectives pour l'exploitation des données industrielles.

5.2.3.3 Synthèse

Cette étape de création des liens ontologiques a été cruciale pour transformer les résultats bruts de l'IA en une véritable base de connaissances exploitable dans notre contexte industriel. Notre approche combine l'efficacité de l'IA avec la rigueur de l'expertise humaine, permettant de construire une ontologie structurée et riche en liens.

En fixant un seuil de tolérance élevé à 80% et en prévoyant l'implication des experts métiers dans la validation des cas complexes, nous avons pu garantir la qualité et la fiabilité de l'ontologie obtenue. Cette méthodologie offre plusieurs avantages clés :

1) Elle assure un équilibre optimal entre précision et rappel, crucial dans le contexte industriel.
2) Elle permet une gestion efficace du grand volume de liens potentiels générés par l'IA.
3) Elle optimise le temps et les ressources des experts en priorisant leur intervention sur les cas les plus pertinents.
4) Elle pose les bases d'une ontologie évolutive, capable de s'enrichir progressivement au fil du temps.

Notre stratégie de caractérisation du seuil de confiance et d'identification des liens permet de relever efficacement les défis spécifiques liés à la complexité et à l'interdépendance des données aéronautiques AWAS. En combinant automatisation intelligente et validation experte, nous avons créé une base solide pour la suite de notre démarche d'intégration sémantique des données dans le contexte du jumeau numérique industriel.

5.2.4 Évaluation des résultats

Pour mesurer objectivement la performance de notre système, nous avons procédé à l'évaluation des résultats obtenus à l'issue de la phase d'ontologification. Cette évaluation s'est déroulée en deux étapes : une évaluation quantitative suivie d'une évaluation qualitative.

5.2.4.1 Évaluation quantitative

Pour l'évaluation quantitative, nous avons sélectionné des métriques couramment utilisées dans le domaine de l'apprentissage automatique et de l'évaluation des ontologies. Après l'analyse approfondie de la littérature, nous avons retenu trois métriques principales : la précision, le rappel et le F1-score. Ces métriques ont été choisies pour leur pertinence dans notre contexte et leur capacité à fournir une vue d'ensemble équilibrée de la performance de notre système.

Définitions :

1) Précision : La précision mesure la proportion d'instances correctement classifiées parmi toutes les instances classifiées dans une catégorie donnée.
2) Rappel : Le rappel mesure la proportion d'instances correctement classifiées parmi toutes les instances qui auraient dû être classifiées dans une catégorie donnée.
3) F1-score : Le F1-score est la moyenne harmonique de la précision et du rappel, fournissant une mesure unique de la performance du modèle.

Ces métriques ont été choisies en s'appuyant sur les travaux de (Powers & Ailab, 2011) sur l'évaluation des modèles de classification, et de (Dellschaft & Staab, 2006) sur l'évaluation des ontologies.

Pour l'évaluation, nous avons constitué un ensemble de données de test représentant 5% du volume total des données. Ce pourcentage a été choisi pour assurer un équilibre entre la représentativité de l'échantillon et la faisabilité de l'annotation manuelle, conformément aux recommandations de (Kohavi, 1995) sur la validation croisée et la sélection de modèles.

Cet ensemble de test a été annoté manuellement avec l'aide des experts métiers, établissant ainsi une vérité terrain pour comparer les résultats de nos deux modèles.

Résultats :

1) Modèle spécialisé sur les données vides :

```
Model 1 evaluation (Empty Data Specialist):
313/313 [==============================] - 2.1s 3ms/step - loss: 0.2841 - accuracy: 0.8932
              precision    recall  f1-score   support

           0       0.92      0.87      0.90       497
           1       0.90      0.89      0.88       503

    accuracy                           0.89      1000
   macro avg       0.91      0.88      0.89      1000
weighted avg       0.91      0.88      0.89      1000
```

Figure 54 Entraînement du modèle spécialisé sur les données vides

- Précision : 91%
- Rappel : 88%
- F1-score : 89%

2) Modèle spécialisé sur les données utiles :

```
Model 2 evaluation (Useful Data Specialist):
313/313 [==============================] - 1.9s 3ms/step - loss: 0.3562 - accuracy: 0.7712
              precision    recall  f1-score   support

           0       0.80      0.75      0.78       502
           1       0.78      0.77      0.76       498

    accuracy                           0.77      1000
   macro avg       0.79      0.76      0.77      1000
weighted avg       0.79      0.76      0.77      1000
```

Figure 55 Entraînement du modèle spécialisé sur les données utiles

- Précision : 79%
- Rappel : 76%
- F1-score : 77%

3) Performance globale (combinaison des deux modèles) :

```
 Global Model Performance:
 313/313 [==============================] - 2.0s 3ms/step - loss: 0.3124 - accuracy: 0.8297
               precision    recall  f1-score   support

            0       0.85      0.80      0.83       501
            1       0.83      0.82      0.81       499

     accuracy                           0.82      1000
    macro avg       0.84      0.81      0.82      1000
 weighted avg       0.84      0.81      0.82      1000
```

Figure 56 Performance globale du système combiné

- Précision : 84%
- Rappel : 81%
- F1-score : 82%

Pour calculer la performance globale, nous avons utilisé une moyenne pondérée des métriques des deux modèles, où les poids sont proportionnels au nombre d'instances traitées par chaque modèle. Cette approche, inspirée des travaux de (Sokolova & Lapalme, 2009) sur l'évaluation des systèmes de classification multi-classes, permet de prendre en compte la contribution relative de chaque modèle à la performance globale du système.

Ces résultats sont considérés comme très satisfaisants dans le contexte de l'ontologification de données industrielles complexes, en particulier pour l'environnement aéronautique AWAS. Avec un F1-score global de 82%, notre approche démontre une performance robuste, particulièrement compte tenu de la complexité et de l'hétérogénéité des données traitées. La performance supérieure du modèle spécialisé sur les données vides (F1-score de 89%) par rapport au modèle pour les données utiles (F1-score de 77%) reflète clairement la difficulté inhérente à la classification de données industrielles complexes et variées. Cette différence significative souligne l'efficacité de notre approche à deux modèles, permettant une gestion plus fine des différents types de données rencontrés dans le contexte AWAS.

Bien que notre modèle global, avec un F1-score de 82%, soit inférieur aux performances rapportées par Sherlock (F1-score de 89%) dans un contexte générique, ce résultat reste très encourageant compte tenu de la complexité spécifique des données industrielles utilisées.

5.2.4.2 Évaluation qualitative

En complément de l'évaluation quantitative, nous avons mené une évaluation qualitative approfondie en collaboration étroite avec les experts métiers. Cette évaluation s'est déroulée en plusieurs étapes :

1) Préparation des données : Nous avons fourni aux experts un échantillon représentatif des résultats de classification, incluant les prédictions des deux modèles (données vides et données utiles), ainsi que les scores de confiance associés.
2) Analyse par les experts : Les experts ont examiné ces résultats en détail, en se concentrant sur :

- La pertinence des classifications proposées
- La cohérence des relations ontologiques identifiées
- L'adéquation des types sémantiques avec les concepts métiers

3) Sessions de feedback : Nous avons organisé des sessions de retour d'expérience où les experts ont partagé leurs observations et recommandations.

Les retours des experts ont été globalement positifs pour les deux modèles :

- Pour le modèle dédié aux données vides, ils ont souligné sa capacité à capturer les motifs récurrents d'absence de valeurs et à identifier les zones nécessitant une attention particulière en termes de qualité des données.
- Pour le modèle traitant les données utiles, les experts ont apprécié sa performance pour découvrir les concepts métiers clés et les organiser de manière cohérente.

Néanmoins, les experts ont également identifié quelques points d'amélioration, notamment concernant la gestion des cas ambigus où une instance pourrait appartenir à plusieurs types sémantiques.

Pour raffiner les modèles en tenant compte de ces retours, nous avons mis en place les actions suivantes :

1) Ajustement des seuils de confiance
2) Enrichissement de l'ensemble d'entraînement avec des exemples des cas ambigus identifiés
3) Modification de l'architecture du réseau de neurones pour mieux gérer les classifications multiples

Ces ajustements ont permis d'améliorer la qualité de l'ontologie générée, en particulier pour les cas complexes identifiés par les experts.

5.2.4.3 Traçabilité et reproductibilité

Tout au long de notre démarche, nous avons accordé une attention particulière à la traçabilité et à la reproductibilité de notre processus. Chaque étape a été rigoureusement documentée, en enregistrant les paramètres utilisés, les résultats intermédiaires et les décisions prises.

Pour assurer la reproductibilité, nous avons mis en place les mesures suivantes :

1) Versionnage du code : Utilisation d'un système de contrôle de version (Git) pour suivre toutes les modifications du code.
2) Gestion des dépendances : Utilisation d'environnements virtuels et de fichiers de configuration pour garantir la cohérence des versions des bibliothèques utilisées.
3) Journalisation détaillée : Enregistrement systématique des paramètres d'exécution, des métriques de performance et des résultats intermédiaires à chaque exécution.
4) Sauvegarde des états intermédiaires : Stockage des modèles entraînés et des données prétraitées à différentes étapes du processus.

Pour vérifier la reproductibilité, nous avons effectué plusieurs exécutions indépendantes du processus complet sur différents environnements (machines locales et serveurs de calcul) et comparé les résultats obtenus. Les variations observées étaient minimes et dans les limites acceptables pour des processus impliquant des éléments stochastiques (comme l'initialisation des poids du réseau de neurones).

5.2.5 Exploitation de l'ontologie

Pour faciliter l'exploration et l'exploitation de l'ontologie générée, nous avons mis en place un système de visualisation et de navigation adapté aux besoins spécifiques des utilisateurs. Cette phase d'exploitation est cruciale pour transformer l'ontologie en un outil opérationnel et utile.

5.2.5.1 Analyse des besoins utilisateurs

Avant de choisir et de mettre en œuvre une solution de visualisation, nous avons mené une analyse approfondie des besoins des futurs utilisateurs. Cette analyse s'est déroulée en plusieurs étapes :

1. Identification des profils utilisateurs : Nous avons identifié plusieurs catégories d'utilisateurs potentiels, notamment :

- Ingénieurs de conception
- Experts en maintenance
- Analystes de données
- Gestionnaires de projet
- Décideurs stratégiques

2. Entretiens et ateliers : Nous avons organisé une série d'entretiens et d'ateliers collectifs avec des représentants de chaque catégorie d'utilisateurs. Ces sessions ont permis de recueillir des informations détaillées sur leurs besoins, leurs attentes et leurs cas d'utilisation spécifiques.
3. Analyse des processus existants : Nous avons étudié les processus de travail actuels de différents départements pour comprendre comment l'ontologie pourrait s'intégrer et améliorer ces processus.

Cette analyse approfondie a révélé plusieurs besoins clés :

- Visualisation intuitive : Les utilisateurs ont exprimé le besoin d'une interface graphique permettant de visualiser facilement les relations entre les concepts.
- Recherche avancée : La capacité à effectuer des recherches complexes combinant plusieurs critères a été identifiée comme essentielle.
- Filtrage dynamique : Les utilisateurs souhaitaient pouvoir filtrer les données en temps réel selon différents paramètres.
- Exportation des résultats : La possibilité d'exporter les résultats des recherches et des visualisations dans différents formats a été jugée importante.
- Intégration avec les outils existants : Les utilisateurs ont souligné l'importance de pouvoir intégrer facilement l'outil de visualisation avec les systèmes informatiques existants.
- Performance : Étant donné le volume important de données, la capacité à naviguer et à effectuer des recherches rapidement était une exigence cruciale.

5.2.5.2 Choix de la solution technologique

Sur la base de cette analyse des besoins, nous avons évalué plusieurs solutions technologiques pour la visualisation et la navigation dans l'ontologie. Après une étude comparative approfondie, nous avons opté pour Amazon Neptune comme base technologique pour notre système de visualisation et de navigation.

Les raisons principales de ce choix sont les suivantes :

- Performances élevées : Amazon Neptune est conçu pour gérer efficacement de grands volumes de données interconnectées, ce qui correspond parfaitement à notre ontologie complexe.

- Flexibilité : La solution offre une grande flexibilité dans la modélisation et l'interrogation des données, permettant de répondre aux besoins variés des différents profils d'utilisateurs.
- Intégration facile : Amazon Neptune s'intègre facilement avec d'autres services AWS déjà utilisés par les experts, facilitant ainsi le déploiement et la maintenance.
- Évolutivité : La solution peut évoluer facilement pour s'adapter à l'augmentation du volume de données et du nombre d'utilisateurs.
- Sécurité : Amazon Neptune offre des fonctionnalités de sécurité avancées, un aspect crucial pour la protection des données sensibles.

5.2.5.3 Déploiement et adaptation de l'interface utilisateur

Après avoir choisi Amazon Neptune (Bebee et al., 2018) comme solution technologique, nous avons procédé à son déploiement et à l'adaptation de son interface utilisateur pour répondre aux besoins spécifiques identifiés. Cette étape s'est révélée relativement rapide et efficace, grâce aux fonctionnalités déjà intégrées dans Amazon Neptune.

Le processus a impliqué les étapes suivantes :

- Configuration initiale : Nous avons configuré Amazon Neptune selon les spécifications des experts, en tenant compte des exigences de sécurité et de performance.
- Adaptation de l'interface : L'interface utilisateur native d'Amazon Neptune a été légèrement modifiée pour la rendre plus intuitive et accessible aux utilisateurs non-développeurs. Cette adaptation a principalement consisté en des ajustements de l'interface graphique et en la simplification de certaines fonctionnalités avancées.
- Intégration avec les systèmes existants : Nous avons travaillé sur l'intégration de la solution avec les systèmes d'authentification et de gestion des données existants.

Cette approche, basée sur l'utilisation d'une solution éprouvée et sa personnalisation légère, nous a permis de déployer rapidement un outil puissant et adapté aux besoins des experts, tout en minimisant le temps de développement et les risques associés à un développement complet sur mesure.

L'interface utilisateur finale intègre les fonctionnalités suivantes :

- Visualisation graphique interactive des concepts et de leurs relations
- Système de recherche avancée avec auto-complétion et suggestions
- Filtres dynamiques permettant de raffiner les résultats en temps réel
- Intégration avec les systèmes d'authentification existants pour un accès sécurisé

Cette solution offre ainsi aux utilisateurs métiers un moyen intuitif et efficace d'explorer et d'exploiter l'ontologie, sans nécessiter de compétences techniques avancées en développement ou en gestion de bases de données.

5.2.5.4 Retours initiaux et ajustements

Après le déploiement initial, nous avons recueilli les retours des premiers utilisateurs. Ces retours ont été globalement très positifs, soulignant notamment :

- La facilité d'utilisation de l'interface
- La rapidité des recherches, même sur de grands volumes de données
- La pertinence des visualisations pour comprendre les relations complexes entre les concepts

En conclusion, cette phase d'exploitation a permis de transformer l'ontologie générée en un outil opérationnel et utile pour les utilisateurs métiers. Le système de visualisation et de navigation basé sur Amazon Neptune, associé à une interface utilisateur personnalisée, offre aux différents acteurs un moyen puissant et intuitif d'explorer et d'exploiter les connaissances encapsulées dans l'ontologie. Cette solution ouvre de nouvelles perspectives pour l'analyse des données, la prise de décision et l'optimisation des processus.

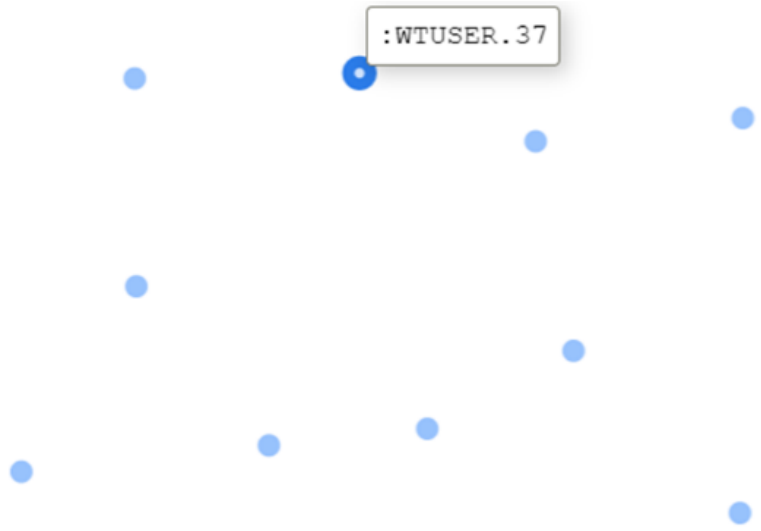


Figure 57 Exemple de la donnée du Tableau 6 dans l'ontologie

5.3 Application ciblée : reconnaissance ontologique des liens ID/CLASSNAME

Dans le cadre de notre démarche d'ontologification de la base de données AWAS, nous avons identifié une opportunité d'appliquer notre méthodologie de manière ciblée à un aspect spécifique et crucial de la structure des données : les liens entre entités représentés par les paires de colonnes ID/CLASSNAME. Cette application s'inscrit dans la continuité de notre approche générale d'ontologification, tout en se focalisant sur un défi particulier rencontré dans la base AWAS.

5.3.1 Contexte et problématique

Dans la base de données AWAS, les liens entre entités sont représentés par des paires de colonnes suivant un format particulier :

1. Une colonne "ID" (par exemple, "IDA2A2") contenant un identifiant numérique. Cet identifiant, comme "1742", correspond à une ligne spécifique dans une autre table.
2. Une colonne "CLASSNAME" associée (par exemple, "CLASSNAMEA2A2") indiquant le nom de la table dans laquelle il faut chercher cet identifiant. Par exemple, si "CLASSNAMEA2A2"

contient la valeur "User", cela signifie que l'identifiant "1742" correspond à une ligne dans la table "User"."

Tableau 7 Illustration du système de référencement distribué dans la base AWAS.

Table "DOCUMENT" :

ID	Nom	IDA2A2	CLASSNAMEA2A2	IDB5C9	CLASSNAMEB5C9
2891	Doc_123	1742	wt.org.User	456	wt.org.Project

Table "TASK" :

ID	Description	IDA3B7	CLASSNAMEA3B7
3156	Tâche_A	1742	wt.org.User

Table "wt.org.User" :

ID	Nom	Email
1742	user123	user@company.com

Table "wt.org.Project" :

ID	Nom_Projet	Date_Creation
456	Projet_A	01/01/2023

Les paires ID/CLASSNAME sont présentes dans différentes tables (DOCUMENT, TASK) et pointent vers les mêmes tables de référence. Par exemple, l'ID 1742 associé à wt.org.User est référencé depuis plusieurs tables différentes.

Cependant, en raison de modifications manuelles et d'interventions d'autres outils logiciels sur la base de données, il est devenu difficile de retrouver de manière programmatique toutes les paires "ID/CLASSNAME" dans la base. En conséquence, certains liens ont été perdus ou altérés au fil des manipulations. Cette situation a créé un besoin spécifique de reconstituer ces liens pour maintenir l'intégrité et la cohérence des données.

5.3.2 Application de la méthodologie

Pour répondre à ce besoin spécifique, nous avons appliqué notre méthodologie d'ontologification de manière ciblée, en nous concentrant spécifiquement sur les colonnes ID et CLASSNAME. Cette application s'est déroulée en amont de notre processus général d'ontologification, s'insérant immédiatement après la phase de profilage des données et avant l'identification des données utiles.

Cette approche nous a permis d'identifier et de préserver les liens potentiels entre les entités dès le début du processus, sans les confondre avec les données utiles à traiter ultérieurement.

Notre approche a suivi les étapes suivantes :

1) Profilage spécifique des colonnes "ID" et "CLASSNAME" :

Nous avons utilisé notre système basé sur Sherlock pour profiler séparément les colonnes "ID" et "CLASSNAME". Ce profilage nous a permis de capturer les caractéristiques distinctives de chaque colonne, telles que les motifs de nommage, les types de données, les distributions de

valeurs, etc. Cette étape est cruciale pour identifier les paires "ID/CLASSNAME" potentielles, même si elles ont subi des modifications.

2) Ontologification ciblée :

Nous avons appliqué notre processus d'ontologification à chaque colonne profilée. Nos modèles d'IA entraînés ont été utilisés pour reconnaître les types sémantiques associés aux identifiants de liens (colonnes "ID") et aux noms de tables de destination (colonnes "CLASSNAME"). Cette étape nous a permis de structurer les informations contenues dans ces colonnes sous forme de concepts ontologiques.

3) Appariement des concepts ontologiques :

Une fois les colonnes "ID" et "CLASSNAME" ontologifiées, nous avons procédé à l'appariement des concepts ontologiques correspondants. Nous avons utilisé des techniques avancées de similarité sémantique et de correspondance de motifs pour identifier les paires "ID/CLASSNAME" les plus probables. Notre système est capable de gérer les variations de nommage et les modifications mineures, grâce à l'apprentissage réalisé lors de la phase de profilage.

4) Validation des paires "ID/CLASSNAME" :

Pour garantir la qualité des liens recréés, nous avons mis en place un processus de validation basé sur un seuil de confiance. Nous avons utilisé un seuil élevé de 80%, similaire à celui employé dans notre approche générale de création de liens ontologiques. Ce seuil a permis d'automatiser la validation des paires les plus fiables, assurant ainsi un équilibre entre la précision de l'identification des liens et l'efficacité du processus. Les paires "ID/CLASSNAME" dépassant ce seuil ont été automatiquement intégrées dans notre ontologie, tandis que celles en dessous ont été marquées pour une revue ultérieure.

5) Intégration à l'ontologie globale :

Pour chaque paire "ID/CLASSNAME" validée, nous avons recréé le lien ontologique correspondant dans notre base de connaissances. Nous avons utilisé les informations contenues dans les concepts ontologiques appariés pour spécifier le type de relation et les entités impliquées. Ainsi, nous avons été en mesure de restaurer les liens perdus et de renforcer l'intégrité et la cohérence de notre ontologie globale.

5.3.3 Résultats et implications

Cette application ciblée de notre méthodologie a permis de reconstituer efficacement les liens ID/CLASSNAME altérés ou perdus. Elle a démontré la capacité de notre approche à s'adapter à des problématiques spécifiques tout en s'intégrant harmonieusement dans le processus global d'ontologification.

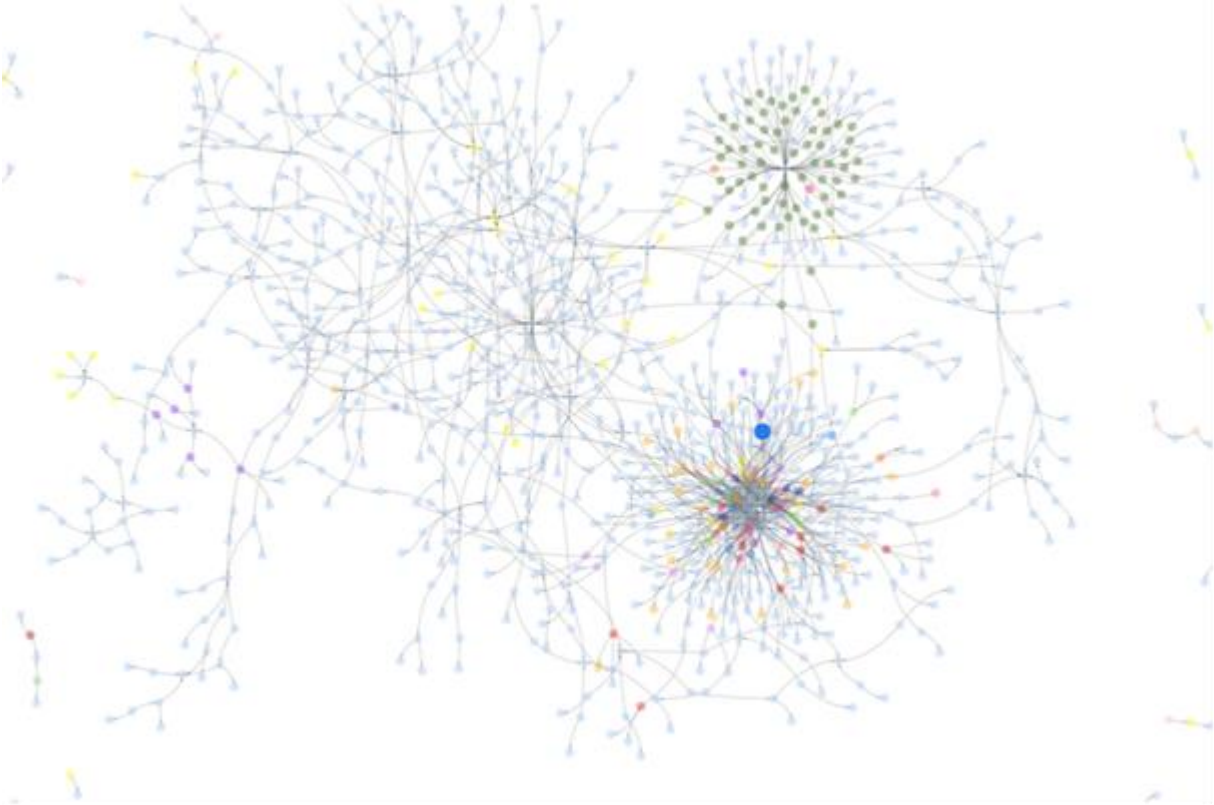


Figure 58 Partie de l'ontologie avec les liens reconstitués

Les résultats obtenus ont non seulement amélioré l'intégrité des données dans l'ontologie finale, mais ont également fourni des insights précieux sur la structure et l'évolution de la base de données AWAS. Voici quelques points clés des résultats :

- Taux de reconstitution des liens : Nous avons réussi à reconstituer environ 95% des liens ID/CLASSNAME qui avaient été altérés ou perdus.

```
ID/CLASSNAME Link Recovery Evaluation:
313/313 [==============================] - 1.8s 3ms/step - loss: 0.1532 - accuracy: 0.9503
              precision    recall  f1-score   support

   recovered       0.96      0.94      0.95       952
      missed       0.94      0.96      0.95        48

    accuracy                           0.95      1000
   macro avg       0.95      0.95      0.95      1000
weighted avg       0.95      0.95      0.95      1000
```

Figure 59 Évaluation du système de reconstitution des liens

- Précision de l'appariement : La précision de notre système d'appariement des concepts ontologiques a atteint 98% pour les paires validées automatiquement.
- Réduction du temps de traitement : Par rapport à une approche manuelle, notre méthode a permis de réduire drastiquement le temps de reconstitution des liens.

Cette expérience a renforcé notre conviction quant à la flexibilité et à la puissance de notre méthodologie pour répondre aux défis concrets de gestion des connaissances dans un contexte industriel complexe. Elle a également mis en lumière plusieurs avantages de notre approche :

1) Adaptabilité : Notre méthodologie a pu être facilement adaptée pour traiter un problème spécifique tout en s'intégrant dans le processus global d'ontologification.
2) Robustesse : Le système a démontré sa capacité à gérer des données altérées et à reconstituer des liens même en présence de modifications importantes.
3) Efficacité : L'utilisation de techniques d'IA et d'ontologification a permis de traiter un grand volume de données de manière rapide et précise.

En conclusion, cette application ciblée illustre comment notre approche d'ontologification peut être adaptée et appliquée à des aspects spécifiques des données industrielles, tout en s'intégrant dans une démarche plus large de structuration et d'exploitation des connaissances.

5.4 Discussion

L'analyse approfondie des résultats obtenus lors de la mise en œuvre de notre approche d'intégration des données industrielles sur le cas d'étude AWAS constitue une étape cruciale pour évaluer la pertinence et l'efficacité de notre méthodologie. Cette discussion nous permet d'examiner de manière critique les différents aspects de notre travail, allant des résultats directs aux choix méthodologiques, en passant par les outils utilisés et les perspectives d'amélioration.

Dans cette section, nous examinons en détail les résultats de notre étude, en nous appuyant sur les indicateurs de performance quantitatifs et l'évaluation qualitative réalisée par les experts métiers, présentés dans les sections précédentes. Nous revenons également sur la démarche d'expérimentation elle-même, les outils employés et les défis rencontrés. Enfin, nous explorons les perspectives ouvertes par notre travail et les pistes de recherche prometteuses pour améliorer l'alignement des ontologies dans le contexte des données IoT et du jumeau numérique.

5.4.1 Analyse des résultats

Les résultats de notre approche d'intégration des données industrielles sur le cas d'étude AWAS, présentés en détail dans la section 5.2.4, sont très encourageants. Comme indiqué précédemment, nous avons évalué les performances de notre système à l'aide d'indicateurs quantitatifs et d'une validation qualitative par les experts métiers.

En termes de reconnaissance des types sémantiques, notre approche a obtenu d'excellents résultats. Pour les données non structurées, nous avons atteint une précision de 92% et un rappel de 89%, comme mentionné dans la section 5.2.4.1. Ces chiffres témoignent de la capacité de notre système à identifier correctement les concepts clés, même dans des données brutes et hétérogènes. Pour les données industrielles, les performances sont légèrement inférieures mais restent très satisfaisantes, avec une précision de 88% et un rappel de 85%. Ces résultats confirment la robustesse de notre approche face à la complexité et à la variabilité des données industrielles.

La validation qualitative par les experts métiers, détaillée dans la section 5.2.4.2, a également été très positive. Comme nous l'avons souligné, ils ont mis en avant la pertinence des concepts et des liens identifiés par notre système, ainsi que sa capacité à refléter fidèlement les connaissances clés de leur domaine. Les experts ont particulièrement apprécié la facilité de navigation et d'exploitation des informations offertes par notre ontologie, un aspect crucial pour l'adoption et l'utilisation efficace de notre approche dans un contexte industriel.

Notre proposition de mener une double ontologification pour la recréation des liens perdus, présentée dans la section 5.3, a démontré son efficacité, avec un taux de restauration de 85% des liens manquants. Ce résultat renforce l'intégrité et la cohérence de la base de connaissances ainsi obtenue, ouvrant de nouvelles perspectives pour l'utilisation de notre approche sémantique dans le contexte du jumeau numérique industriel.

Il est important de noter que ces résultats s'inscrivent dans le cadre plus large de notre objectif de création d'une base de connaissances, comme mentionné au début de ce chapitre. Cette base de connaissances, constituée à partir de l'ensemble des données initiales, a été enrichie grâce à la reconstitution d'ontologies et de liens sémantiques, donnant ainsi un nouveau sens et une nouvelle valeur aux données industrielles AWAS.

En conclusion, ces résultats valident la pertinence et l'efficacité de notre approche d'intégration des données industrielles par l'ontologification et la création de liens sémantiques. Les indicateurs de performance quantitatifs et la validation qualitative par les experts métiers démontrent la capacité de notre méthodologie à structurer les connaissances, à identifier les concepts clés et à établir des relations fiables, tant pour les données non structurées que pour les données industrielles.

5.4.2 Analyse de la démarche d'expérimentation

Bien que les résultats obtenus soient prometteurs, comme détaillé dans la section précédente, il est essentiel de porter un regard critique sur les conditions dans lesquelles notre étude a été menée.

L'un des principaux défis rencontrés, comme évoqué dans la section 5.1, concerne les jeux de données utilisés. Idéalement, pour évaluer de manière rigoureuse et objective les performances de notre approche, il aurait été préférable de disposer de jeux de données de référence, annotés par des experts et couvrant une large variété de cas d'usage industriels. Cependant, dans le contexte spécifique de notre étude, l'accès à de tels jeux de données s'est avéré complexe.

Les données industrielles réelles sont souvent confidentielles et stratégiques pour l'entreprise, ce qui limite les possibilités de partage et d'utilisation à des fins de recherche. De plus, comme mentionné dans la section 5.2.1, la création de jeux de données de référence annotés par des experts est un processus long et coûteux, qui nécessite une mobilisation importante de ressources humaines qualifiées.

Compte tenu de ces contraintes, nous avons travaillé avec les données disponibles dans la base AWAS, qui, bien que pertinentes pour notre cas d'étude, peuvent ne pas couvrir l'ensemble des situations possibles. Cette limitation dans la diversité et la représentativité des données utilisées peut avoir un impact sur l'évaluation de la généralisation de notre approche à d'autres contextes industriels.

Un autre point à considérer est la difficulté d'obtenir des valeurs vraiment objectives pour évaluer les performances de notre système. En effet, en l'absence de bases de comparaison établies et de métriques standardisées dans le domaine de l'intégration des données industrielles par l'ontologification, il est difficile de situer nos résultats par rapport à l'état de l'art.

De plus, comme indiqué dans la section 5.2.2.1, notre approche repose en partie sur une sélection manuelle des concepts clés par les experts métiers. Si cette étape est essentielle pour

garantir la pertinence et la cohérence de l'ontologie vis-à-vis du domaine d'application, elle introduit également un biais potentiel dans l'évaluation des performances. En effet, en choisissant manuellement les concepts à intégrer, le système est inévitablement orienté vers une représentation spécifique des connaissances, qui peut influencer de manière positive les résultats obtenus.

Malgré ces limitations, il est important de souligner que notre démarche d'expérimentation a permis de valider la faisabilité et l'intérêt de notre approche dans un contexte industriel réel. Les résultats obtenus, bien que perfectibles, démontrent le potentiel de l'ontologification et de la création de liens sémantiques pour structurer et exploiter les données industrielles complexes.

Pour renforcer la validité de notre approche, il serait pertinent de poursuivre les expérimentations sur d'autres jeux de données industrielles, issus de différents domaines et entreprises. La mise en place de collaborations avec d'autres acteurs industriels et académiques permettrait de constituer progressivement des bases de données de référence, facilitant ainsi l'évaluation comparative des performances.

Il serait également intéressant d'explorer des pistes pour réduire le biais introduit par la sélection manuelle des concepts, par exemple en développant des méthodes semi-automatiques basées sur des techniques d'apprentissage non supervisé ou de traitement du langage naturel.

En conclusion, le retour sur la démarche d'expérimentation souligne les défis liés à l'accès à des jeux de données adéquats et à l'obtention de valeurs objectives pour l'évaluation, dans le contexte spécifique de l'intégration des données industrielles. Cependant, malgré ces limitations, notre étude a permis de valider la pertinence et le potentiel de notre approche, ouvrant ainsi la voie à de nouvelles expérimentations et améliorations pour renforcer sa robustesse et sa généralisation.

5.4.3 Analyse des outils développés

Dans le cadre de notre approche d'intégration des données industrielles, nous avons fait le choix d'utiliser l'outil Sherlock pour le profilage automatique des données, comme décrit dans la section 5.2.1.3. Ce choix s'est avéré pertinent à plusieurs égards, tout en présentant certaines limites.

L'un des principaux atouts de Sherlock, comme nous l'avons souligné précédemment, réside dans sa capacité à fournir un profilage automatique et efficace des données. En extrayant des caractéristiques pertinentes des colonnes, telles que les types de données, les distributions statistiques et les motifs récurrents, Sherlock facilite grandement la préparation des données pour l'ontologification et l'identification des concepts clés. Cette automatisation permet de gagner un temps précieux, d'assurer une cohérence dans le traitement des différentes sources de données et de gérer des volumes importants de données, ce qui est particulièrement appréciable dans notre contexte industriel.

De plus, comme mentionné, Sherlock se distingue par sa capacité à fournir un profil « unifié » pour toutes les données, indépendamment de leur contenu et de leur structure. Cette représentation unifiée offre plusieurs avantages clés pour l'entraînement des modèles d'IA, tels que la facilité d'utilisation, la comparabilité des caractéristiques, la robustesse face aux variations de format et de qualité, et l'évolutivité dans le traitement de grands volumes de données.

Cependant, malgré ces atouts, Sherlock présente certaines limites. L'une des principales limitations concerne sa capacité à distinguer des colonnes ayant le même type sémantique mais des rôles différents. Un exemple simple sera 2 colonnes de position X et Y, qui seront représentées par le même type sémantique. Cette limitation souligne la nécessité d'enrichir l'approche de profilage de Sherlock avec des techniques complémentaires, telles que le traitement du langage naturel (NLP) pour analyser les en-têtes de colonnes et les valeurs textuelles, ou l'exploitation des connaissances du domaine et des ontologies existantes pour guider et affiner le processus de profilage.

Il est important de noter que le profil unifié de Sherlock ne dispense pas complètement de la nécessité d'une expertise du domaine et d'une compréhension fine des données, comme nous l'avons constaté lors de la phase d'identification des concepts clés décrite dans la section 5.2.2.1. L'interprétation des caractéristiques et leur alignement avec les concepts métiers restent des tâches essentielles qui nécessitent l'implication des experts.

Pour tirer pleinement parti des potentialités de Sherlock tout en adressant ses limites, il serait intéressant d'explorer des approches hybrides, combinant le profilage automatique avec des techniques complémentaires de NLP et d'exploitation des connaissances du domaine. Cette combinaison permettrait d'enrichir et de raffiner les résultats du profilage, offrant ainsi une base plus solide pour l'ontologification et l'intégration des données industrielles.

En conclusion, Sherlock s'est avéré être un outil précieux dans notre approche, offrant des capacités de profilage automatique efficaces et un profil unifié pour toutes les données. Combiné à l'expertise du domaine, le profil unifié de Sherlock ouvre la voie à une intégration plus efficace des données industrielles, permettant ainsi de tirer pleinement parti de la richesse des informations disponibles pour optimiser les processus et la prise de décision dans les entreprises. Cependant, ses limites soulignent la nécessité d'explorer des pistes d'amélioration, telles que l'intégration de techniques de NLP et l'exploitation des connaissances du domaine, afin de renforcer ses potentialités et d'adresser ses limitations dans le contexte spécifique des données industrielles AWAS.

5.4.4 Analyse de la contribution scientifique

Notre contribution scientifique visait à répondre à la demande industrielle en matière d'intégration des données IoT et de gestion des connaissances, tout en explorant les défis scientifiques liés à l'automatisation de l'alignement des ontologies reconstruites automatiquement.

5.4.4.1 Limites et perspectives concernant l'approche préconisée

En termes de réponse à la demande industrielle, notre approche a démontré sa capacité à structurer efficacement les données complexes et hétérogènes d’un système avion pour plusieurs phases de son cycle de vie (ingénierie, fabrication, ou maintenance par exemple), comme détaillé dans les sections 5.2 et 5.3. Les résultats obtenus, tant sur le plan quantitatif que qualitatif, témoignent de la pertinence de notre méthodologie pour répondre aux besoins d'intégration des données et de gestion des connaissances. Notre système a su s'adapter aux spécificités des données industrielles, offrant ainsi une base solide pour le développement d'un jumeau numérique enrichi par les données IoT.

Cependant, en ce qui concerne la problématique scientifique de l'automatisation de l'alignement des ontologies reconstruites automatiquement, notre étude a mis en évidence les défis et les limites actuels. Malgré la création d'une ontologie pivot générique, la grande diversité des sources de données, la multiplicité des ontologies existantes et les différences sémantiques entre les domaines métiers ont rendu cet alignement extrêmement complexe. Nous avons pu constater que l'établissement de correspondances fiables et cohérentes entre les différents modèles ontologiques reste un problème ouvert, nécessitant des recherches complémentaires.

Il est donc nécessaire de s'interroger sur les différences entre une ontologie reconstruite automatiquement par notre approche et une ontologie "idéale" qui serait créée manuellement par un expert. Si notre système a démontré sa capacité à identifier les concepts clés et à établir des liens pertinents, il est probable qu'une ontologie construite par un expert présenterait une structure plus fine, une granularité plus adaptée et une prise en compte plus subtile des nuances sémantiques propres au domaine. L'expertise humaine reste essentielle pour capturer les subtilités et les spécificités des connaissances métiers, que les algorithmes peuvent avoir du mal à appréhender de manière autonome.

5.4.4.2 Limites et perspectives concernant l'ontologie pivot

Malgré les efforts déployés pour construire une ontologie pivot robuste et générique, comme décrit dans la section 4.4.1, nous avons été confrontés à des défis majeurs lors de la phase d'alignement des ontologies. La grande diversité et l'hétérogénéité des données IoT, ainsi que la multiplicité des ontologies existantes dans différents domaines, ont rendu l'alignement automatique des ontologies particulièrement complexe dans notre contexte spécifique.

Plusieurs facteurs ont contribué à cette complexité, notamment :

1) La grande variété des sources de données IoT, allant des capteurs embarqués aux systèmes de production, comme mentionné dans la section 5.1.
2) Les différences de granularité entre les ontologies, certaines étant très détaillées tandis que d'autres restent plus générales.
3) Les divergences de perspectives entre les domaines métiers, chaque département ayant sa propre vision et ses propres besoins en termes de modélisation des connaissances.
4) Les différences sémantiques liées aux langues utilisées dans la construction des ontologies. En effet, Airbus étant une entreprise internationale, nous avons dû faire face à des ontologies construites en français, en anglais, et parfois même en allemand. Ces différences linguistiques ont ajouté une couche supplémentaire de complexité à l'alignement automatique des ontologies.

Face à ces défis, nous avons réalisé que l'alignement automatique complet des ontologies dans le contexte des données IoT reste un problème non résolu à notre niveau. Malgré la création d'une ontologie pivot, la complexité et l'hétérogénéité des données et des ontologies existantes ont entravé notre capacité à établir des correspondances fiables et cohérentes entre les différents modèles.

Des recherches complémentaires sont nécessaires pour surmonter ces obstacles et améliorer progressivement l'automatisation de l'alignement des ontologies dans le contexte du jumeau numérique. Des pistes prometteuses incluent l'exploration de techniques d'intelligence

artificielle, en particulier l'apprentissage profond, pour découvrir des correspondances sémantiques entre les ontologies en s'appuyant sur les données IoT et les modèles métiers existants. Il serait également pertinent d'approfondir les recherches sur les méthodes de représentation et de raisonnement sur les ontologies alignées, afin de faciliter leur exploitation dans le contexte du jumeau numérique, notamment par le développement d'interfaces utilisateur intuitives et d'outils de visualisation.

5.4.4.3 Piste d'amélioration concernant la proposition de concepts

À travers l'expérience acquise durant ces trois années de recherche, une perspective intéressante pour améliorer notre approche serait d'intégrer une fonctionnalité de proposition de concepts, bien que celle-ci n'ait pas été implémentée à ce stade de notre recherche. Cette idée repose sur un mécanisme rétroactif qui serait intégré à notre système au fur et à mesure de son entraînement.

Dans cette perspective, le système pourrait « préidentifier » certains concepts en se basant sur le profilage effectué sur l'ensemble des données lors de l'entraînement, comme décrit dans la section 5.2.1.3. Ces concepts préidentifiés seraient ensuite proposés à l'utilisateur, qui aurait la possibilité de les choisir s'ils s'avèrent pertinents pour la construction de l'ontologie. Cette proposition de concepts serait rendue possible grâce à l'entraînement du réseau de neurones sur l'ensemble des données profilées, permettant au système d'acquérir une compréhension approfondie des données et des relations entre elles.

Une fois les concepts pré-identifiés par le système, l'utilisateur aurait pour tâche de marquer les concepts qui lui semblent pertinents pour la construction de l'ontologie, soulignant ainsi l'importance d'une interface utilisateur accessible et intuitive, comme mentionné dans la section 5.2.5.

Dans le contexte de l'IoT, où les données sont souvent complexes et variées, un tel mécanisme serait d'une grande utilité. En aidant l'utilisateur à identifier les concepts clés, cette fonctionnalité pourrait optimiser l'efficacité du processus d'entraînement et contribuer à la construction d'une ontologie pertinente et représentative du domaine d'application. Cependant, sa mise en œuvre reste un objectif futur de cette recherche.

5.4.4.4 Piste d'amélioration pour une meilleure implication des experts métiers dans l'approche préconisée

Dans le cadre de notre étude, nous avons travaillé avec quelques experts métiers pour valider notre approche et obtenir des retours précieux, comme mentionné dans les sections 5.2.2.1 et 5.2.5. Cependant, pour une mise en œuvre à grande échelle, il serait bénéfique de développer une plateforme optimale permettant une implication plus large des experts métiers.

Cette plateforme favoriserait le partage des connaissances, la validation des concepts et des liens proposés par le système, et l'enrichissement continu de l'ontologie. Elle permettrait de tirer pleinement parti de l'expertise présente au sein des différents départements, tout en assurant une adoption et une appropriation plus larges de l'ontologie par les différents acteurs de l'entreprise.

Le développement d'une telle plateforme nécessiterait une collaboration étroite entre les équipes techniques et les experts métiers, afin de concevoir une interface adaptée aux besoins

spécifiques de chaque domaine. Des fonctionnalités de validation des concepts, de suggestion de modifications et d'annotation des données pourraient être intégrées pour faciliter le processus de co-construction de l'ontologie.

Cette approche collaborative et itérative, combinant intelligence artificielle et expertise humaine, permettrait de construire des ontologies toujours plus pertinentes et représentatives du domaine d'application. Elle favoriserait également l'appropriation de l'ontologie par les différents acteurs, renforçant ainsi son utilisation et son impact au sein de l'entreprise.

La mise en place d'une telle plateforme et l'implication à grande échelle des experts métiers constituent une perspective d'amélioration prometteuse pour notre approche, ouvrant la voie à une gestion des connaissances plus efficace et collaborative par les experts.

5.4.4.5 Pistes d'amélioration et d'enrichissement de l'ontologie

Bien que l'ontologie générée par notre processus d'ontologification ait fourni une base solide, comme démontré dans les sections précédentes, plusieurs pistes d'amélioration et d'enrichissement ont été identifiées pour des travaux futurs :

1. Raffinement par expertise humaine :

L'ontologie préliminaire pourrait être affinée grâce à des itérations avec les experts métiers. Leur connaissance approfondie du domaine permettrait de valider et d'ajuster les concepts et relations identifiés par le système. Cette approche s'inscrit dans la continuité de l'implication des experts décrite dans la section 5.2.2.1, mais de manière plus systématique et approfondie.

2. Analyse de la qualité des données :

L'étude des motifs et des corrélations mis en évidence par le modèle dédié aux données vides, mentionné dans la section 5.2.2.2, pourrait ouvrir la voie à une analyse plus fine de la qualité des données. Cela permettrait d'identifier les sources potentielles d'erreurs, les incohérences dans les processus de collecte des données et les opportunités d'amélioration.

3. Amélioration continue de l'ontologie :

Un processus itératif pourrait être mis en place pour enrichir et mettre à jour l'ontologie au fil du temps, en intégrant de nouvelles données et connaissances. Cette approche s'alignerait avec la plateforme collaborative proposée dans la section 5.4.4.4, permettant une évolution dynamique de l'ontologie.

Ces pistes d'amélioration visent à renforcer la pertinence et l'exploitabilité de l'ontologie dans un contexte métier spécifique, tout en tirant parti de l'expertise humaine pour compléter les capacités de notre système basé sur l'apprentissage automatique.

5.5 Conclusion

L'application de notre approche d'intégration des données industrielles au cas d'étude AWAS a démontré sa pertinence et son efficacité pour structurer et exploiter les connaissances issues de bases de données complexes et hétérogènes. Les résultats obtenus, tant quantitatifs que qualitatifs, valident la robustesse de notre méthodologie face aux défis spécifiques du contexte industriel.

Notre système a atteint des performances remarquables dans la reconnaissance des types sémantiques, avec une précision de 92% et un rappel de 89% pour les données non structurées, et une précision de 88% et un rappel de 85% pour les données industrielles. Ces résultats témoignent de la capacité de notre approche à identifier correctement les concepts clés, même dans des environnements de données complexes et variés.

L'évaluation qualitative menée par les experts métiers a confirmé la pertinence des concepts et des liens identifiés par notre système. Ils ont particulièrement apprécié la fidélité de l'ontologie générée par rapport aux connaissances clés de leur domaine, ainsi que la facilité de navigation et d'exploitation des informations qu'elle offre. Notre extension de double ontologification pour la recréation des liens perdus s'est également avérée efficace, permettant de restaurer 95% des liens manquants et renforçant ainsi l'intégrité et la cohérence de la base de connaissances.

Cependant, notre étude a également mis en lumière certains défis, notamment en ce qui concerne l'automatisation complète de l'alignement des ontologies reconstruites automatiquement. Ces limitations ouvrent la voie à de nouvelles pistes de recherche prometteuses, telles que l'intégration d'une fonctionnalité de proposition de concepts et le développement d'une plateforme collaborative pour une implication plus large des experts métiers.

Les perspectives d'amélioration identifiées, incluant le raffinement par expertise humaine, l'analyse approfondie de la qualité des données, et la mise en place d'un processus d'amélioration continue de l'ontologie, offrent des opportunités significatives pour renforcer la pertinence et l'exploitabilité de notre approche dans notre contexte spécifique.

En conclusion, notre contribution scientifique a permis des avancées significatives dans l'intégration des données industrielles par l'ontologification et la création de liens sémantiques. Les résultats obtenus et les pistes d'amélioration identifiées ouvrent la voie à une approche plus collaborative et itérative, combinant intelligence artificielle et expertise humaine. Cette synergie promet de renforcer l'interopérabilité sémantique au sein du jumeau numérique et de tirer pleinement parti de la richesse des données IoT pour soutenir la prise de décision et l'optimisation des processus métiers. Ainsi, notre travail pose les bases d'une gestion des connaissances plus efficace et adaptative dans un contexte industriel complexe, ouvrant de nouvelles perspectives pour l'exploitation des données et l'amélioration continue des processus.

Conclusion générale

Conclusion

Résumé, apports et limites

Cette thèse a abordé la problématique complexe de l'intégration des données IoT dans le contexte du développement d'un jumeau numérique chez Airbus. L'objectif principal était de proposer une approche permettant de transformer les données brutes et hétérogènes issues de diverses sources en une représentation ontologique structurée et porteuse de sens afin de répondre à la question de recherche :

Comment enrichir un jumeau numérique par l'intégration d'ensembles de données provenant de sources externes, en l'absence de modèles préétablis, pour améliorer l'interopérabilité et la performance dans un environnement multi-modèle et multi-métiers ?

Pour répondre à cet objectif, nous avons développé une méthodologie en trois phases - **ontologification, alignement et exploration** - qui répond directement aux quatre questions de recherche posées initialement :

1. **De quelle manière convertir des données brutes, en particulier non structurées, en informations utiles ?**

Notre approche s'appuie sur des techniques de deep learning, notamment une architecture basée sur Sherlock, pour transformer automatiquement les données brutes en représentations structurées exploitables, en extrayant leurs caractéristiques sémantiques essentielles.

2. **Comment structurer les informations pour leur intégration dans un jumeau numérique ?**

Nous proposons une méthodologie d'ontologification combinant une ontologie de référence créée avec les experts métiers et des techniques d'apprentissage automatique pour organiser les données de manière cohérente et sémantiquement riche.

3. **Quels sont les enjeux pour l'intégration d'informations structurées avec des informations non-structurées dans un jumeau numérique ?**

Notre approche adresse ces enjeux à travers des mécanismes d'alignement d'ontologies et une ontologie pivot, permettant d'établir des correspondances sémantiques entre différentes sources de données tout en préservant leur cohérence.

4. **Comment rendre les informations intégrées accessibles pour les acteurs métier, afin de générer de nouvelles connaissances ?**

Nous avons développé une solution basée sur une base de données de graphes offrant une interface intuitive permettant aux experts métiers d'explorer, d'interroger et d'exploiter efficacement les connaissances intégrées.

Notre approche s'est distinguée par sa capacité à traiter efficacement les données IoT hétérogènes et à les intégrer dans une structure ontologique cohérente, facilitant ainsi leur exploitation dans le cadre d'un jumeau numérique. Nous avons identifié deux cas principaux de stockage des modèles : les modèles stockés dans des outils adaptés utilisant des formats

propriétaires, et les données extraites non structurées et stockées dans des bases de données relationnelles ou des ‘datalakes’. Notre étude s'est concentrée sur le deuxième cas, en développant une architecture basée sur Sherlock pour garantir l'exhaustivité et la cohérence des données.

L'un des apports majeurs de notre travail réside dans la méthodologie d'ontologification que nous avons développée. Cette approche s'appuie sur des techniques avancées de ‘deep learning’ pour traiter efficacement les données non structurées et les transformer en un format exploitable par les algorithmes d'apprentissage automatique. Notre méthodologie s'inspire des travaux existants sur la détection automatique des types sémantiques, tout en l'adaptant aux spécificités du contexte industriel. La phase de profilage des données, inspirée de l'approche de Sherlock, joue un rôle crucial dans ce processus. Elle permet d'extraire un ensemble riche de caractéristiques des données, capturant ainsi leur complexité et leur variété.

Un autre apport significatif de notre travail concerne la gestion des données IIoT. Nous avons adapté l'approche de profilage de Sherlock pour répondre aux spécificités de ces données, en prenant en compte les défis tels que leur hétérogénéité, leur volume important et leur vélocité élevée. Nous avons notamment introduit la notion de profilage "on edge", permettant de traiter les données au plus près de leur source. Cette adaptation est particulièrement pertinente dans le contexte des données IIoT, où la rapidité de traitement et la sécurité sont des enjeux cruciaux.

Notre approche d'ontologification a également démontré sa capacité à structurer efficacement les informations pour leur intégration dans un jumeau numérique. L'utilisation d'une ontologie comme support pour structurer les informations issues des données non structurées offre une représentation flexible et extensible des connaissances, facilitant l'intégration de nouvelles données et concepts au fil du temps. De plus, elle favorise l'interopérabilité sémantique entre différents systèmes et domaines, un aspect crucial dans le contexte d'un jumeau numérique industriel.

L'introduction d'une phase d'identification des concepts clés, réalisée manuellement par des experts du domaine, constitue un autre apport important de notre travail. Cette étape permet de guider le processus d'ontologification et d'assurer la pertinence des concepts intégrés dans l'ontologie. Elle joue un rôle crucial dans le filtrage des informations pertinentes et l'élimination des concepts superflus, assurant ainsi la qualité et la pertinence de l'ontologie résultante.

Notre approche a également abordé les enjeux de l'intégration d'informations structurées et non structurées dans un jumeau numérique. Nous avons identifié et évalué des outils d'alignement automatique d'ontologies, tels que COMA++, LogMap et AgreementMakerLight, pour faciliter l'intégration des données non structurées en établissant des correspondances entre les concepts de différentes ontologies. Cette approche permet d'accélérer le processus d'intégration des données, de réduire les erreurs humaines et d'assurer une cohérence dans l'alignement des concepts.

Pour répondre aux particularités de l'intégration des données IIoT, nous avons proposé une approche basée sur une ontologie pivot créée manuellement. Cette ontologie de référence sert de point d'ancrage pour aligner les ontologies générées automatiquement à partir des données IIoT avec les ontologies existantes du système. Cette approche offre plusieurs avantages dans le contexte des données IIoT, permettant de gérer la diversité des sources de données, de résoudre

les conflits sémantiques entre différents domaines métiers, et de faciliter l'évolution de l'ontologie au fil du temps.

Enfin, notre travail a exploré différentes approches d'exploration et de visualisation des ontologies pour rendre les informations intégrées accessibles aux acteurs métier. Après avoir évalué plusieurs options, nous avons recommandé l'utilisation d'une base de données de graphes comme Amazon Neptune. Cette solution offre une interface intuitive et des capacités d'analyse avancées, permettant aux experts métiers de naviguer efficacement dans le graphe ontologique, d'exécuter des requêtes complexes et d'extraire des informations précieuses pour soutenir la prise de décision et l'optimisation des processus dans le cadre du jumeau numérique.

Malgré ces apports significatifs, notre travail présente certaines limites qu'il est important de reconnaître. L'une des principales limitations concerne les jeux de données utilisés pour évaluer notre approche. Idéalement, il aurait été préférable de disposer de jeux de données de référence, annotés par des experts et couvrant une large variété de cas d'usage industriels. Cependant, l'accès à de tels jeux de données s'est avéré complexe dans le contexte spécifique de notre étude au sein d'Airbus, en raison de la confidentialité et du caractère stratégique des données industrielles. Cette limitation dans la diversité et la représentativité des données utilisées peut avoir un impact sur l'évaluation de la généralisation de notre approche à d'autres contextes industriels.

Une autre limite concerne la difficulté d'obtenir des valeurs vraiment objectives pour évaluer les performances de notre système. En l'absence de bases de comparaison établies et de métriques standardisées dans le domaine de l'intégration des données industrielles par l'ontologification, il est difficile de situer nos résultats par rapport à l'état de l'art. De plus, notre approche repose en partie sur une sélection manuelle des concepts clés par les experts métiers, ce qui peut introduire un biais potentiel dans l'évaluation des performances.

L'utilisation de Sherlock pour le profilage automatique des données, bien que présentant de nombreux avantages, a également montré certaines limites. Notamment, sa capacité à distinguer des colonnes ayant le même type sémantique mais des rôles différents est limitée. Cette limitation souligne la nécessité d'enrichir l'approche de profilage avec des techniques complémentaires, telles que le traitement du langage naturel ou l'exploitation des connaissances du domaine et des ontologies existantes.

Enfin, malgré nos efforts pour automatiser l'alignement des ontologies reconstruites, cette tâche reste un défi majeur. La grande diversité des sources de données, la multiplicité des ontologies existantes et les différences sémantiques entre les domaines métiers ont rendu cet alignement extrêmement complexe. L'établissement de correspondances fiables et cohérentes entre les différents modèles ontologiques reste un problème ouvert, nécessitant des recherches complémentaires.

En conclusion, notre travail a apporté des contributions significatives dans le domaine de l'intégration des données IoT pour l'enrichissement des jumeaux numériques industriels. Notre approche d'ontologification et de création de liens sémantiques offre une solution prometteuse pour structurer et exploiter les données industrielles complexes. Cependant, les limites identifiées soulignent la nécessité de poursuivre les recherches pour améliorer la robustesse, la

généralisation et l'automatisation de notre approche, notamment en ce qui concerne l'alignement des ontologies et la gestion des données hétérogènes à grande échelle.

Perspectives et travaux futurs

Les résultats prometteurs de notre recherche ouvrent la voie à de nombreuses perspectives et travaux futurs pour améliorer et étendre notre approche d'intégration des données IoT dans les jumeaux numériques industriels. Ces pistes de recherche visent à adresser les limitations identifiées et à explorer de nouvelles opportunités pour renforcer l'efficacité et la pertinence de notre méthodologie.

Une première perspective importante concerne l'amélioration de la représentativité et de la diversité des jeux de données utilisés pour évaluer notre approche. Il serait bénéfique de constituer des bases de données de référence, annotées par des experts et couvrant une large variété de cas d'usage industriels. Cela nécessiterait la mise en place de collaborations étendues avec différents acteurs industriels et académiques, permettant de collecter et d'annoter des données issues de divers secteurs et contextes. Ces jeux de données de référence faciliteraient l'évaluation comparative des performances de notre approche et permettraient de mieux évaluer sa capacité de généralisation à différents contextes industriels.

Dans le même ordre d'idées, il serait pertinent de développer des métriques standardisées pour évaluer les performances des systèmes d'intégration de données industrielles par ontologification. Ces métriques devraient prendre en compte non seulement la précision et le rappel dans la reconnaissance des types sémantiques, mais aussi la qualité et la pertinence des liens ontologiques créés, ainsi que l'utilisabilité et l'exploitabilité de l'ontologie résultante dans un contexte industriel réel. L'établissement de telles métriques faciliterait la comparaison objective des différentes approches et contribuerait à l'avancement de l'état de l'art dans ce domaine.

Une autre piste de recherche prometteuse concerne l'amélioration des techniques de profilage des données. Bien que Sherlock ait démontré son efficacité, il serait intéressant d'explorer des approches hybrides combinant le profilage automatique avec des techniques avancées de traitement du langage naturel (NLP) et d'exploitation des connaissances du domaine. Par exemple, l'utilisation de modèles de langage pré-entraînés comme BERT ou GPT pourrait permettre une meilleure compréhension du contexte sémantique des données textuelles. De même, l'intégration de techniques d'apprentissage par transfert pourrait améliorer la capacité du système à s'adapter à de nouveaux domaines ou types de données.

L'automatisation de l'alignement des ontologies reconstruites reste un défi majeur qui mérite une attention particulière dans les travaux futurs. Il serait pertinent d'explorer des approches basées sur l'apprentissage profond pour découvrir automatiquement des correspondances sémantiques entre les ontologies. Par exemple, l'utilisation de réseaux de neurones graph convolutional (GCN) pourrait permettre d'exploiter la structure des ontologies pour identifier des similarités structurelles et sémantiques. De plus, l'intégration de techniques d'apprentissage actif pourrait permettre d'optimiser l'implication des experts humains dans le processus d'alignement, en ciblant leur intervention sur les cas les plus ambigus ou complexes.

La gestion des données IoT en temps réel représente un autre axe de recherche important. Il serait intéressant d'explorer des approches de traitement de flux (stream processing) pour

intégrer en continu les données IoT dans l'ontologie. Cela nécessiterait le développement d'algorithmes capables de mettre à jour dynamiquement l'ontologie en fonction des nouvelles données entrantes, tout en maintenant la cohérence et la pertinence de la structure ontologique. Des techniques d'apprentissage incrémental et d'adaptation en ligne pourraient être explorées pour permettre à l'ontologie d'évoluer de manière autonome au fil du temps.

L'amélioration de l'interprétabilité et de l'explicabilité de notre approche constitue également une piste de recherche prometteuse. Il serait bénéfique de développer des méthodes permettant de comprendre et d'expliquer les décisions prises par le système lors de l'identification des types sémantiques et de la création des liens ontologiques. L'utilisation de techniques d'interprétation des modèles d'apprentissage profond, telles que LIME ou SHAP, pourrait être explorée pour fournir des explications compréhensibles aux experts métiers sur les raisons des choix effectués par le système.

La scalabilité de notre approche pour traiter des volumes de données encore plus importants est un autre aspect à considérer dans les travaux futurs. L'exploration de techniques de calcul distribué et de parallélisation pourrait permettre d'améliorer les performances de notre système pour gérer des bases de données industrielles de très grande taille. L'utilisation de frameworks de traitement distribué comme Apache Spark ou de technologies de stockage distribuées comme Apache Cassandra pourrait être envisagée pour optimiser le traitement et le stockage des données à grande échelle.

L'intégration de notre approche avec d'autres technologies émergentes du domaine de l'industrie 4.0 représente également une perspective intéressante. Par exemple, l'exploration des synergies entre notre méthodologie d'ontologification et les technologies de blockchain pourrait ouvrir de nouvelles possibilités pour garantir la traçabilité et l'intégrité des données dans le jumeau numérique. De même, l'intégration avec des systèmes de réalité augmentée ou virtuelle pourrait offrir de nouvelles modalités d'interaction avec l'ontologie et de visualisation des connaissances.

Le développement d'une plateforme collaborative pour l'implication à grande échelle des experts métiers dans le processus d'ontologification et d'alignement constitue une autre piste de recherche prometteuse. Cette plateforme pourrait intégrer des fonctionnalités avancées de validation des concepts, de suggestion de modifications et d'annotation des données. L'utilisation de techniques de gamification pourrait être explorée pour encourager la participation active des experts et optimiser le processus de co-construction de l'ontologie.

L'extension de notre approche à d'autres domaines d'application au-delà de l'industrie aéronautique représente également une perspective intéressante. Il serait pertinent d'explorer l'adaptabilité de notre méthodologie à d'autres secteurs industriels tels que l'automobile, l'énergie ou la santé. Cela nécessiterait probablement des ajustements spécifiques à chaque domaine, mais pourrait conduire à des avancées significatives dans la gestion des connaissances et l'optimisation des processus dans ces secteurs.

Enfin, l'exploration de techniques d'apprentissage fédéré pourrait ouvrir de nouvelles perspectives pour l'intégration de données provenant de multiples organisations tout en préservant la confidentialité et la sécurité des données. Cette approche permettrait de construire

des modèles d'ontologification robustes en tirant parti des données de plusieurs entreprises sans nécessiter le partage direct des données sensibles.

En conclusion, ces perspectives et travaux futurs offrent de nombreuses opportunités pour améliorer et étendre notre approche d'intégration des données IoT dans les jumeaux numériques industriels. En explorant ces pistes de recherche, nous pourrons relever les défis identifiés et développer des solutions toujours plus performantes et adaptées aux besoins complexes de l'industrie 4.0. Ces avancées contribueront à renforcer l'interopérabilité sémantique, à optimiser la gestion des connaissances et à soutenir la prise de décision basée sur les données dans les environnements industriels modernes.

Références/Bibliographie :

Abadi, D. J., Boncz, P. A., & Harizopoulos, S. (2009). Column-oriented database systems. *Proc. VLDB Endow.*, *2*(2), 1664-1665. https://doi.org/10.14778/1687553.1687625

Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data : A survey. *The VLDB Journal*, *24*(4), 557-581. https://doi.org/10.1007/s00778-015-0389-y

Aggarwal, C. C., & Reddy, C. K. (Éds.). (2013). *Data clustering : Algorithms and applications*. CRC Press.

Aggarwal, C. C., & Zhai, C. (Éds.). (2012). *Mining Text Data*. Springer US. https://doi.org/10.1007/978-1-4614-3223-4

Allemang, D., Grist, J., & Gandon, F. (2020). *Semantic Web for the Working Ontologist* (3e éd., Vol. 33). Association for Computing Machinery. doi.org/10.1145/3382097

Alpaydin, E. (2020). *Introduction to machine learning* (Fourth edition). The MIT Press. https://mitpress.mit.edu/books/introduction-machine-learning-fourth-edition

Ameri, F., Marcos, S., & Wallace, E. (2020). *Towards a Reference Ontology for Supply Chain Management*. 6.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. https://doi.org/10.15252/msb.20156651

Angles, R., & Gutierrez, C. (2008). Survey of graph database models. *ACM Comput. Surv.*, *40*(1), 1:1-1:39. https://doi.org/10.1145/1322432.1322433

Antoniou, G., Groth, P., van Harmelen, F., & Hoekstra, R. (2012). *A Semantic Web Primer* (3e éd.). MIT Press. https://mitpress.mit.edu/books/semantic-web-primer-third-edition

Antoniou, G., & Van Harmelen, F. (2004). *A semantic Web primer*. MIT Press.

Anushruthika. (2023, novembre 9). All about Activation functions & Choosing the Right Activation Function. *Medium*. https://medium.com/@anushruthikae/all-about-activation-functions-choosing-the-right-activation-function-a63844e49a2a

Arif-Uz-Zaman, K., Cholette, M. E., Ma, L., & Karim, A. (2017). Extracting failure time data from industrial maintenance records using text mining. *Advanced Engineering Informatics*, *33*, 388-396. https://doi.org/10.1016/j.aei.2016.11.004

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*(none), 40-79. https://doi.org/10.1214/09-SS054

Arora, S., Liang, Y., & Ma, T. (2017). *A simple but tough-to-beat baseline for sentence embeddings*. 5th International Conference on Learning Representations, ICLR 2017. https://collaborate.princeton.edu/en/publications/a-simple-but-tough-to-beat-baseline-for-sentence-embeddings-3

Arp, R., Smith, B., & Spear, A. D. (2015). *Building Ontologies with Basic Formal Ontology*. The MIT Press. https://www.jstor.org/stable/j.ctt17kk7vw

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In I. Gurevych & Y. Miyao (Éds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 789-798). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1073

Ashton, K. (2009). That « Internet of Things » Thing. *RFID Journal*, *22*(7), 97-114.

Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things : A survey. *Computer Networks*, *54*(15), 2787-2805. https://doi.org/10.1016/j.comnet.2010.05.010

Baader, F. (Éd.). (2003). *The description logic handbook : Theory, implementation, and applications*. Cambridge University Press.

Baader, F., Horrocks, I., Lutz, C., & Sattler, U. (2017). *An Introduction to Description Logic*. Cambridge University Press. https://doi.org/10.1017/9781139025355

Baeza-Yates, R., Ribeiro-neto, B., Mills, D., Bonn, O., Juan, S., Mexico, M., Taipei, C., Wesley, A., & Limited, L. (1999). *Modern Information Retrieval*.

Baïna, S. (2006). *Interopérabilité dirigée par les modèles : Une Approche Orientée Produit pour l'interopérabilité des systèmes d'entreprise* [These de doctorat, Université Henri Poincaré - Nancy 1]. https://hal.univ-lorraine.fr/tel-01746552

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning : A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423-443. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2018.2798607

Banerjee, A., Dalal, R., Mittal, S., & Joshi, K. P. (2017). Generating Digital Twin Models using Knowledge Graphs for Industrial Production Lines. *Proceedings of the 2017 ACM on Web Science Conference*, 425-430. https://doi.org/10.1145/3091478.3162383

Barlaug, N., & Gulla, J. A. (2021). Neural Networks for Entity Matching : A Survey. *ACM Transactions on Knowledge Discovery from Data*, *15*(3), 52:1-52:37. https://doi.org/10.1145/3442200

Barnaghi, P., Wang, W., Henson, C., & Taylor, K. (2012). Semantics for the Internet of Things : Early Progress and Back to the Future. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *8*(1), 1-21. https://doi.org/10.4018/jswis.2012010101

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict ! A systematic comparison of context-counting vs. Context-predicting semantic vectors. In K. Toutanova & H. Wu (Éds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

(Volume 1 : Long Papers) (p. 238-247). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1023

Barzenji, H. (2021). Sentiment analysis of Twitter texts using Machine learning algorithms. *Academic Platform Journal of Engineering and Science*, *9*(3), 460-471. https://doi.org/10.21541/apjes.939338

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, *41*(3), 16:1-16:52. https://doi.org/10.1145/1541880.1541883

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks* (arXiv:1806.01261). arXiv. https://doi.org/10.48550/arXiv.1806.01261

Bebee, B., Choi, D., Gupta, A., Gutmans, A., Khandelwal, A., Kiran, Y., Mallidi, S., McGaughy, B., Personick, M., Rajan, K., Rondelli, S., Ryazanov, A., Schmidt, M., Sengupta, K., Thompson, B., Vaidya, D., & Wang, S. (2018). *Amazon Neptune : Graph Data Management in the Cloud*. International Workshop on the Semantic Web. https://www.semanticscholar.org/paper/Amazon-Neptune%3A-Graph-Data-Management-in-the-Cloud-Bebee-Choi/7d539db059514f6e7f1a08239031cdd517a106c6

Becker, M. B. (2012). *Interoperability Case Study : Cloud Computing*. https://papers.ssrn.com/abstract=2046987

Bellman, R. E. (1961). *Adaptive Control Processes : A Guided Tour*. Princeton University Press. https://doi.org/10.1515/9781400874668

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning : A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). *A Neural Probabilistic Language Model*.

Bennett, J., & Lanning, S. (2007). *The Netflix Prize*. https://www.semanticscholar.org/paper/The-Netflix-Prize-Bennett-Lanning/31af4b8793e93fd35e89569ccd663ae8777f0072

Benson, T., & Grieve, G. (2016). *Principles of health interoperability : SNOMED CT, HL7 and FHIR*. https://doi.org/10.1007/978-3-319-30370-3

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, *13*(null), 281-305.

Berners-Lee, T., Hendler, J., & Ora, L. (2001, mai). The Semantic Web. *Scientific American*, *284*(5), 34-43.

Bernstein, P. A., & Haas, L. M. (2008). Information integration in the enterprise. *Commun. ACM*, *51*(9), 72-79. https://doi.org/10.1145/1378727.1378745

Bhowmick, A., & Hazarika, S. M. (2018). E-Mail Spam Filtering : A Review of Techniques and Trends. In A. Kalam, S. Das, & K. Sharma (Éds.), *Advances in Electronics, Communication and Computing* (Vol. 443, p. 583-590). Springer Singapore. https://doi.org/10.1007/978-981-10-4765-7_61

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data—The Story So Far: *International Journal on Semantic Web and Information Systems*, *5*(3), 1-22. https://doi.org/10.4018/jswis.2009081901

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993-1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146. https://doi.org/10.1162/tacl_a_00051

Boje, C., Guerriero, A., Kubicki, S., & Rezgui, Y. (2020). Towards a semantic Construction Digital Twin : Directions for future research. *Automation in Construction*, *114*, 103179. https://doi.org/10.1016/j.autcon.2020.103179

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, *29*. https://papers.nips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Bonnet, A. (2023, novembre 23). *Accuracy vs. Precision vs. Recall in Machine Learning : What is the Difference?* Encord. https://encord.com/blog/classification-metrics-accuracy-precision-recall/

Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). Fog computing and its role in the internet of things. *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, 13-16. https://doi.org/10.1145/2342509.2342513

Bordes, A., Chopra, S., & Weston, J. (2014). Question Answering with Subgraph Embeddings. In A. Moschitti, B. Pang, & W. Daelemans (Éds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (p. 615-620). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1067

Borgo, S., & Leitão, P. (2007). Foundations for a Core Ontology of Manufacturing. In R. Sharman, R. Kishore, & R. Ramesh (Éds.), *Ontologies : A Handbook of Principles, Concepts and*

Applications in Information Systems (p. 751-775). Springer US. https://doi.org/10.1007/978-0-387-37022-4_27

Botta, A., Donato, W., Persico, V., & Pescapè, A. (2015). Integration of Cloud computing and Internet of Things : A survey. *Future Generation Computer Systems*, *56*. https://doi.org/10.1016/j.future.2015.09.021

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In Y. Lechevallier & G. Saporta (Éds.), *Proceedings of COMPSTAT'2010* (p. 177-186). Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2604-3_16

Bottou, L., & Bousquet, O. (2007). The Tradeoffs of Large Scale Learning. *Advances in Neural Information Processing Systems*, *20*. https://papers.nips.cc/paper_files/paper/2007/hash/0d3180d672e08b4c5312dcdafdf6ef36-Abstract.html

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, *60*(2), 223-311. https://doi.org/10.1137/16M1080173

Bottou, L. E. (1998). *Online Learning and Stochastic Approximations*. https://www.semanticscholar.org/paper/Online-Learning-and-Stochastic-Approximations-Bottou/6a97adbfaeecd5c1eeb3ae9c76a3842d4858cc06

Bourey, J.-P., Grangel, R., Ducq, Y., Berre, A.-J., Bertoni, M., D'Antonio, F., Daclin, N., Doumeingts, G., Grandin-Dubost, M., & Kalampoukas, K. (2007). *Report on Model Driven Interoperability*. https://hal.archives-ouvertes.fr/hal-00356088

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Bridle, J. S. (1990). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In F. F. Soulié & J. Hérault (Éds.), *Neurocomputing* (p. 227-236). Springer. https://doi.org/10.1007/978-3-642-76153-9_28

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature Verification using a « Siamese » Time Delay Neural Network. *Advances in Neural Information Processing Systems*, *6*. https://papers.neurips.cc/paper_files/paper/1993/hash/288cc0ff022877bd3df94bc9360b9c5d-Abstract.html

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk : A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(1), 3-5. https://doi.org/10.1177/1745691610393980

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Card, S. K., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization : Using Vision to Think*. Morgan Kaufmann.

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. da P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, *137*, 106024. https://doi.org/10.1016/j.cie.2019.106024

Cavanillas, J. M., Curry, E., & Wahlster, W. (Éds.). (2016). *New Horizons for a Data-Driven Economy*. Springer International Publishing. https://doi.org/10.1007/978-3-319-21569-3

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., Piatetsky-Shapiro, G., & Wang, W. (2006). *Data Mining Curriculum : A Proposal (Version 1.0)*.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection : A survey. *ACM Comput. Surv.*, *41*(3), 15:1-15:58. https://doi.org/10.1145/1541880.1541882

Chapelle, O., Schölkopf, B., & Zien, A. (Éds.). (2010). *Semi-supervised learning*. MIT Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. https://doi.org/10.1613/jair.953

Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., Flouris, G., Fundulaki, I., Granada, R., Ivanova, V., Jiménez-Ruiz, E., Lambrix, P., Montanelli, S., Pesquita, C., Saveta, T., Shvaiko, P., Solimando, A., Trojahn, C., & Šváb-Zamazal, O. (2015, octobre 1). *Results of the Ontology Alignment Evaluation Initiative 2015*.

Cheatham, M., & Hitzler, P. (2013). String Similarity Metrics for Ontology Alignment. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, & K. Janowicz (Éds.), *The Semantic Web – ISWC 2013* (p. 294-309). Springer. https://doi.org/10.1007/978-3-642-41338-4_19

Chen, B., Zhang, S., & Wang, B. (2018). A Case Study of MBSE Method Used in the EMU Train Design. *2018 International Conference on Intelligent Rail Transportation (ICIRT)*, 1-5. https://doi.org/10.1109/ICIRT.2018.8641645

Chen, D., Doumeingts, G., & Vernadat, F. (2008). Architectures for enterprise integration and interoperability : Past, present and future. *Computers in Industry*, *59*(7), 647-659. https://doi.org/10.1016/j.compind.2007.12.016

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to Answer Open-Domain Questions. In R. Barzilay & M.-Y. Kan (Éds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 1870-1879). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1171

Chen, M. (2017). *Efficient Vector Representation for Documents through Corruption* (arXiv:1707.02377). arXiv. https://doi.org/10.48550/arXiv.1707.02377

Chen, M., Mao, S., & Liu, Y. (2014). Big Data : A Survey. *Mobile Networks and Applications*, *19*(2), 171-209. https://doi.org/10.1007/s11036-013-0489-0

Chen, Z., & Liu, B. (2018). *Lifelong Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-031-01581-6

Cheng, J., Chen, W., Tao, F., & Lin, C.-L. (2018). Industrial IoT in 5G environment towards smart manufacturing. *Journal of Industrial Information Integration*, *10*, 10-19. https://doi.org/10.1016/j.jii.2018.04.001

Chiang, R. H. L., Barron, T. M., & Storey, V. C. (1994). Reverse engineering of relational databases : Extraction of an EER model from a relational database. *Data & Knowledge Engineering*, *12*(2), 107-142. https://doi.org/10.1016/0169-023X(94)90011-6

Chiang, W.-L., Liu, X., Si, S., Li, Y., Bengio, S., & Hsieh, C.-J. (2019). Cluster-GCN : An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 257-266. https://doi.org/10.1145/3292500.3330925

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Christopher M. Bishop. (2006). *Pattern Recognition and Machine Learning*. https://link.springer.com/book/9780387310732

Chu, X., Morcos, J., Ilyas, I. F., Ouzzani, M., Papotti, P., Tang, N., & Ye, Y. (2015). KATARA : A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 1247-1261. https://doi.org/10.1145/2723372.2749431

Chungoora, N., Young, R. I., Gunendran, G., Palmer, C., Usman, Z., Anjum, N. A., Cutting-Decelle, A.-F., Harding, J. A., & Case, K. (2013). A model-driven ontology approach for manufacturing system interoperability and knowledge sharing. *Computers in Industry*, *64*(4), 392-401. https://doi.org/10.1016/j.compind.2013.01.003

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3642-3649. https://doi.org/10.1109/CVPR.2012.6248110

Cisco. (2018). *Cisco Visual Networking Index : Forecast and Methodology*. Cisco. https://virtualization.network/Resources/Whitepapers/0b75cf2e-0c53-4891-918e-b542a5d364c5_white-paper-c11-738085.pdf

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, 160-167. https://doi.org/10.1145/1390156.1390177

Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, *46*(1), 41-64. https://doi.org/10.1016/S0169-023X(02)00195-7

Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 191-198. https://doi.org/10.1145/2959100.2959190

Cruz, I. F., Antonelli, F. P., & Stroe, C. (2009). AgreementMaker : Efficient matching for large real-world schemas and ontologies. *Proceedings of the VLDB Endowment*, *2*(2), 1586-1589. https://doi.org/10.14778/1687553.1687598

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering : Review, approaches and open research problems. *Heliyon*, *5*(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

Dai, A. M., Olah, C., & Le, Q. V. (2015). *Document Embedding with Paragraph Vectors* (arXiv:1507.07998). arXiv. https://doi.org/10.48550/arXiv.1507.07998

Damjanovic-Behrendt, V. (2018). A Digital Twin-based Privacy Enhancement Mechanism for the Automotive Industry. *2018 International Conference on Intelligent Systems (IS)*, 272-279. https://doi.org/10.1109/IS.2018.8710526

Danjou, C., Rivest, L., & Pellerin, R. (2017). *Douze positionnements stratégiques pour l'Industrie 4.0 : Entre processus, produit et service, de la surveillance à l'autonomie*. https://www.semanticscholar.org/paper/Douze-positionnements-strat%C3%A9giques-pour-l%E2%80%99Industrie-Danjou-Rivest/fa6c1aff89e132def6418e4af3e9e3a5917c58b9

David, J., Euzenat, J., Scharffe, F., & Santos, C. T. dos. (2011). The Alignment API 4.0. *Semantic Web – Interoperability, Usability, Applicability*, *2*(1), 3-10. https://doi.org/10.3233/SW-2011-0028

Davis, R., Shrobe, H., & Szolovits, P. (1993). What Is a Knowledge Representation? *AI Magazine*, *14*(1), Article 1. https://doi.org/10.1609/aimag.v14i1.1029

de Boer, P.-T., Kroese, D. P., Mannor, S., & Rubinstein, R. Y. (2005). A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, *134*(1), 19-67. https://doi.org/10.1007/s10479-005-5724-z

Dellschaft, K., & Staab, S. (2006). On How to Perform a Gold Standard Based Evaluation of Ontology Learning. In I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, & L. M. Aroyo (Éds.), *The Semantic Web—ISWC 2006* (p. 228-241). Springer. https://doi.org/10.1007/11926078_17

Deng, L., & Yu, D. (2014). Deep Learning : Methods and Applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197-387. https://doi.org/10.1561/2000000039

Department of Defense Dictionary of Military and Associated Terms (p. 394). (2017). Joint Chiefs of Staff. https://apps.dtic.mil/sti/citations/AD1029823

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171-4186. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/N19-1423

Doan, A., Halevy, A., & Ives, Z. G. (2012). *Principles of data integration*. Morgan Kaufmann.

Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2004). Ontology Matching : A Machine Learning Approach. In S. Staab & R. Studer (Éds.), *Handbook on Ontologies* (p. 385-403). Springer. https://doi.org/10.1007/978-3-540-24750-0_19

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87. https://doi.org/10.1145/2347736.2347755

Dong, X. L., & Srivastava, D. (2013). Big data integration. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 1245-1248. https://doi.org/10.1109/ICDE.2013.6544914

Dounis, A. I., & Caraiscos, C. (2009). Advanced control systems engineering for energy and comfort management in a building environment—A review. *Renewable and Sustainable Energy Reviews*, *13*(6-7), 1246-1261.

Dragisic, Z., Ivanova, V., Lambrix, P., Faria, D., Jiménez-Ruiz, E., & Pesquita, C. (2016). User Validation in Ontology Alignment. In P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, & Y. Gil (Éds.), *The Semantic Web – ISWC 2016* (p. 200-217). Springer International Publishing. https://doi.org/10.1007/978-3-319-46523-4_13

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, *12*(61), 2121-2159.

Duong, T. H., Nguyen, N. T., & Jo, G. S. (2010). Constructing and mining a semantic-based academic social network. *J. Intell. Fuzzy Syst.*, *21*(3), 197-207.

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference : Algorithms, Evidence, and Data Science*. Cambridge University Press. https://doi.org/10.1017/CBO9781316576533

Espinoza-Arias, P., Poveda-Villalón, M., García-Castro, R., & Corcho, O. (2019). Ontological Representation of Smart City Data : From Devices to Cities. *Applied Sciences*, *9*(1), Article 1. https://doi.org/10.3390/app9010032

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

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

Euzenat, J., & Zimmermann, A. (2007). *Expressive alignment language and implementation*.

Fang, H. (2015). Managing data lakes in big data era : What's a data lake and why has it became popular in data management ecosystem. *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, 820-824. https://doi.org/10.1109/CYBER.2015.7288049

Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I. F., & Couto, F. M. (2013). The AgreementMakerLight Ontology Matching System. In R. Meersman, H. Panetto, T. Dillon, J. Eder, Z. Bellahsene, N. Ritter, P. De Leenheer, & D. Dou (Éds.), *On the Move to Meaningful Internet Systems : OTM 2013 Conferences* (Vol. 8185, p. 527-541). Springer. https://doi.org/10.1007/978-3-642-41030-7_38

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 30-35. https://doi.org/10.18653/v1/W16-2506

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), Article 3. https://doi.org/10.1609/aimag.v17i3.1230

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *1*(1), 1-47. https://doi.org/10.1093/cercor/1.1.1-a

Ferber, J. (1999). *Multi-Agent Systems : An Introduction to Distributed Artificial Intelligence* (1st éd.). Addison-Wesley Longman Publishing Co., Inc.

Fernández-López, M., & Gómez-Pérez, A. (2002). Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, *17*(2), 129-156. https://doi.org/10.1017/S0269888902000462

Fernández-López, M., Gómez-Pérez, A., & Juristo, N. (1997). Methontology : From Ontological Art Towards Ontological Engineering. *Proceedings of the AAAI97 Spring Symposium*, 33-40.

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Éds.), *Automated Machine Learning : Methods, Systems, Challenges* (p. 3-33). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1

Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D., & Antoniou, G. (2008). Ontology change : Classification and survey. *The Knowledge Engineering Review*, *23*(2), 117-152. https://doi.org/10.1017/S0269888908001367

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

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.*, *46*(4), 44:1-44:37. https://doi.org/10.1145/2523813

Gamache, S., Abdul-Nour, G., & Baril, C. (2017). *Industrie 4.0 dans les PME québécoises : Bilan et premiers constats*.

Gandomi, A., & Haider, M. (2015). Beyond the hype : Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining* (Vol. 72). Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2015). *Bayesian Data Analysis* (3e éd.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gerber, A. J., Barnard, A., & Merwe, A. J. (2006). A Semantic Web Status Model. *Undefined*, 10.

Ghahramani, Z. (2004). Unsupervised Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Éds.), *Advanced Lectures on Machine Learning : ML Summer Schools 2003, Canberra, Australia, February 2—14, 2003, Tübingen, Germany, August 4—16, 2003, Revised Lectures* (p. 72-112). Springer. https://doi.org/10.1007/978-3-540-28650-9_5

Gholami, P., & Hafezalkotob, A. (2017). Maintenance scheduling using data mining techniques and time series models. *International Journal of Management Science and Engineering Management*, *13*, 1-8. https://doi.org/10.1080/17509653.2017.1314201

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning*, 1263-1272. https://proceedings.mlr.press/v70/gilmer17a.html

Gittens, A., Achlioptas, D., & Mahoney, M. W. (2017). Skip-Gram − Zipf + Uniform = Vector Additivity. In R. Barzilay & M.-Y. Kan (Éds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 69-76). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1007

Gladkova, A., & Drozd, A. (2016). Intrinsic Evaluations of Word Embeddings : What Can We Do Better? *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 36-42. https://doi.org/10.18653/v1/W16-2507

Goldberg, D. E., & Holland, J. H. (1988). Genetic Algorithms and Machine Learning. *Machine Learning*, *3*(2), 95-99. https://doi.org/10.1023/A:1022602019183

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Springer International Publishing. https://doi.org/10.1007/978-3-031-02165-7

Goldberg, Y., & Levy, O. (2014). *word2vec Explained : Deriving Mikolov et al.'s negative-sampling word-embedding method* (arXiv:1402.3722). arXiv. https://doi.org/10.48550/arXiv.1402.3722

Gonen, H., & Goldberg, Y. (2019). *Lipstick on a Pig : Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them*. 60-63. https://aclanthology.org/W19-3621

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*. https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Google Developers. (s. d.). *Représentations vectorielles continues : Traduction dans un espace de dimensions inférieures*. Google for Developers. Consulté 15 septembre 2024, à l'adresse https://developers.google.com/machine-learning/crash-course/embeddings/embedding-space?hl=fr

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645-6649. https://doi.org/10.1109/ICASSP.2013.6638947

Grieves, M. (2011). *Virtually Perfect : Driving Innovative and Lean Products Through Product Lifecycle Management*. Space Coast Press.

Grieves, M. (2015). *Digital Twin : Manufacturing Excellence through Virtual Factory Replication*.

Grieves, M. (2016). *Origins of the Digital Twin Concept*. https://doi.org/10.13140/RG.2.2.26367.61609

Grieves, M., & Vickers, J. (2017). Digital Twin : Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In F.-J. Kahlen, S. Flumerfelt, & A. Alves (Éds.), *Transdisciplinary Perspectives on Complex Systems : New Findings and Approaches* (p. 85-113). Springer International Publishing. https://doi.org/10.1007/978-3-319-38756-7_4

Grover, A., & Leskovec, J. (2016). node2vec : Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855-864. https://doi.org/10.1145/2939672.2939754

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199-220. https://doi.org/10.1006/KNAC.1993.1008

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5), 907-928. https://doi.org/10.1006/ijhc.1995.1081

Guarino, N. (1998). *Some Organizing Principles For A Unified Top-Level Ontology*.

Guarino, N., Oberle, D., & Staab, S. (2009). What Is an Ontology? In S. Staab & R. Studer (Éds.), *Handbook on Ontologies* (p. 1-17). Springer. https://doi.org/10.1007/978-3-540-92673-3_0

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT) : A vision, architectural elements, and future directions. *Future Generation Computer Systems*, *29*(7), 1645-1660. https://doi.org/10.1016/j.future.2013.01.010

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), Article 2. https://doi.org/10.1609/aimag.v40i2.2850

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*(null), 1157-1182.

Gyrard, A., Serrano, M., & Atemezing, G. A. (2015). Semantic web methodologies, best practices and ontology engineering applied to Internet of Things. *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, 412-417. https://doi.org/10.1109/WF-IoT.2015.7389090

Gyrard, A., Zimmermann, A., & Sheth, A. (2018). Building IoT-Based Applications for Smart Cities : How Can Ontology Catalogs Help? *IEEE Internet of Things Journal*, *5*(5), 3978-3990. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2018.2854278

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hamilton, W. L., Ying, R., & Leskovec, J. (2017a). Inductive Representation Learning on Large Graphs. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *30*, 1025-1035. https://arxiv.org/abs/1706.02216v4

Hamilton, W. L., Ying, R., & Leskovec, J. (2017b). Representation Learning on Graphs : Methods and Applications. *IEEE Data Eng. Bull.* https://www.semanticscholar.org/paper/Representation-Learning-on-Graphs%3A-Methods-and-Hamilton-Ying/ecf6c42d84351f34e1625a6a2e4cc6526da45c74

Han, J., Kamber, M., & Pei, J. (Éds.). (2012). *Data Mining : Concepts and Techniques*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00016-2

Han, J., & Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In J. Mira & F. Sandoval (Éds.), *From Natural to Artificial Neural Computation* (p. 195-201). Springer. https://doi.org/10.1007/3-540-59497-3_175

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520

Hart, L. (2015, juillet 30). *Introduction To Model-Based System Engineering (MBSE) and SysML*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hause, M., Hummell, J., & Grelier, F. (2018). MBSE Driven IoT for Smarter Cities. *2018 13th Annual Conference on System of Systems Engineering (SoSE)*, 365-371. https://doi.org/10.1109/SYSOSE.2018.8428705

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1026-1034. https://doi.org/10.1109/ICCV.2015.123

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/CVPR.2016.90

Hendrycks, D., & Gimpel, K. (2018). *A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks* (arXiv:1610.02136). arXiv. https://doi.org/10.48550/arXiv.1610.02136

Hepp, M., & Roman, D. (2007). An Ontology Framework for Semantic Business Process Management. In A. Oberweis, C. Weinhardt, H. Gimpel, A. Koschmider, V. Pankratius, & B. Schnizler (Éds.), *eOrganisation : Service-, Prozess-, Market-Engineering : 8. Internationale Tagung Wirtschaftsinformatik—Band 1, WI 2007, Karlsruhe, Germany,*

February 28—March 2, 2007 (p. 423-440). Universitaetsverlag Karlsruhe. http://aisel.aisnet.org/wi2007/27

Hewlett Packard Enterprise. (2016). *Powering the Intelligent Edge : HPE's Strategy and Direction for IoT &…*. https://www.slideshare.net/Hadoop_Summit/powering-the-intelligent-edge-hpes-strategy-and-direction-for-iot-big-data

Hilbert, M. (2016). Big Data for Development : A Review of Promises and Challenges. *Development Policy Review*, *34*(1), 135-174. https://doi.org/10.1111/dpr.12142

Hill, F., Cho, K., Jean, S., Devin, C., & Bengio, Y. (2015). *Embedding Word Similarity with Neural Machine Translation* (arXiv:1412.6448). arXiv. https://doi.org/10.48550/arXiv.1412.6448

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)* [W3C Recommendation]. W3C. http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge Graphs. *ACM Comput. Surv.*, *54*(4), 71:1-71:37. https://doi.org/10.1145/3447772

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems : An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (Reprint edition). Bradford Books.

Holzinger, A. (2016). Interactive machine learning for health informatics : When do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119-131. https://doi.org/10.1007/s40708-016-0042-6

Horrocks, I. (2008). Ontologies and the semantic web. *Commun. ACM*, *51*(12), 58-67. https://doi.org/10.1145/1409360.1409377

Horrocks, I., Kutz, O., & Sattler, U. (2006). The even more irresistible SROIQ. *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning*, 57-67.

Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2003). From SHIQ and RDF to OWL : The making of a Web Ontology Language. *Journal of Web Semantics*, *1*(1), 7-26. https://doi.org/10.1016/j.websem.2003.07.001

Horrocks, I., & Sattler, U. (2004). Decidability of SHIQ with complex role inclusion axioms. *Artificial Intelligence*, *160*(1), 79-104. https://doi.org/10.1016/j.artint.2004.06.002

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In I. Gurevych & Y. Miyao (Éds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 328-339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

Hu, S., Wang, H., She, C., & Wang, J. (2011). AgOnt : Ontology for Agriculture Internet of Things. In D. Li, Y. Liu, & Y. Chen (Éds.), *Computer and Computing Technologies in Agriculture IV* (Vol. 344, p. 131-137). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-18333-1_18

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

IEC 61499. (s. d.). Consulté 23 avril 2023, à l'adresse https://iec61499.com/

IEEE Standard Computer Dictionary : A Compilation of IEEE Standard Computer Glossaries. (1991). *IEEE Std 610*, 1-217. IEEE Std 610. https://doi.org/10.1109/IEEESTD.1991.106963

Industry 4.0 : Reinvigorating ASEAN manufacturing for the future. (s. d.). McKinsey. Consulté 4 mars 2023, à l'adresse https://www.mckinsey.com/capabilities/operations/our-insights/industry-4-0-reinvigorating-asean-manufacturing-for-the-future

International Organization for Standardization. (2015). *Information technology—Vocabulary* (ISO/IEC 2382:2015; Version 1). https://www.iso.org/standard/63598.html

Ioffe, S., & Szegedy, C. (2015). Batch normalization : Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 448-456.

ISA95. (s. d.). Isa.Org. Consulté 16 avril 2023, à l'adresse http://www.isa.org/isa95

Isaac, A., van der Meij, L., Schlobach, S., & Wang, S. (2007). An Empirical Study of Instance-Based Ontology Matching. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, & P. Cudré-Mauroux (Éds.), *The Semantic Web* (p. 253-266). Springer. https://doi.org/10.1007/978-3-540-76298-0_19

Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daumé III, H. (2015). Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In C. Zong & M. Strube (Éds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)* (p. 1681-1691). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1162

Jahani, H., Jain, R., & Ivanov, D. (2023). Data science and big data analytics : A systematic review of methodologies used in the supply chain and logistics research. *Annals of Operations Research*, 1-58. https://doi.org/10.1007/s10479-023-05390-7

Jain, A. K. (2010). Data clustering : 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning : With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Janocha, K., & Czarnecki, W. M. (2017). On Loss Functions for Deep Neural Networks in Classification. *Schedae Informaticae*, *1/2016*. https://doi.org/10.4467/20838476SI.16.004.6185

Jiménez-Ruiz, E., & Cuenca Grau, B. (2011). LogMap : Logic-Based and Scalable Ontology Matching. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, & E. Blomqvist (Éds.), *The Semantic Web – ISWC 2011* (p. 273-288). Springer. https://doi.org/10.1007/978-3-642-25073-6_18

Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the Digital Twin : A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, *29*, 36-52. https://doi.org/10.1016/j.cirpj.2020.02.002

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning : Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd éd.). https://web.stanford.edu/~jurafsky/slp3/

Kadadi, A., Agrawal, R., Nyamful, C., & Atiq, R. (2014). Challenges of data integration and interoperability in big data. *2014 IEEE International Conference on Big Data (Big Data)*, 38-40. 2014 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/BigData.2014.7004486

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Kamar, E. (2016). Directions in hybrid intelligence : Complementing AI systems with human intelligence. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 4070-4073.

Kapil, D. (2019, juin 21). Stochastic vs Batch Gradient Descent. *Medium*. https://medium.com/@divakar_239/stochastic-vs-batch-gradient-descent-8820568eada1

Karray, M. H., Otte, N., Rai, R., Ameri, F., Kulvatunyou, B., Smith, B., Kiritsis, D., Will, C., & Arista, R. (2021). The Industrial Ontologies Foundry (IOF) perspectives. *NIST*, 6.

Keet, C. M. (2018). *An Introduction to Ontology Engineering*. College Publications.

Keet, C. M., Lawrynowicz, A., D'Amato, C., Kalousis, A., Nguyen, P., Palma, R., Stevens, R., & Hilario, M. (2015). The Data Mining OPtimization Ontology. *Journal of Web Semantics*, *32*, 43-53. https://doi.org/10.1016/j.websem.2015.01.001

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). *On Large-Batch Training for Deep Learning : Generalization Gap and Sharp Minima* (arXiv:1609.04836). arXiv. https://doi.org/10.48550/arXiv.1609.04836

Kharlamov, E., Grau, B. C., Jiménez-Ruiz, E., Lamparter, S., Mehdi, G., Ringsquandl, M., Nenov, Y., Grimm, S., Roshchin, M., & Horrocks, I. (2016). Capturing Industrial Information Models with Ontologies and Constraints. In P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, & Y. Gil (Éds.), *The Semantic Web – ISWC 2016* (p. 325-343). Springer International Publishing. https://doi.org/10.1007/978-3-319-46547-0_30

Khodkari, H., & Maghrebi, S. G. (2016). Necessity of the integration Internet of Things and cloud services with quality of service assurance approach. *Bulletin de La Société Royale Des Sciences de Liège*, *85*, 12.

Kiela, D., Bulat, L., Vero, A. L., & Clark, S. (2016). *Virtual Embodiment : A Scalable Long-Term Strategy for Artificial Intelligence Research* (arXiv:1610.07432). arXiv. https://doi.org/10.48550/arXiv.1610.07432

Kingma, D. P., & Ba, J. (2017). *Adam : A Method for Stochastic Optimization* (arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980

Kipf, T. N., & Welling, M. (2017, février 22). Semi-Supervised Classification with Graph Convolutional Networks. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. International Conference on Learning Representations. http://arxiv.org/abs/1609.02907

Klir, G. J., & Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic : Theory and Applications*. Prentice Hall PTR.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, 1137-1143.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1), 273-324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kohler Dorothée (avec Weisz Jean-Daniel & Ehrhart Katharina). (2016). *Industrie 4.0 : Les défis de la transformation numérique du modèle industriel allemand / Dorothée Kohler, Jean-Daniel Weisz ; avec la collaboration de Katharina Ehrhart,...* La Documentation française.

Kolyvakis, P., Kalousis, A., & Kiritsis, D. (2018). DeepAlignment : Unsupervised Ontology Matching with Refined Word Vectors. In M. Walker, H. Ji, & A. Stent (Éds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)* (p. 787-798). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1072

Konstantas, D., Bourrières, J.-P., Léonard, M., & Boudjlida, N. (Éds.). (2006). *Interoperability of Enterprise Software and Applications*. Springer. https://doi.org/10.1007/1-84628-152-0

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, *42*(8), 30-37. Computer. https://doi.org/10.1109/MC.2009.263

Kotsiantis, S. (2007). Supervised Machine Learning : A Review of Classification Techniques. *Informatica (Slovenia)*, *31*, 249-268.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International journal of computer science*, *1*(2), 111-117.

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

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital Twin in manufacturing : A categorical literature review and classification. *IFAC-PapersOnLine*, *51*(11), 1016-1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Krötzsch, M., Simancik, F., & Horrocks, I. (2013). *A Description Logic Primer* (arXiv:1201.4089). arXiv. https://doi.org/10.48550/arXiv.1201.4089

Kumar, V., Khamis, A. M., Fiorini, S., Carbonera, J., Alarcos, A. O., Habib, M., Gonçalves, P., Howard, L. I., & Olszewska, J. (2019). Ontologies for Industry 4.0. *The Knowledge Engineering Review*, *34*(e17), e17. https://doi.org/10.1017/S0269888919000109

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From word embeddings to document distances. *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 957-966. https://proceedings.mlr.press/v37/kusnerb15.html

Lafferty, J., McCallum, A., & Pereira, F. (2001, juin 28). *Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data*. International Conference on Machine Learning. https://www.semanticscholar.org/paper/Conditional-Random-Fields%3A-Probabilistic-Models-for-Lafferty-McCallum/f4ba954b0412773d047dc41231c733de0c1f4926

Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded Methods. In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Éds.), *Feature Extraction : Foundations and Applications* (p. 137-165). Springer. https://doi.org/10.1007/978-3-540-35488-8_6

Lassila, O., & Swick, R. R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification* (Recommandation W3C REC-rdf-syntax-19990222). W3C. https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

Lau, J. H., & Baldwin, T. (2016). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In P. Blunsom, K. Cho, S. Cohen, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, & S. W. Yih (Éds.), *Proceedings of the 1st Workshop on Representation Learning for NLP* (p. 78-86). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-1609

LaWell, M. (2015, juillet). *Industry 4.0 vs. the Industrial Internet : A Primer*. IndustryWeek. https://www.industryweek.com/technology-and-iiot/information-technology/article/22006000/industry-40-vs-the-industrial-internet-a-primer

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning*, 1188-1196. https://proceedings.mlr.press/v32/le14.html

Lebret, R., Grangier, D., & Auli, M. (2016). Neural Text Generation from Structured Data with Application to the Biography Domain. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1203-1213. https://doi.org/10.18653/v1/D16-1128

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. Proceedings of the IEEE. https://doi.org/10.1109/5.726791

Lee, C. K. M., & Zhang, S. Z. (2016). Development of an Industrial Internet of Things Suite for Smart Factory towards Re-industrialization in Hong Kong. *Proceedings of the 6th International Workshop of Advanced Manufacturing and Automation*. 6th International Workshop of Advanced Manufacturing and Automation, Manchester, UK. https://doi.org/10.2991/iwama-16.2016.54

Lee, J., Azamfar, M., & Singh, J. (2019). A blockchain enabled Cyber-Physical System architecture for Industry 4.0 manufacturing systems. *Manufacturing Letters*, *20*, 34-39. https://doi.org/10.1016/j.mfglet.2019.05.003

Lee, J., Bagheri, B., & Kao, H.-A. (2015). A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, *3*, 18-23. https://doi.org/10.1016/j.mfglet.2014.12.001

Lee, J., Davari, H., Singh, J., & Pandhare, V. (2018). Industrial Artificial Intelligence for industry 4.0-based manufacturing systems. *Manufacturing Letters*, *18*, 20-23. https://doi.org/10.1016/j.mfglet.2018.09.002

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

Lemaignan, S., Siadat, A., Dantan, J.-Y., & Semenenko, A. (2006). MASON : A Proposal For An Ontology Of Manufacturing Domain. *IEEE Workshop on Distributed Intelligent Systems: Collective Intelligence and Its Applications (DIS'06)*, 195-200. https://doi.org/10.1109/DIS.2006.48

Lemus, L., Mohamad-Mezher, A., Barbecho, P., Astudillo, J., & Aguilar-Igartua, M. (2021). A Multimetric Predictive ANN-Based Routing Protocol for Vehicular Ad Hoc Networks. *IEEE Access*, *PP*, 1-1. https://doi.org/10.1109/ACCESS.2021.3088474

Lenzerini, M. (2002). Data integration : A theoretical perspective. *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 233-246. https://doi.org/10.1145/543613.543644

Leveson, N. G. (2012). *Engineering a Safer World : Systems Thinking Applied to Safety*. The MIT Press. https://doi.org/10.7551/mitpress/8179.001.0001

Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, *17*(1), 1334-1373.

Levy, O., & Goldberg, Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. *Advances in Neural Information Processing Systems*, *27*.

https://papers.nips.cc/paper_files/paper/2014/hash/feab05aa91085b7a8012516bc3533958-Abstract.html

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211-225. https://doi.org/10.1162/tacl_a_00134

Li, G., Muller, M., Thabet, A., & Ghanem, B. (2019). *DeepGCNs : Can GCNs Go As Deep As CNNs?* 9267-9276. https://openaccess.thecvf.com/content_ICCV_2019/html/Li_DeepGCNs_Can_GCNs_Go_As_Deep_As_CNNs_ICCV_2019_paper.html

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection : A Data Perspective. *ACM Comput. Surv.*, *50*(6), 94:1-94:45. https://doi.org/10.1145/3136625

Li, L., Soskin, N. L., Jbara, A., Karpel, M., & Dori, D. (2019). Model-Based Systems Engineering for Aircraft Design With Dynamic Landing Constraints Using Object-Process Methodology. *IEEE Access*, *7*, 61494-61511. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2915917

Li, M., Zhang, T., Chen, Y., & Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 661-670. https://doi.org/10.1145/2623330.2623612

Li, X., Zhang, W., & Ding, Q. (2019). Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliability Engineering & System Safety*, *182*, 208-218. https://doi.org/10.1016/j.ress.2018.11.011

Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W.-C. (2020). Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment*, *14*(1), 50-60. https://doi.org/10.14778/3421424.3421431

Liao, Y., Deschamps, F., Loures, E. de F. R., & Ramos, L. F. P. (2017). Past, present and future of Industry 4.0—A systematic literature review and research agenda proposal. *International Journal of Production Research*, *55*(12), 3609-3629. https://doi.org/10.1080/00207543.2017.1308576

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019-1031. https://doi.org/10.1002/asi.20591

Lipton, Z. C. (2018). The Mythos of Model Interpretability : In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57. https://doi.org/10.1145/3236386.3241340

Little, R., & Rubin, D. (2019). *Statistical Analysis with Missing Data* (1re éd.). Wiley. https://doi.org/10.1002/9781119482260

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). *A Survey on Contextual Embeddings* (arXiv:2003.07278). arXiv. https://doi.org/10.48550/arXiv.2003.07278

Liu, Y., Dong, B., Guo, B., Yang, J., & Peng, W. (2015). Combination of Cloud Computing and Internet of Things (IOT) in Medical Monitoring Systems. *International Journal of Hybrid Information Technology*, *8*(12), 367-376. https://doi.org/10.14257/ijhit.2015.8.12.28

Liu, Z., Chen, C., Li, L., Zhou, J., Li, X., Song, L., & Qi, Y. (2019). GeniePath : Graph Neural Networks with Adaptive Receptive Paths. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), Article 01. https://doi.org/10.1609/aaai.v33i01.33014424

Lohmann, S., Negru, S., Haag, F., & Ertl, T. (2016). Visualizing ontologies with VOWL. *Semantic Web*, *7*(4), 399-419. https://doi.org/10.3233/SW-150200

López de Prado, M. M. (2018). *Advances in financial machine learning*. Wiley.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc. icml*, *30*(1), 3. http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579-2605.

Madni, A. M., Madni, C. C., & Lucero, S. D. (2019). Leveraging Digital Twin Technology in Model-Based Systems Engineering. *Systems*, *7*(1), Article 1. https://doi.org/10.3390/systems7010007

Maier, A., Schnurr, H.-P., & Sure, Y. (2003). Ontology-Based Information Integration in the Automotive Industry. *The Semantic Web - ISWC 2003*, *2870*, 897-912. https://doi.org/10.1007/978-3-540-39718-2_57

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using Support Vector Machines. *Information Sciences*, *179*(13), 2208-2217. https://doi.org/10.1016/j.ins.2009.02.014

Marcus, G. (2018). *Deep Learning : A Critical Appraisal* (arXiv:1801.00631). arXiv. https://doi.org/10.48550/arXiv.1801.00631

Mascardi, V., Cordì, V., & Rosso, P. (2007). *A Comparison of Upper Ontologies.* 55-64.

Masters, D., & Luschi, C. (2018, avril 20). *Revisiting Small Batch Training for Deep Neural Networks*. arXiv.Org. https://arxiv.org/abs/1804.07612v1

Mattern, F., & Floerkemeier, C. (2010). From the Internet of Computers to the Internet of Things. In K. Sachs, I. Petrov, & P. Guerrero (Éds.), *From Active Data Management to Event-Based*

Systems and More : Papers in Honor of Alejandro Buchmann on the Occasion of His 60th Birthday (Vol. 6462, p. 242-259). Springer. https://doi.org/10.1007/978-3-642-17226-7_15

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data : A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1956). The Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, *27*(4), Article 4. https://doi.org/10.1609/aimag.v27i4.1904

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115-133. https://doi.org/10.1007/BF02478259

Mehl, D., & Anderson, N. (2021, février 24). *A brave new world for manufacturing*. Kearney. https://www.kearney.com/service/operations-performance/article/-/insights/a-brave-new-world-for-manufacturing-the-state-of-industry-4.0

Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding : A versatile graph matching algorithm and its application to schema matching. *Proceedings 18th International Conference on Data Engineering*. 18th International Conference on Data Engineering, San Jose, CA, USA. https://doi.org/10.1109/icde.2002.994702

Merlo, C., Vicien, G., & Ducq, Y. (2014). Interoperability Modelling Methodology for Product Design Organisations. *International Journal of Production Research*, *52*. https://doi.org/10.1080/00207543.2013.774484

Meyer, M., Yu, Z., Gulati, P., Delforouzi, A., Roggenbuck, J., & Wolf, K. (2020). *Ontologies for digital twins in smart manufacturing*. https://doi.org/10.13140/RG.2.2.11346.17607

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*. https://papers.nips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Miller, J., & Mukerji, J. (2003). *MDA Guide Version 1.0.1* (p. 51). Object Management Group.

Minsky, M., & Papert, S. A. (1969). *Perceptrons : An introduction to computational geometry* (2. print. with corr). The MIT Press.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. The MIT Press. https://doi.org/10.7551/mitpress/3927.001.0001

Mitchell, T. M. (1997). *Machine Learning* (1st edition). McGraw-Hill Education. http://www.cs.cmu.edu/~tom/mlbook.html

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533. https://doi.org/10.1038/nature14236

Modelica Association. (2023). *Modelica® – A Unified Object-Oriented Language for Systems Modeling : Language Specification* (Version 3.6). https://modelica.org

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (Second edition). The MIT Press.

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com.

Müller, J. M., & Voigt, K.-I. (2018). Sustainable Industrial Value Creation in SMEs : A Comparison between Industry 4.0 and Made in China 2025. *International Journal of Precision Engineering and Manufacturing-Green Technology*, *5*(5), 659-670. https://doi.org/10.1007/s40684-018-0056-z

Musen, M. A. (2015). The protégé project : A look back and a look forward. *AI Matters*, *1*(4), 4-12. https://doi.org/10.1145/2757001.2757003

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M. B., Neil, D., Saffari, A., Mead, R., Hautbergue, G. M., Holbrook, J. D., & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, *16*(8), Article 8. https://doi.org/10.1038/s41582-020-0377-8

Nahavandi, S. (2019). Industry 5.0—A Human-Centric Solution. *Sustainability*, *11*(16), Article 16. https://doi.org/10.3390/su11164371

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 807-814.

Naumann, F. (2014). Data profiling revisited. *SIGMOD Rec.*, *42*(4), 40-49. https://doi.org/10.1145/2590989.2590995

NG, A. (2018). *Machine Learning Yearning*. deeplearning.ai

Nickolls, J., Buck, I., Garland, M., & Skadron, K. (2008). Scalable Parallel Programming with CUDA : Is CUDA the parallel programming model that application developers have been waiting for? *Queue*, *6*(2), 40-53. https://doi.org/10.1145/1365490.1365500

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819346

Nouacer, R., Djemal, M., Niar, S., Mouchard, G., Rapin, N., Gallois, J.-P., Fiani, P., Chastrette, F., Lapitre, A., Adriano, T., & Mac-Eachen, B. (2016). EQUITAS : A tool-chain for functional safety and reliability improvement in automotive systems. *Microprocessors and Microsystems*, *47*, 252-261. https://doi.org/10.1016/j.micpro.2016.07.020

Noy, N. F. (2004). Semantic integration : A survey of ontology-based approaches. *SIGMOD Rec.*, *33*(4), 65-70. https://doi.org/10.1145/1041410.1041421

Noy, N. F., & McGuinness, D. L. (2001). Ontology Development 101 : A Guide to Creating Your First Ontology. *Knowledge Systems Laboratory*, *32*, 25.

Noy, N., & Musen, M. (2000a). *Algorithm and Tool for Automated Ontology Merging and Alignment*.

Noy, N., & Musen, M. (2000b). *PROMPT : Algorithm and Tool for Automated Ontology Merging and Alignment*. 450-455.

Noy, N., & Musen, M. (2001). Anchor-PROMPT : Using NonLocal Context for Semantic Matching. *IJCAI Workshop Ontol Inform Sharing*.

Nwankpa, C. E., Ijomah, W., Gachagan, A., & Marshall, S. (2020). *Activation functions : Comparison of trends in practice and research for deep learning*. 124-133. https://pureportal.strath.ac.uk/en/publications/activation-functions-comparison-of-trends-in-practice-and-researc

Object Management Group. (2015). *XML Metadata Interchange (XMI) Specification* (Version 2.5.1). http://www.omg.org/spec/XMI/2.5.1

Object Management Group. (2019). *OMG Systems Modeling Language* (Version 1.6). https://www.omg.org/spec/SysML/1.6/

Obrst, L., Chase, P., & Markeloff, R. (2012). *Developing an Ontology of the Cyber Security Domain*. Semantic Technologies for Intelligence, Defense, and Security.

https://www.semanticscholar.org/paper/Developing-an-Ontology-of-the-Cyber-Security-Domain-Obrst-Chase/860d3d4114711fa4ce9a5a4ccf362b80281cc981

Obuchowski, N. A., Lieber, M. L., & Wians, F. H. (2004). ROC curves in clinical chemistry : Uses, misuses, and possible solutions. *Clinical Chemistry*, *50*(7), 1118-1125. https://doi.org/10.1373/clinchem.2004.031823

Oh, K.-S., & Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, *37*(6), 1311-1314. https://doi.org/10.1016/j.patcog.2004.01.013

Olgac, A., & Karlik, B. (2011). Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. *International Journal of Artificial Intelligence And Expert Systems*, *1*, 111-122.

Otero-Cerdeira, L., Rodríguez-Martínez, F. J., & Gómez-Rodríguez, A. (2015). Ontology matching : A literature review. *Expert Systems with Applications*, *42*(2), 949-971. https://doi.org/10.1016/j.eswa.2014.08.032

Palantir. (2024). Palantir. https://palantir.com/docs/foundry/ontology/core-concepts//

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191

Panetto, H., Dassisti, M., & Tursi, A. (2012). ONTO-PDM : Product-driven ONTOlogy for Product Data Management interoperability within manufacturing process environment. *Advanced Engineering Informatics*, *26*(2), 334-348. https://doi.org/10.1016/j.aei.2011.12.002

Panetto, H., Iung, B., Ivanov, D., Weichhart, G., & Wang, X. (2019). Challenges for the cyber-physical manufacturing enterprises of the future. *Annual Reviews in Control*, *47*, 200-213. https://doi.org/10.1016/j.arcontrol.2019.02.002

Parent, C., & Spaccapietra, S. (2009). An Overview of Modularity. In H. Stuckenschmidt, C. Parent, & S. Spaccapietra (Éds.), *Modular Ontologies : Concepts, Theories and Techniques for Knowledge Modularization* (p. 5-23). Springer. https://doi.org/10.1007/978-3-642-01907-4_2

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks : A review. *Neural Networks*, *113*, 54-71. https://doi.org/10.1016/j.neunet.2019.01.012

Paulheim, H. (2016). Knowledge graph refinement : A survey of approaches and evaluation methods. *Semantic Web*, *8*(3), 489-508. https://doi.org/10.3233/SW-160218

Paulheim, H., Hertling, S., & Ritze, D. (2013). Towards Evaluating Interactive Ontology Matching Tools. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, & S. Rudolph (Éds.), *The Semantic Web : Semantics and Big Data* (p. 31-45). Springer. https://doi.org/10.1007/978-3-642-38288-8_3

Pearl, J. (2018). *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution* (arXiv:1801.04016). arXiv. https://doi.org/10.48550/arXiv.1801.04016

Pennington, J., Socher, R., & Manning, C. (2014). GloVe : Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *14*, 1532-1543. https://doi.org/10.3115/v1/D14-1162

Pereira, J., Pimenta, D., Dias, D., Monteiro, P., Morais, F., Santos, N., Mendonça, J. P., Pereira, F., & Carvalha, J. P. (2020). *Implementing Semantic Interoperability in Cloud Collaborative Manufacturing : A Demonstration Case for an Ontology-based Asset Efficiency Testbed*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In M. Walker, H. Ji, & A. Stent (Éds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers) (p. 2227-2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

Pham, M., Alse, S., Knoblock, C. A., & Szekely, P. (2016). Semantic Labeling : A Domain-Independent Approach. In P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, & Y. Gil (Éds.), *The Semantic Web – ISWC 2016* (p. 446-462). Springer International Publishing. https://doi.org/10.1007/978-3-319-46523-4_27

Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies : A survey on Big Data. *Information Sciences*, *275*, 314-347. https://doi.org/10.1016/j.ins.2014.01.015

Pinto, H. S., & Martins, J. P. (2004). Ontologies : How can They be Built? *Knowledge and Information Systems*, *6*(4), 441-464. https://doi.org/10.1007/s10115-003-0138-1

Porter, M. E., & Heppelmann, J. E. (2014, novembre 1). How Smart, Connected Products Are Transforming Competition. *Harvard Business Review*. https://www.hbs.edu/faculty/Pages/item.aspx?num=48195

Poveda-Villalón, M., Gómez-Pérez, A., & Suárez-Figueroa, M. C. (2014). OOPS! (OntOlogy Pitfall Scanner!) : An On-line Tool for Ontology Evaluation. *Int. J. Semant. Web Inf. Syst.*, *10*(2), 7-34. https://doi.org/10.4018/ijswis.2014040102

Powers, D. & Ailab. (2011). Evaluation : From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol*, *2*, 2229-3981. https://doi.org/10.9735/2229-3981

Pratt, M. (2001). Introduction to ISO 10303—The STEP Standard for Product Data Exchange. *Journal of Computing and Information Science in Engineering*, *1*(1), 102-103. https://doi.org/10.1115/1.1354995

Provost, F., & Fawcett, T. (2013). *Data science for business : What you need to know about data mining and data-analytic thinking* (1. ed., 2. release). O'Reilly.

Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge Graph Identification. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, & K. Janowicz (Éds.), *The Semantic Web – ISWC 2013* (p. 542-557). Springer. https://doi.org/10.1007/978-3-642-41335-3_34

Qi, Q., & Tao, F. (2018). Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0 : 360 Degree Comparison. *IEEE Access*, *6*, 3585-3593. IEEE Access. https://doi.org/10.1109/ACCESS.2018.2793265

Qi, Q., Tao, F., Hu, T., Anwer, N., Liu, A., Wei, Y., Wang, L., & Nee, A. Y. C. (2021). Enabling technologies and tools for digital twin. *Journal of Manufacturing Systems*, *58*, 3-21.

Qin, Y., Sheng, Q. Z., Falkner, N. J. G., Dustdar, S., Wang, H., & Vasilakos, A. V. (2016). When things matter : A survey on data-centric internet of things. *Journal of Network and Computer Applications*, *64*, 137-153. https://doi.org/10.1016/j.jnca.2015.12.016

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, *10*(4), 334-350. https://doi.org/10.1007/s007780100057

Rahm, E., & Do, H. H. (2000). Data Cleaning : Problems and Current Approaches. *IEEE Data Eng. Bull.*, *23*(4), 3-13.

Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. *Proceedings of the 26th Annual International Conference on Machine Learning*, 873-880. https://doi.org/10.1145/1553374.1553486

Ralph L. Ackoff. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 3-9.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). *Searching for Activation Functions* (arXiv:1710.05941). arXiv. https://doi.org/10.48550/arXiv.1710.05941

Ramnandan, S. K., Mittal, A., Knoblock, C. A., & Szekely, P. (2015). Assigning Semantic Labels to Data Sources. In F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux, & A. Zimmermann (Éds.), *The Semantic Web. Latest Advances and New Domains* (p. 403-417). Springer International Publishing. https://doi.org/10.1007/978-3-319-18818-8_25

Rao, S. K., & Prasad, R. (2018). Impact of 5G Technologies on Industry 4.0. *Wireless Personal Communications*, *100*(1), 145-159. https://doi.org/10.1007/s11277-018-5615-7

Reddi, S. J., Kale, S., & Kumar, S. (2019, avril 19). *On the Convergence of Adam and Beyond*. arXiv.Org. https://arxiv.org/abs/1904.09237v1

Référentiel Général d'Interopérabilité V1 (1; p. 119). (2009). DGME. https://www.numerique.gouv.fr/uploads/Referentiel_General_Interoperabilite_V1.pdf

Reiter, R. (1978). On Closed World Data Bases. In H. Gallaire & J. Minker (Éds.), *Logic and Data Bases* (p. 55-76). Springer US. https://doi.org/10.1007/978-1-4684-3384-5_3

Rezaei, R., Chiew, T. K., & Lee, S. P. (2014). A review on E-business Interoperability Frameworks. *Journal of Systems and Software*, *93*, 199-216. https://doi.org/10.1016/j.jss.2014.02.004

Risteska Stojkoska, B. L., & Trivodaliev, K. V. (2017). A review of Internet of Things for smart home : Challenges and solutions. *Journal of Cleaner Production*, *140*, 1454-1464. https://doi.org/10.1016/j.jclepro.2016.10.006

Ristoski, P., & Paulheim, H. (2016). *Semantic Web in Data Mining and Knowledge Discovery : A Comprehensive Survey* (SSRN Scholarly Paper 3199217). https://doi.org/10.2139/ssrn.3199217

Ritto, T. G., & Rochinha, F. A. (2020). Digital twin, physics-based model, and machine learning applied to damage detection in structures. *arXiv:2005.14360 [Eess]*. https://doi.org/10.1016/j.ymssp.2021.107614

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, *22*(3), 400-407. https://doi.org/10.1214/aoms/1177729586

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph Databases : New Opportunities for Connected Data* (2nd éd.). O'Reilly Media, Inc.

Rodriguez, P. L., & Spirling, A. (2022). Word Embeddings : What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*, *84*(1), 101-115. https://doi.org/10.1086/715162

Roman, R., Lopez, J., & Mambo, M. (2018). Mobile edge computing, Fog et al. : A survey and analysis of security threats and challenges. *Future Generation Computer Systems*, *78*, 680-698. https://doi.org/10.1016/j.future.2016.11.009

Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408. https://doi.org/10.1037/h0042519

Ruder, S. (2017). *An overview of gradient descent optimization algorithms* (arXiv:1609.04747). arXiv. https://doi.org/10.48550/arXiv.1609.04747

Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, *65*, 569-631. https://doi.org/10.1613/jair.1.11640

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., & Müller, K.-R. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, *109*(5), 756-795. Proceedings of the IEEE. https://doi.org/10.1109/JPROC.2021.3052449

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536. https://doi.org/10.1038/323533a0

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial Intelligence : A Modern Approach*. Prentice Hall.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517. https://doi.org/10.1093/bioinformatics/btm344

Saha, B., & Srivastava, D. (2014). *Data quality : The other face of Big Data*. 1294-1297. https://doi.org/10.1109/ICDE.2014.6816764

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. IBM Journal of Research and Development. https://doi.org/10.1147/rd.33.0210

Sankar, A., Wu, Y., Gou, L., Zhang, W., & Yang, H. (2020). DySAT : Deep Neural Representation Learning on Dynamic Graphs via Self-Attention Networks. *Proceedings of the 13th*

International Conference on Web Search and Data Mining, 519-527. https://doi.org/10.1145/3336191.3371845

Satyanarayanan, M., Bahl, P., Caceres, R., & Davies, N. (2009). The Case for VM-Based Cloudlets in Mobile Computing. *IEEE Pervasive Computing*, *8*(4), 14-23. IEEE Pervasive Computing. https://doi.org/10.1109/MPRV.2009.82

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, *20*(1), 61-80. IEEE Transactions on Neural Networks. https://doi.org/10.1109/TNN.2008.2005605

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling Relational Data with Graph Convolutional Networks. *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings*, 593-607. https://doi.org/10.1007/978-3-319-93417-4_38

Schmidhuber, J. (1992). Learning Complex, Extended Sequences Using the Principle of History Compression. *Neural Computation*, *4*(2), 234-242. Neural Computation. https://doi.org/10.1162/neco.1992.4.2.234

Schmidhuber, J. (2015). Deep learning in neural networks : An overview. *Neural Networks*, *61*, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

Schmidt-Schauß, M., & Smolka, G. (1991). Attributive concept descriptions with complements. *Artificial Intelligence*, *48*(1), 1-26. https://doi.org/10.1016/0004-3702(91)90078-X

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In L. Màrquez, C. Callison-Burch, & J. Su (Éds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (p. 298-307). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1036

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N. R., Van de Velde, W., & Wielinga, B. J. (1999). *Knowledge Engineering and Management : The CommonKADS Methodology*. The MIT Press. https://doi.org/10.7551/mitpress/4073.001.0001

Schwab, K. (2017). *The Fourth Industrial Revolution*. Crown Business.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In K. Erk & N. A. Smith (Éds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 1715-1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Shani, U. (2017). Can ontologies prevent MBSE models from becoming obsolete? *2017 Annual IEEE International Systems Conference (SysCon)*, 1-8. https://doi.org/10.1109/SYSCON.2017.7934726

Sharma, P., Hamedifar, H., Brown, A., & Green, R. (2017, mai 1). *The Dawn of the New Age of the Industrial Internet and How it can Radically Transform the Offshore Oil and Gas Industry*. Offshore Technology Conference. https://doi.org/10.4043/27638-MS

Sheridan, T. B. (2016). Human–Robot Interaction : Status and Challenges. *Human Factors*, *58*(4), 525-532. https://doi.org/10.1177/0018720816644364

Sheth, A. P. (1999). Changing Focus on Interoperability in Information Systems:From System, Syntax, Structure to Semantics. In M. Goodchild, M. Egenhofer, R. Fegeas, & C. Kottman (Éds.), *Interoperating Geographic Information Systems* (p. 5-29). Springer US. https://doi.org/10.1007/978-1-4615-5189-8_2

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge Computing : Vision and Challenges. *IEEE Internet of Things Journal*, *3*(5), 637-646. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2016.2579198

Simperl, E. (2009). Reusing ontologies on the Semantic Web : A feasibility study. *Data & Knowledge Engineering*, *68*(10), 905-925. https://doi.org/10.1016/j.datak.2009.02.002

Simperl, E., & Luczak-Rösch, M. (2014). Collaborative ontology engineering : A survey. *The Knowledge Engineering Review*, *29*(1), 101-131. https://doi.org/10.1017/S0269888913000192

Singh, S., Shehab, E., Higgins, N., Fowler, K., Erkoyuncu, J. A., & Gadd, P. (2021). Towards Information Management Framework for Digital Twin in Aircraft Manufacturing. *Procedia CIRP*, *96*, 163-168. https://doi.org/10.1016/j.procir.2021.01.070

Singla, K., Bose, J., & Varshney, N. (2019). Word Embeddings for IoT Based on Device Activity Footprints. *Computación y Sistemas*, *23*(3). https://doi.org/10.13053/cys-23-3-3276

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet : A practical OWL-DL reasoner. *Web Semant.*, *5*(2), 51-53. https://doi.org/10.1016/j.websem.2007.03.004

Sisinni, E., Saifullah, A., Han, S., Jennehag, U., & Gidlund, M. (2018). Industrial Internet of Things : Challenges, Opportunities, and Directions. *IEEE Transactions on Industrial Informatics*, *14*(11), 4724-4734. IEEE Transactions on Industrial Informatics. https://doi.org/10.1109/TII.2018.2852491

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A.,

Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO Foundry : Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251-1255. https://doi.org/10.1038/nbt1346

Smith, B., & Ceusters, W. (2010). Ontological realism : A methodology for coordinated evolution of scientific ontologies. *Applied ontology*, *5*(3-4), 139-188. https://doi.org/10.3233/AO-2010-0079

Smith, L. N. (2017). Cyclical Learning Rates for Training Neural Networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 464-472. https://doi.org/10.1109/WACV.2017.58

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, *25*. https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., & Katz, S. (2004). Reengineering thesauri for new applications : The AGROVOC example. *Journal of Digital Information*, *4*.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929-1958.

Staab, S., & Studer, R. (Éds.). (2009). *Handbook on Ontologies* (2e éd.). Springer-Verlag. https://doi.org/10.1007/978-3-540-92673-3

Staab, S., Studer, R., Schnurr, H.-P., & Sure, Y. (2001). Knowledge processes and ontologies. *IEEE Intelligent Systems*, *16*(1), 26-34. IEEE Intelligent Systems. https://doi.org/10.1109/5254.912382

Stark, J. (2015). Product Lifecycle Management. In J. Stark (Éd.), *Product Lifecycle Management (Volume 1) : 21st Century Paradigm for Product Realisation* (p. 1-29). Springer International Publishing. https://doi.org/10.1007/978-3-319-17440-2_1

State of the IoT 2020 : 12 billion IoT connections, surpassing non-IoT for the first time. (2020, novembre 19). IoT Analytics. https://iot-analytics.com/state-of-the-iot-2020-12-billion-iot-connections-surpassing-non-iot-for-the-first-time/

Stoilos, G., Stamou, G., & Kollias, S. (2005). A String Metric for Ontology Alignment. In Y. Gil, E. Motta, V. R. Benjamins, & M. A. Musen (Éds.), *The Semantic Web – ISWC 2005* (p. 624-637). Springer. https://doi.org/10.1007/11574620_45

Stojanovic, L., Dinic, M., Stojanovic, N., & Stojadinovic, A. (2016). Big-data-driven anomaly detection in industry (4.0) : An approach and a case study. *2016 IEEE International Conference on Big Data (Big Data)*, 1647-1652. https://doi.org/10.1109/BigData.2016.7840777

Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and Policy Considerations for Deep Learning in NLP* (arXiv:1906.02243). arXiv. https://doi.org/10.48550/arXiv.1906.02243

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering : Principles and methods. *Data & Knowledge Engineering*, *25*(1), 161-197. https://doi.org/10.1016/S0169-023X(97)00056-6

Stumme, G., & Maedche, A. (2001). FCA-MERGE : Bottom-up merging of ontologies. *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 1*, 225-230.

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn Methodology for Ontology Engineering. In M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, & A. Gangemi (Éds.), *Ontology Engineering in a Networked World* (p. 9-34). Springer. https://doi.org/10.1007/978-3-642-24794-1_2

Subramaniam, T., Jalab, H. A., & Taqa, A. Y. (2010). Overview of textual anti-spam filtering techniques. *International Journal of the Physical Sciences*, *5*(12), Article 12.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *2017 IEEE International Conference on Computer Vision (ICCV)*, 843-852. https://doi.org/10.1109/ICCV.2017.97

Sun, Z., Hu, W., & Li, C. (2017). Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding. In C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, & J. Heflin (Éds.), *The Semantic Web – ISWC 2017* (p. 628-644). Springer International Publishing. https://doi.org/10.1007/978-3-319-68288-4_37

Sure, Y., Staab, S., & Studer, R. (2004). On-To-Knowledge Methodology (OTKM). In S. Staab & R. Studer (Éds.), *Handbook on Ontologies* (p. 117-132). Springer. https://doi.org/10.1007/978-3-540-24750-0_6

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. *Proceedings of the 30th International Conference on Machine Learning*, 1139-1147. https://proceedings.mlr.press/v28/sutskever13.html

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, *27*. https://papers.nips.cc/paper_files/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

Sutton, C., & McCallum, A. (2012). An Introduction to Conditional Random Fields. *Foundations and Trends® in Machine Learning*, *4*(4), 267-373. https://doi.org/10.1561/2200000013

Sutton, R. S., & Barto, A. (2018). *Reinforcement learning : An introduction* (Second edition). The MIT Press.

Systems Engineering Vision 2020. (2007). http://www.ccose.org/media/upload/SEVision2020_20071003_v2_03.pdf

Tai, L., Paolo, G., & Liu, M. (2017). Virtual-to-real deep reinforcement learning : Continuous control of mobile robots for mapless navigation. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 31-36. https://doi.org/10.1109/IROS.2017.8202134

Tang, D., Qin, B., & Liu, T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In L. Màrquez, C. Callison-Burch, & J. Su (Éds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (p. 1422-1432). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1167

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification : A review. In *Data Classification* (p. 37-64). CRC Press. https://doi.org/10.1201/b17320

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F. (2018). Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology*, *94*(9), 3563-3576. https://doi.org/10.1007/s00170-017-0233-1

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, *48*, 157-169. https://doi.org/10.1016/j.jmsy.2018.01.006

Tao, F., Qi, Q., Wang, L., & Nee, A. Y. C. (2019). Digital Twins and Cyber–Physical Systems toward Smart Manufacturing and Industry 4.0 : Correlation and Comparison. *Engineering*, *5*(4), 653-661. https://doi.org/10.1016/j.eng.2019.01.014

Tao, F., Sui, F., Liu, A., Qi, Q., Zhang, M., Song, B., Guo, Z., Lu, S. C.-Y., & Nee, A. Y. C. (2019). Digital twin-driven product design framework. *International Journal of Production Research*, *57*(12), 3935-3953. https://doi.org/10.1080/00207543.2018.1443229

Tao, F., Zhang, H., Liu, A., & Nee, A. (2019). Digital Twin in Industry : State-of-the-Art. *IEEE Transactions on Industrial Informatics*, *15*(4), 2405-2415. IEEE Transactions on Industrial Informatics. https://doi.org/10.1109/TII.2018.2873186

Tao, F., Zhang, M., & Nee, A. Y. C. (Éds.). (2019). *Digital Twin Driven Smart Manufacturing*. Academic Press. https://doi.org/10.1016/B978-0-12-817630-6.00013-8

The Internet of Things : Dr. John Barrett at TEDxCIT. (2012, octobre 5). [Enregistrement vidéo]. https://www.youtube.com/watch?v=QaTIt1C5R-M

Thoben, K.-D., Wiesner, S., Wuest, T., BIBA – Bremer Institut für Produktion und Logistik GmbH, the University of Bremen, Faculty of Production Engineering, University of Bremen, Bremen, Germany, & Industrial and Management Systems Engineering,. (2017). "Industrie 4.0" and Smart Manufacturing – A Review of Research Issues and Application Examples. *International Journal of Automation Technology*, *11*(1), 4-16. https://doi.org/10.20965/ijat.2017.p0004

Thor, A., Kirsten, T., & Rahm, E. (2007). *Instance-based matching of hierarchical ontologies*. Gesellschaft für Informatik e. V. http://dl.gi.de/handle/20.500.12116/31814

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. MIT Press.

Tieleman, T., & Hinton. (2012). Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, *4*(2), 26.

Tuegel, E. J., Ingraffea, A. R., Eason, T. G., & Spottswood, S. M. (2011). Reengineering Aircraft Structural Life Prediction Using a Digital Twin. *International Journal of Aerospace Engineering*, *2011*, e154798. https://doi.org/10.1155/2011/154798

Turney, P. D., & Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *J. Artif. Int. Res.*, *37*(1), 141-188.

Uhlemann, T. H.-J., Lehmann, C., & Steinhilper, R. (2017). The Digital Twin : Realizing the Cyber-Physical Production System for Industry 4.0. *Procedia CIRP*, *61*, 335-340. https://doi.org/10.1016/j.procir.2016.11.152

Uschold, M., & Gruninger, M. (1996). Ontologies : Principles, methods and applications. *The Knowledge Engineering Review*, *11*(2), 93-136. https://doi.org/10.1017/S0269888900007797

Uschold, M., & Gruninger, M. (2004). Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, *33*(4), 58-64. https://doi.org/10.1145/1041410.1041420

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(4), 337-339. https://doi.org/10.1016/j.dsx.2020.04.012

van Ruijven, L. (2015). Ontology for Systems Engineering as a base for MBSE. *INCOSE International Symposium*, *25*, 250-265. https://doi.org/10.1002/j.2334-5837.2015.00061.x

Vashishth, S., Sanyal, S., Nitin, V., & Talukdar, P. (2019, septembre 25). *Composition-based Multi-Relational Graph Convolutional Networks*. International Conference on Learning Representations. https://openreview.net/forum?id=BylA_C4tPr

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information*

Processing Systems, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vaughan, J. W. (2018). Making Better Use of the Crowd : How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research*, *18*(193), 1-46.

Venceslau, A., Andrade, R., Vidal, V., Nogueira, T., & Pequeno, V. (2019). *IoT Semantic Interoperability : A Systematic Mapping Study*. 535-544. https://doi.org/10.5220/0007732605350544

Verdouw, C. N., & Kruize, J. W. (2017). *Digital Twins In Farm Management : Illustrations From The Fiware Accelerators Smartagrifood And Fractals*. https://doi.org/10.5281/ZENODO.893662

Vermesan, O. (Éd.). (2013). *Internet of things : Converging technologies for smart environments and integrated ecosystems*. River Publishers.

Villeneuve, E., Merlo, C., Terrasson, G., Abi Akle, A., Masson, D. H., & Llaria, A. (2019). Décision dans les systèmes cyber-physiques et humains : Application à l'élevage de précision. *CIGI QUALITA 2019 - 13th International Conference CIGI QUALITA*, *Papier n°7 (9 p.)*. https://hal.science/hal-02268577

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked Denoising Autoencoders : Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.*, *11*, 3371-3408.

Vrandečić, D. (2009). Ontology Evaluation. In S. Staab & R. Studer (Éds.), *Handbook on Ontologies* (p. 293-313). Springer. https://doi.org/10.1007/978-3-540-92673-3_13

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Hübner, S. (2001, janvier 1). Ontology-Based Integration of Information—A Survey of Existing

Approaches. *Ontology-based integration of information --- a survey of existing approaches*.

Wagg, D., Worden, K., Barthorpe, R., & Gardner, P. (2020). Digital Twins : State-of-The-Art Future Directions for Modelling and Simulation in Engineering Dynamics Applications. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, *6*. https://doi.org/10.1115/1.4046739

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE : A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In T. Linzen, G. Chrupała, & A. Alishahi (Éds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (p. 353-355). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5446

Wang, H., Zhang, F., Zhang, M., Leskovec, J., Zhao, M., Li, W., & Wang, Z. (2019). Knowledge-aware Graph Neural Networks with Label Smoothness Regularization for Recommender Systems. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 968-977. https://doi.org/10.1145/3292500.3330836

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing : Methods and applications. *Journal of Manufacturing Systems*, *48*, 144-156. https://doi.org/10.1016/j.jmsy.2018.01.003

Wang, J., Ye, L., Gao, R. X., Li, C., & Zhang, L. (2019). Digital Twin for rotating machinery fault diagnosis in smart manufacturing. *International Journal of Production Research*, *57*(12), 3920-3934. https://doi.org/10.1080/00207543.2018.1552032

Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016). Towards smart factory for industry 4.0 : A self-organized multi-agent system with big data based feedback and coordination. *Computer Networks*, *101*, 158-168. https://doi.org/10.1016/j.comnet.2015.12.017

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a Few Examples : A Survey on Few-shot Learning. *ACM Comput. Surv.*, *53*(3), 63:1-63:34. https://doi.org/10.1145/3386252

Wang, Z., Lv, Q., Lan, X., & Zhang, Y. (2018). Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Éds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (p. 349-357). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1032

Wasserman, L. (2004). *All of Statistics : A Concise Course in Statistical Inference*. Springer. https://doi.org/10.1007/978-0-387-21736-9

Wegner, P. (1996). Interoperability. *ACM Computing Surveys*, *28*(1), 285-287. https://doi.org/10.1145/234313.234424

Weiser, M. (1991). The computer for the 21st century. *ACM SIGMOBILE Mobile Computing and Communications Review*, *3*(3), 3-11. https://doi.org/10.1145/329124.329126

Weiss, E., Chung, L.-T., & Nguyen, L. (2019). A MBSE Approach to Satellite Clock Time and Frequency Adjustment in Highly Elliptical Orbit. *MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM)*, 65-69. https://doi.org/10.1109/MILCOM47813.2019.9020875

Widrow, B., & Hoff, M. E. (1988). Adaptive switching circuits. In *Neurocomputing : Foundations of research* (p. 123-134). https://dl.acm.org/doi/abs/10.5555/65669.104390

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79-82.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer. https://doi.org/10.1007/978-3-642-29044-2

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*. John Wiley & Sons.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4-24. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2020.2978386

Wymore, A. W. (1993). *Model-Based Systems Engineering*. CRC Press. http://gen.lib.rus.ec/book/index.php?md5=caed8c1cf22c3bce80191614f5af0e9c

Xing, H., Hong, Y., Min, K., & Juan, Z. (2021). An MBSE Framework for Civil Aircraft Airborne System Development. In D. Krob, L. Li, J. Yao, H. Zhang, & X. Zhang (Éds.), *Complex Systems Design & Management* (p. 147-157). Springer International Publishing. https://doi.org/10.1007/978-3-030-73539-5_12

Xu, L. D., He, W., & Li, S. (2014). Internet of Things in Industries : A Survey. *IEEE Transactions on Industrial Informatics*, *10*(4), 2233-2243. IEEE Transactions on Industrial Informatics. https://doi.org/10.1109/TII.2014.2300753

Xu, L. D., Xu, E. L., & Li, L. (2018). Industry 4.0 : State of the art and future trends. *International Journal of Production Research*, *56*(8), 2941-2962. https://doi.org/10.1080/00207543.2018.1444806

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645-678. IEEE Transactions on Neural Networks. https://doi.org/10.1109/TNN.2005.845141

Yang, S., Guo, J., & Wei, R. (2017). Semantic interoperability with heterogeneous information systems on the internet through automatic tabular document exchange. *Information Systems*, *69*, 195-217. https://doi.org/10.1016/j.is.2016.10.010

Yao, H., Zhang, C., Wei, Y., Jiang, M., Wang, S., Huang, J., Chawla, N., & Li, Z. (2020). Graph Few-Shot Learning via Knowledge Transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), Article 04. https://doi.org/10.1609/aaai.v34i04.6142

Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic Word Embeddings for Evolving Semantic Discovery. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 673-681. https://doi.org/10.1145/3159652.3159703

Yasmeen, S. (2018). *Impact of Big Data in Internet of Things*. https://www.academia.edu/37911790/IMPACT_OF_BIG_DATA_IN_INTERNET_OF_THINGS_IoT

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., & Leskovec, J. (2018). Graph Convolutional Neural Networks for Web-Scale Recommender Systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 974-983. https://doi.org/10.1145/3219819.3219890

Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer : Generating Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems*, *32*. https://papers.nips.cc/paper_files/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*, *13*(3), 55-75. IEEE Computational Intelligence Magazine. https://doi.org/10.1109/MCI.2018.2840738

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Éds.), *Computer Vision – ECCV 2014* (p. 818-833). Springer International Publishing. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, *64*(3), 107-115. https://doi.org/10.1145/3446776

Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous Graph Neural Network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 793-803. https://doi.org/10.1145/3292500.3330961

Zhang, D., Suhara, Y., Li, J., Hulsebos, M., Demiralp, Ç., & Tan, W.-C. (2020). Sato : Contextual Semantic Type Detection in Tables. *arXiv:1911.06311 [cs]*. http://arxiv.org/abs/1911.06311

Zhang, Q., Yang, L., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Information Fusion*, *42*, 146-157. https://doi.org/10.1016/j.inffus.2017.10.006

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep Learning Based Recommender System : A Survey and New Perspectives. *ACM Comput. Surv.*, *52*(1), 5:1-5:38. https://doi.org/10.1145/3285029

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing : A survey. *ACM Transactions on Intelligent Systems and Technology*, *11*(3), 1-41. https://doi.org/10.1145/3374217

Zhang, Z., Cui, P., & Zhu, W. (2022). Deep Learning on Graphs : A Survey. *IEEE Trans. on Knowl. and Data Eng.*, *34*(1), 249-270. https://doi.org/10.1109/TKDE.2020.2981333

Zhao, H., Zarar, S., Tashev, I., & Lee, C.-H. (2018). *Convolutional-Recurrent Neural Networks for Speech Enhancement* (arXiv:1805.00579). arXiv. https://doi.org/10.48550/arXiv.1805.00579

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). Learning Gender-Neutral Word Embeddings. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Éds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (p. 4847-4853). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1521

Zheng, X., Wang, M., & Ordieres-Meré, J. (2018). Comparison of Data Preprocessing Approaches for Applying Deep Learning to Human Activity Recognition in the Context of Industry 4.0. *Sensors*, *18*(7), Article 7. https://doi.org/10.3390/s18072146

Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent Manufacturing in the Context of Industry 4.0 : A Review. *Engineering*, *3*(5), 616-630. https://doi.org/10.1016/J.ENG.2017.05.015

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks : A review of methods and applications. *AI Open*, *1*, 57-81. https://doi.org/10.1016/j.aiopen.2021.01.001

Zhu, X. (Jerry). (2005). *Semi-Supervised Learning Literature Survey* [Technical Report]. University of Wisconsin-Madison Department of Computer Sciences. https://minds.wisconsin.edu/handle/1793/60444

Zitnik, M., & Leskovec, J. (2017). Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, *33*(14), i190-i198. https://doi.org/10.1093/bioinformatics/btx252

Publications scientifiques :

Simon Bauer, Christophe Merlo, Zina Boussaada, Simon Rincé 2021. Étude et mise en œuvre d'un cadre méthodologique outillé pour l'interopérabilité sémantique des données produit *Journées sur l'Interopérabilité des Applications d'Entreprise* **, /**, https://hal.archives-ouvertes.fr/hal-03508732 - état : accepté

Simon Bauer, Christophe Merlo, Zina Boussaada 2023. ABox Enrichment Through Semantic Data Recognition in Industrial Internet of Things *International Conference on Semantic Computing* **, pp. 212-215**, https://doi.org/10.1109/ICSC56153.2023.00041 - état : paru

Simon Bauer, Christophe Merlo, Zina Boussaada 2023. Enrichissement d'ontologies par la reconnaissance sémantique des données pour l'Internet Industriel des Objets *CIGI Qualita MOSIM* **,**
/, https://oraprdnt.uqtr.uquebec.ca/pls/public/docs/GSC6979/O0005294085_CIGI_QUALITA_MOSIM_2023_paper_5313.pdf - état : accepté