Implémentation d'une interopérabilité fédérée supportée par la transformation automatisée à la volée de modèles de données hétérogènes : application aux problèmes d'appariement des schémas

Mustapha Labreche

Université Fédérale

Toulouse Midi-Pyrénées

THÈSE

IMT Mines Albi-Carmaux
École Mines-Télécom

en vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

délivré par
IMT – École Nationale Supérieure des Mines d'Albi-Carmaux

---

présentée et soutenue par

Mustapha LABRECHE

le 18/04/2023

Implémentation d'une interopérabilité fédérée supportée par la transformation automatisée à la volée de modèles de données hétérogènes : application aux problèmes d'appariement des schémas

---

École doctorale et discipline ou spécialité :
EDSYS : Génie Industriel et Informatique

Unité de recherche :
Centre Génie Industriel, IMT Mines Albi

Directeurs de thèse :
Xavier LORCA, Professeur, IMT Mines Albi
Aurélie MONTARNAL, Maître-Assistante, IMT Mines Albi

Autres membres du jury :
Anne LAURENT, Professeure, LIRMM, Université de Montpellier *(Rapporteure)*
Nicolas DACLIN, Maître-Assistant HDR, IMT Mines Alès *(Rapporteur)*
Jean-Pierre BOUREY, Professeur, École Centrale de Lille *(Président)*
Chihab HANACHI, Professeur, Université Toulouse Capitole *(Examinateur)*
Sébastien WEILL, Directeur Produits et Recherche & Développement, Forterro, Sylob *(Invité)*

Sommaire

Remerciements

Ce n'est pas par tradition que cette page figure au préambule de ce rapport, mais c'est plutôt un devoir moral qui m'incite à le faire et comme le proverbe le dit, *celui qui ne félicite pas ses bienfaiteurs n'a pas de bien*.

C'est avec grand plaisir que je réserve donc cette page pour exprimer ma profonde gratitude envers toutes les personnes qui m'ont soutenu tout au long de ces trois années et demie. Ce travail est devenu une réalité, grâce à vous !

Je suis tout d'abord extrêmement reconnaissant envers mon équipe d'encadrement, **Pr. Xavier Lorca** et **Dr. Aurélie Montarnal**. Je vous remercie pour tous vos conseils, votre soutien continu, votre accompagnement, vos efforts pour coordonner et concrétiser les idées abstraites que j'ai apportées. Merci d'avoir participé à ce travail, et à tout ce que vous m'avez appris.

Je tiens à exprimer mes sincères remerciements à l'entreprise Forterro Sylob, **Sébastien Weill**, **Jean-Pierre Adi**, **Didier Artau**, **Benoît Wambergue** et toutes les personnes avec qui j'ai pu travailler. Merci pour votre précieux accompagnement, et vos grands efforts pour coordonner ce travail de recherche, ainsi que le temps que vous avez consacré à assister à toutes les réunions.

Je remercie tous les autres membres du jury d'avoir accepté d'évaluer ce travail. Une mention spéciale va au président du Jury, **Pr. Jean-Pierre Bourey** pour avoir accepté de présider le jury de ma thèse, aux rapporteurs, **Pr. Anne Laurent** et **Dr. Nicolas Daclin** pour leur intérêt et à **Pr. Chihab Hanachi** d'avoir examiné cette thèse. Vos retours et vos commentaires constructifs ont été précieux pour moi.

Je tiens également à te remercier **Pr. Frédérick Benaben** pour ton précieux soutien depuis nos premiers échanges de mail y a trois ans et à tes conseils et retours sur ce vaste domaine de l'interopérabilité qui ont été particulièrement utiles pour la suite. Je te remercie par ailleurs **Sébastien Truptil** d'avoir participé au début de cette aventure et d'avoir fait en sorte que ce travail commence.

Mes remerciements vont à **Claude Laffore** et à **Marlène Boval** pour avoir organisé et coordonné toutes les tâches administratives et logistiques liées à ma thèse et ma soutenance.

Un grand merci à **Nafe, Rodolphe, Oussema, Romain, Cheick, Mehran, Samer, Ghassen, Sam, Walid, Ibrahim, Abdallah, Khaled, Elyes, Thibaut, Robin** et tous ceux que j'ai rencontrés au CGI pour leur soutien dans mon quotidien.

Je tiens à exprimer ma profonde gratitude envers mes **parents** qui m'ont soutenu tout au long de mes études et de mon parcours. Vos prières, votre soutien et vos encouragements.

Je garde toujours le meilleur pour la fin. Merci pour ton soutien inconditionnel pendant les moments difficiles. Merci d'avoir toujours cru en moi. Merci pour ta patience et ta disponibilité tous les jours depuis des années. Tu es un pilier, une force, un soutien et je suis reconnaissant de t'avoir ... **Asma**.

Cadre général de la thèse

Introduction générale

Les entreprises au travers de leurs politiques et démarches stratégiques de gestion gravissent plusieurs étapes, de la création à la transmission en passant par la croissance, le développement et la maturité. Ces étapes sont caractérisées par l'évolution de l'activité de l'entreprise qui fait face à divers défis qu'elle doit relever avec efficacité afin d'assurer sa pérennité et ainsi améliorer ses capacités de compétition et de collaboration. Pour garantir cette efficacité, les entreprises font appel à des ressources humaines, des ressources robotiques et intelligentes (machines et équipements) ainsi que des ressources technologiques. Alors que l'industrie 4.0 se concentre sur l'intégration des ressources robotiques, intelligentes et technologiques, l'industrie 5.0 appel aux ressources humaines afin de renforcer et de compléter les autres ressources pour un contrôle et une optimisation efficace des activités et des décisions de l'entreprise.

Ces ressources sont alors en communication continue, ouvrant des canaux d'échange d'information hétérogènes, d'interactions complexes et de circulation de flux entre ce qu'on appelle les systèmes d'information d'entreprise. Ces systèmes reçoivent, stockent, envoient, gèrent, comprennent, transforment et qualifient diverses informations, fonctions et processus (décisionnels, opérationnels et supports) présents à plusieurs niveaux de l'entreprise (BENABEN, 2012). Les stratégies de compétitivité reposent alors sur les performances, la fiabilité, la robustesse et la facilité d'utilisation des systèmes d'information à l'image du progiciel de gestion intégré (ERP pour Enterprise Resource Planning), le système de gestion des produits ou le système de gestion de la chaîne logistique.

Un ERP est un système d'information conçu pour intégrer et optimiser les processus et les transactions nécessaires dans l'organisation d'une entreprise. La société *Forterro France*, leader français de l'édition et de l'intégration de logiciels ERP pour les métiers de l'industrie, propose des solutions ERP complètes dédiées aux Startups, Très Petites Entreprises (TPE), Petites et Moyennes Entreprises (PME) et des Entreprises de Taille Intermédiaire (ETI). La société propose des solutions personnalisables depuis plus de 30 ans pour accompagner ses clients et répondre à leurs besoins spécifiques dans plusieurs secteurs d'activités tels l'aéronautique, l'automobile, l'agroalimentaire, et bien d'autres. La solution ERP intègre alors plusieurs modules dotés de fonctionnalités avancées pour la gestion des ventes, la gestion de production, la maintenance ou la finance qui doivent non seulement se connecter entre eux, mais aussi avec d'autres systèmes externes dans le but d'échanger, partager ou transférer des informations.

En rendant possibles des interactions plus complexes, la mise en relation des systèmes permet par conséquent de répondre plus efficacement aux besoins de communication et de

coordination des activités améliorant les prises de décisions. Ainsi, l'interopérabilité est un élément essentiel pour ces interactions et est généralement définie comme la capacité de (deux ou) plusieurs systèmes ou composants à échanger des informations et à utiliser les informations qui ont été échangées (GERACI, 1991). Cependant, avec le temps, l'interopérabilité s'est accompagnée d'un besoin d'adaptation à des environnements technologiques actuels et futurs et des exigences industrielles en cesse d'évolution et dont l'hétérogénéité augmente à mesure que de plus en plus de données sont impliquées.

Ces observations concrètes s'accompagnent d'une constatation dans la littérature de l'absence d'un cadre théorique bien défini de l'interopérabilité. Ceci rend la tâche de la construction et la mise en relation des systèmes délicate et nous amène à poser une première question :

1. *QR 1 : Comment définir un cadre d'interopérabilité ?*

L'émergence d'outils proposés ou développés dans des contextes et pour des objectifs différents ainsi que l'hétérogénéité des données causée par la diversité de leurs natures, types, sources et destinations sont des facteurs qui rendent aussi le processus d'établissement de l'interopérabilité difficile. Les solutions d'interopérabilité proposées dans la littérature varient dans leurs architectures sur trois types d'approche (unifiée, intégrée ou fédérée) et l'adoption d'une approche par rapport à une autre dépend du contexte et les caractéristiques des systèmes impliqués (FERNANDES et al., 2020). Ainsi, de nombreuses solutions proposées ces dernières années pour résoudre le problème de l'interopérabilité sont fondées sur des approches unifiées ou intégrées.

Compte tenu des changements et des évolutions technologiques et industriels continus, les exigences actuelles et futures d'interopérabilité poussent à réfléchir à des solutions plus dynamiques, se basant sur des mécanismes qui n'imposent pas de modifications complexes dans les architectures et les caractéristiques des systèmes. (D. CHEN, DOUMEINGTS et F. VERNADAT, 2008) sur la base de (*ISO 14258 :* 1998) considère l'interopérabilité fédérée comme l'approche qui n'impose aucun modèle ou format standard établi. L'interopérabilité est alors mise en œuvre à la volée : on doit dynamiquement s'adapter aux environnements changeants. Cette approche qui est toujours considérée comme un défi de recherche majeur (CHARALABIDIS et al., 2008) ; (ZACHAREWICZ et al., 2020), offre la possibilité de s'adapter et d'interpréter automatiquement les connaissances à la volée, facilitant ainsi l'échange, le partage et l'utilisation de données. On peut alors se poser une seconde question :

2. *QR2 : Comment mettre en œuvre l'interopérabilité fédérée ?*

Dans ce contexte, l'entreprise Forterro à travers sa marque Sylob finance le projet de recherche à l'origine de cette thèse qui s'inscrit dans le champ de l'optimisation et la facilitation de l'utilisation et le déploiement de ses solutions dans l'objectif de permettre l'interopérabilité entre différents systèmes. Ce projet est en collaboration avec le laboratoire de recherche public Centre de Génie Industriel (CGI) au sein de l'IMT Mines d'Albi qui apporte une expertise scientifique dans la science des données, l'ingénierie des modèles, l'ingénierie à base de connaissances et la recherche opérationnelle. En effet, l'interopérabilité est une des nombreuses briques dans le mur des axes de recherche opérée et étudiée par les chercheurs du Centre de Génie Industriel. Le projet Mediation Information System Engineering (MISE) est un de ces projets interne dont l'objectif est la conception d'un système d'information de médiation pour supporter la collaboration et répondre aux problèmes d'interopérabilité.

MISE 1.0 fournit une méthodologie pour développer une architecture collaborative qui offre une capacité d'interopérabilité aux partenaires en se basant sur la démarche de l'architecture dirigée par les modèles (MDA pour Model Driven Architecture). MDA se base sur des techniques de modélisation en décrivant trois modèles, chacun avec un niveau d'abstraction plus élevé que le suivant et des techniques de transformation de modèles permettant le passage d'un modèle à un autre :

— Le modèle métier indépendant de l'informatisation (CIM pour Computation Independent Model). CIM spécifie les besoins fonctionnels du système ainsi que son environnement. Il sert de référence pour décrire le rôle du système indépendamment des détails liés à son implémentation et surtout éliminer la brèche entre les experts du domaine et les experts de la conception et du développement.
— Le modèle indépendant de la plateforme (PIM pour Platform Independent Model). PIM est un modèle d'analyse et de conception et est utilisé pour répondre à la question suivante : comment le système fait-il ce qui a été défini dans CIM ? Le rôle du PIM est de donner une vision structurelle et dynamique du système et ses interactions avec d'autres systèmes et montre les détails spécifiques de manière indépendante de la plateforme.
— Le modèle spécifique à la plate-forme cible (PSM pour Platform Specific Model). PSM décrit l'implémentation d'une application sur une plateforme particulière, il est donc lié à une plateforme d'exécution, pour cela, il combine les spécifications dans le modèle PIM avec les détails qui décrivent comment le système utilise un type particulier de plateforme. À la différence d'un PIM, un PSM n'a de sens que pour un développeur ayant une connaissance approfondie de la plateforme considérée.

Les travaux de (TOUZI et al., 2009) définissent et formalisent les règles de transformation entre des modèles appartenant aux niveaux CIM et PIM pour enrichir l'aspect interopérabilité entre ces modèles en s'appuyant sur le fait que les utilisateurs fournissent les informations relatives au modèle CIM. Cette hypothèse est ensuite levée lors des travaux de (RAJSIRI et al., 2010) en étudiant l'extraction et la transformation automatique des connaissances pour construire le modèle CIM. Cette itération s'achève avec les travaux de (TRUPTIL et al., 2008) qui complète le processus de transformation du PIM en PSM en implémentant une ontologie collaborative issue d'un métamodèle décrivant la situation de collaboration. En somme, MISE 1.0 extrait et transforme des connaissances en utilisant des métamodèles spécifiques à un domaine et fournit donc un seul processus collaboratif. Ces connaissances sont ensuite enrichies par des informations techniques sur les applications et fonctions, pour finalement, construire un modèle de workflow prêt à être exécuté.

MISE 2.0 (MU et al., 2017) ; (MU, BÉNABEN et PINGAUD, 2015) vient renforcer et améliorer les processus développés dans MISE 1.0 en adoptant un seul métamodèle qui retranscrit une situation de collaboration (MU et al., 2011). MISE 2.0 rassemble la connaissance relative à la situation de collaboration sous forme de processus décisionnels, opérationnels et de support, puis la transfère dans une cartographie du processus collaboratif grâce au métamodèle assisté par des règles de transformation de modèle. Cette cartographie est réutilisée pour déployer le workflow collaboratif technique. Ce déploiement est soutenu par des mécanismes de réconciliation sémantique *à la volée* en utilisant les annotations sémantiques des activités métier et des services techniques (BENABEN et al., 2013) ; (BOISSEL-DALLIER et al., 2015).

La dernière itération MISE 3.0 apporte un levier d'automatisation et de dynamique par des améliorations dans le travail continu, la mesure des performances, la surveillance intelligente et le déploiement du MISE en nuage (BENABEN et al., 2012). L'extraction des connaissances et la mise à jour des modèles se fait en permanence grâce à une architecture orientée événements (BARTHE-DELANOË et al., 2014) et le choix des éléments du modèle collaboratif est assisté par un système d'aide à la décision en incluant les aspects non fonctionnels. Ainsi, MISE 3.0 combine la détection automatique et la détection humaine pour améliorer l'agilité du système collaboratif. Au cours de cette dernière itération, (RAMÈTE et al., 2012) étudie l'application du MISE à la gestion de crise routière. (MONTARNAL, 2015) étudie le développement d'une plateforme cloud et la caractérisation de la dynamique collaborative et l'amélioration des opportunités de collaboration entre des partenaires qui ne se connaissent pas. L'aspect dynamique est aussi étudié dans (BIDOUX, 2016) par la prise en compte des ressources disponibles. La plateforme cloud est soutenue par l'utilisation d'un outil de réconciliation syntaxique et sémantique entre des modèles et des métamodèles par la mise

en place d'une transformation de modèles automatisée (T. WANG, TRUPTIL et BENABEN, 2017). Ce dernier travail constitue la première étape à l'origine de cette thèse.

Contributions

C'est sous l'ombre de ces questions de recherche et dans ce contexte industriel et scientifique que s'inscrivent les travaux de cette thèse. L'objectif principal est de proposer une approche générique et adaptable visant à implémenter l'interopérabilité fédérée entre des bases de données afin d'assister et de faciliter l'échange de données dans les processus de migration. L'approche s'appuiera sur des techniques d'extraction et de représentation des données à la volée afin d'élever leur niveau d'abstraction, ensuite, grâce à des techniques de modélisation et d'optimisation, l'approche proposera des liens potentiels construisant ainsi l'interconnexion entre les bases de données.

Pour atteindre cet objectif, l'organisation générale du manuscrit est résumée dans la Figure 1. Le manuscrit de thèse se consacre d'abord dans le Chapitre 1 à **la proposition un cadre général d'interopérabilité**. En effet, la littérature sur l'interopérabilité est riche et dense, néanmoins, l'état de l'art montre que des ramifications sont apparues qui ne convergent plus et plusieurs points de débat qui seront détailler rendent la mise en place d'une interopérabilité difficile. Ainsi, les problèmes pratiques de la mise en place de l'interopérabilité sont liés à des manquements dans les concepts théoriques, fournir un cadre cohérent d'idées, de notions et de connaissances organisées et structurées est alors nécessaire et devrait donc enrichir les fondements scientifiques de l'interopérabilité (JARDIM-GONCALVES et al., 2013). Ensuite, la thèse s'orientera dans le Chapitre 2 vers **les approches d'interopérabilité fédérée, leurs définition et implémentations**. Dans la littérature, l'utilisation de modèles pivots reste prévalente, mais est insuffisante pour aborder l'ensemble du cadre d'interopérabilité. En confrontant ces approches aux besoins actuels et futurs et en s'appuyant sur le cadre général proposé, on notera que l'interopérabilité fédérée est une voie prometteuse, mais souvent laissée de côté. La thèse détaillera alors dans le Chapitre 3 **la mise en œuvre de l'approche fédérée** pour les problèmes d'appariement des schémas des bases de données. Cette approche sera basée sur une modélisation à la volée et la création de liens par la théorie des graphes et les modèles d'optimisation. L'implémentation de cette approche permettra de prendre en compte diverses dimensions liées aux problèmes d'appariement et établira ainsi l'interopérabilité des données qui est une déclinaison de l'interopérabilité en général aux niveaux des données. Enfin, des **tests de l'approche proposée** seront effectués sur des cas d'étude dans le Chapitre 4.

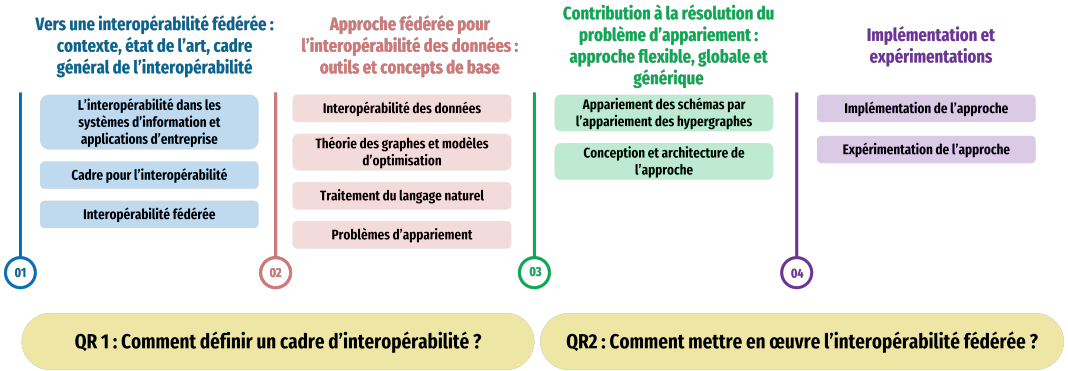


FIGURE 1 – Organisation générale du manuscrit de la thèse.

1

Vers une interopérabilité fédérée : contexte, état de l'art, cadre général pour l'interopérabilité

1.1 Introduction

L'interopérabilité est une notion essentielle et un concept multidimensionnel étudié dans le cadre de divers domaines d'application et notamment les systèmes d'information. Les besoins en termes d'interopérabilité sont devenus sans précédent et plusieurs aspects et concepts sont apparus et ont été développés pour gérer la complexité croissante liée à l'établissement de l'interopérabilité. Cependant, l'exploration de la littérature montre des fragmentations et plusieurs points de débat rendent les efforts morcelés. Sur la base de l'état

de l'art, l'ambition dans ce chapitre est de transposer ce constat et de proposer un cadre général qui décrit les différents aspects et étapes de l'interopérabilité dans la littérature et les rassembler de manière systématique en s'appuyant et couvrant les tentatives d'encapsulation déjà existantes. L'application de ce cadre pour les défis actuels et futurs accentue la nécessité d'avoir une approche capable d'absorber les changements environnementaux et technologiques et d'y répondre au mieux. Cette approche est l'approche fédérée qui devra répondre à certaines exigences. Cependant, on va montrer que l'implémentation actuelle de l'interopérabilité fédérée ne couvre pas toutes les exigences. En somme, la vocation de ce chapitre n'est pas de remettre en question l'état de l'art existant, mais plutôt de montrer les limites de la littérature à répondre aux besoins actuels en termes d'interopérabilité en général et d'interopérabilité fédérée.

L'organisation du chapitre sera comme suit, tout d'abord, la Section 1.2 introduira la notion des systèmes d'information, leur type, et leur rôle dans une organisation. Puis, la Section 1.3 décrira l'importance et la place de l'interopérabilité dans l'environnement des systèmes d'information. Ensuite, la Section 1.4 présentera l'étude de l'état de l'art sur l'interopérabilité et ses concepts et qui conclura sur un cadre général pour l'interopérabilité. La Section 1.6 présentera une étude de l'état de l'art sur l'interopérabilité fédérée et son apport face aux besoins actuels après une analyse des types de solution utilisés dans la Section 1.5. Enfin, la Section 1.7 conclura le chapitre.

1.2 Systèmes d'information

Pour introduire la notion de système d'information, il est important de définir les notions de systèmes et organisation.

1.2.1 Le système et l'organisation

Un système est défini comme étant « *un groupe d'éléments interagissant régulièrement ou un groupe d'éléments interdépendants formant un tout unifié* » (MERRIAM-WEBSTER, 2022) ou « *un groupe organisé ou connecté de choses* » (DICTIONARIES, 2022). L'ensemble d'éléments ou de composants interdépendants interagissent dans le but d'atteindre un objectif commun selon certains principes ou règles (L. JOHNSON, 2021). Un système possède une hiérarchie dont la description des niveaux inférieurs fournit des détails sur la façon dont le système fonctionne et atteint son objectif alors que la description des niveaux supérieurs montre le rôle du système dans son environnement (*ISO 15704 :* 2000). Un système est alors caractérisé par la nature de ses éléments, leurs rôles dans le système, leurs degrés d'interactions avec son environnement (système ouvert, fermé ou isolé) ainsi que sa frontière (critères d'appartenance d'un élément au système).

Une organisation (ou une entreprise) peut être considérée comme un système ou un ensemble de systèmes « *partageant une mission, des buts et des objectifs définis pour offrir un résultat tel qu'un produit ou un service* » (*ISO 15704 :* 2000). Ces systèmes sont dynamiques, intentionnels et densément connectés (CHECKLAND, 2000) et sont censés interagir de manière cohérente pour atteindre des objectifs communs. En général, une entreprise est structurée autour de trois niveaux (BENABEN, 2012) :

1. **Informations :** est un ensemble de faits bruts qui sont traités de sorte qu'ils donnent une valeur ajoutée pour l'entreprise. Ces faits couvrent des informations statiques ou dynamiques sur les clients, les produits ou les comptes de l'entreprise. À la base de ces faits bruts, on a la notion de **données** qui peuvent être de divers types (numérique, audio, images, etc.). Ces données sont stockées, organisées et structurées pour créer l'**information** et apporter une valeur supplémentaire au-delà de la valeur des données individuelles. En définissant des liens entre ces informations par la compréhension de

celles-ci et la prise en compte des moyens par lesquels ces informations peuvent être rendues utiles pour soutenir une tâche spécifique ou prendre une décision, on crée de la **connaissance**.

2. **Fonctions :** qui concernent les compétences et capacités pour accomplir des tâches et atteindre des objectifs. Une fonction peut donc être informatisée ou utiliser des matériels et ressources physiques ou un savoir humain.
3. **Processus :** consistent en un ensemble d'activités ou de fonctions, connexes et ordonnées qui aboutissent à un objectif à partir d'éléments en entrée. Ces processus peuvent être classés en trois catégories définies par (*ISO 9001 :* 2015) :
   - *Processus décisionnels* : qui sont caractérisés par la reconnaissance du problème, la recherche d'information, comparaison des choix, la prise et l'évaluation de la décision. Les processus décisionnels participent donc à l'élaboration des politiques et la gestion de l'organisation ;
   - *Processus opérationnels* : qui concernent les processus contribuant à la production d'un produit ou d'un service et apportant de la valeur à ce que l'organisation propose ;
   - *Processus supports* : qui impliquent des processus qui permettent à l'organisation de bien fonctionner, mais qui ne fournissent pas directement de valeur à ce que l'organisation propose.

En outre, deux notions complémentaires non négligeables, à savoir, la première notion, le *système d'information* qui couvre non seulement les différentes activités et tâches de l'entreprise, mais aussi intervient dans les fonctions et processus avec différents degrés d'implication de l'*humain* qui constitue donc la seconde notion complémentaire.

1.2.2 Les systèmes d'information dans l'organisation

En général, un système d'information (SI) « *fait référence à un système de personnes, de fichiers de données et d'activités qui traitent les données et les informations dans une organisation, et il inclut les processus manuels et automatisés de l'organisation* » (PAUL, 2010). Une autre définition plus précise identifie un système d'information comme « *une composante de deux systèmes, un système de traitement de l'information qui inclut les acteurs, les données et les processus et le système informatique qui inclut les ressources matérielles, les bases de données et les fonctions* » (BENABEN, 2012) ; (MORLEY et al., 2005). Il faut noter que plusieurs définitions et classifications des systèmes d'information ont été rapportées dans la littérature (DAVIS, 2000) ; (Y. DWIVEDI et al., 2009) ; (LAND, 1985) ; (LYYTINEN et NEWMAN, 2006) ; (PATHER, 2017) ; (SYMONS, 1991) et les champs de recherche relatifs sont de plus en plus larges. On retrouve par exemple : les SI et l'architecture de l'entreprise (KASEMSAP, 2018), la gestion stratégique des SI (BOONSTRA, 2013), la capacité et le contrôle des SI (CRAM, BROHMAN et R. B. GALLUPE, 2016), l'agilité des entreprises (LYYTINEN et ROSE, 2006), ou encore, les variables, les bases et les méthodes utilisées dans le développement des SI (HEVNER et CHATTERJEE, 2010) ; (SANCHEZ, TERLIZZI et al., 2017) ; (SUH, 2021). Il existe donc une variété d'interprétations des concepts et des définitions des SI (ALTER, 2008) suivant plusieurs angles de vue (BOELL et CECEZ-KECMANOVIC, 2015).

Les SI gèrent les flux d'informations d'une organisation et soutiennent non seulement la prise de décision, la coordination et le contrôle, mais aussi l'analyse des problèmes, la visualisation des sujets complexes, la création de nouveaux processus ou des produits. Étant donné qu'un SI associe des personnes, des ressources de données ou d'information, des logiciels, des réseaux de communications, des dispositifs physiques et du matériel, des politiques et des procédures, la combinaison de ces éléments interdépendants engendre une complexité et rend le développement d'un seul SI qui couvre l'organisation. De ce fait, plusieurs types de SI ont été développés (K. C. LAUDON et J. P. LAUDON, 2004) ; (O'BRIEN et MARAKAS, 2006) :

— **Système d'information pour dirigeants (Executive Information System (EIS) ou Executive Support System (ESS))** : ces systèmes sont développés et destinés aux dirigeants afin de les aider à prendre des décisions stratégiques, de rendre compte des situations et d'expliquer l'état général des activités de l'organisation en mettant à disposition des informations dans une version simplifiée (ou détaillée si nécessaire) et générées à partir de données agrégées et analysées provenant d'une variété de sources internes et externes.
— **Système d'information de gestion (Management Information System (MIS))** : ces systèmes soutiennent la prise de décisions commerciales destinées aux gestionnaires grâce à des rapports de gestion sur le suivi et la performance en résumant et agrégeant les activités de base de l'entreprise en utilisant des données fournies par d'autres systèmes internes. Ces systèmes ne sont généralement pas flexibles et n'ont peu ou pas de capacité d'analyse, mais fournissent des données qui contribuent à générer des informations par le biais de procédures prédéfinies.
— **Système d'aide à la décision (Decision Support Systems (DSS))** : les systèmes d'aide à la décision aident et assistent la prise de décision à plusieurs niveaux de l'organisation en se basant sur la modélisation, le calcul, et la sélection de la meilleure possibilité pour des problèmes mal structurés ou semi-structurés.
— **Système de traitement des transactions (Transaction Processing System (TPS))** : les systèmes de traitement des transactions sont des systèmes commerciaux de base qui se trouvent au plus bas de la hiérarchie organisationnelle et qui servent le niveau opérationnel de l'organisation. Ces systèmes exécutent, enregistrent, rassemblent, organisent et stockent les transactions nécessaires à la conduite des activités de l'entreprise et transfèrent ces données aux autres systèmes.

Ces systèmes dits *"traditionnels"* effectuent des opérations régulières, isolées et parfois fermées à la communication. Les organisations utilisent d'autres systèmes qui connectent alors les diverses activités et permettent de partager et coordonner les tâches à plusieurs niveaux. Ces applications ouvrent la possibilité que les différents types de systèmes d'une organisation fonctionnent ensemble comme un seul système d'entreprise (K. C. LAUDON et J. P. LAUDON, 2018).

1.2.3 Les applications d'entreprise

Les applications collectent, traitent, stockent, coordonnent et diffusent les données (informations ou connaissances), en vue de la prise de décision, du contrôle, de la coordination ou de l'analyse, grâce à un ensemble connexe de fonctions et de processus visant à améliorer les performances de l'organisation et sont souvent présentées en quatre grandes catégories :

1. **Progiciel de gestion intégré (Enterprise Resource Planning (ERP))** : ce sont des logiciels professionnels aux multiples fonctions, composés d'un ensemble de programmes paramétrables et destinés à être utilisés par une large clientèle pour intégrer divers processus et activités de l'entreprise dans un système unique. Les informations sont donc centralisées et partagées dans toute l'organisation. Ainsi, les dirigeants, cadres et gestionnaires sont en mesure d'utiliser les informations pour effectuer des tâches plus précises et plus rapides à court, moyen ou long terme. Cette application offre un avantage concurrentiel, opérationnel et stratégique. L'histoire des ERP commence dans les années 60 par le développement des systèmes de planification des besoins en composants (Materials Resources Planning ou Materials Requirements Planning (MRP)) afin d'évaluer les besoins en matériaux et gérer les stocks et la production (SIRIGINIDI, 2000), suivi ensuite dans les années 70 par la généralisation de l'idée et les premiers éditeurs de systèmes voient le jour (KLAUS, ROSEMANN et GABLE, 2000). Dans les années 80, une seconde génération de ces systèmes, à savoir le système de planification des ressources de production, voit le jour (Manufacturing Resources

Planning (MRP II)) en intégrant d'autres fonctions au processus de planification, ajoutant plus de contrôle des stocks et des processus de fabrication (RASHID, HOSSAIN et PATRICK, 2002). Dans les années 90, les éditeurs proposent une nouvelle génération d'outils appelée progiciel de gestion intégré (ERP) qui intègrent plus de fonctions internes à l'entreprise ainsi que la gestion des relations externes avec les fournisseurs et les clients, offrant ainsi une source de données unique pour toute l'organisation qui est une interface connectée à une base de données (BENTO et COSTA, 2013) ; (BENTO, COSTA et APARICIO, 2017) ; (BOERSMA et KINGMA, 2005). À partir des années 2000, les ERP II font leur apparition avec une connexion à internet et permettant un accès dématérialisé aux données, facilitant la mise en œuvre de telles applications aux petites et moyennes entreprises (MALHOTRA et TEMPONI, 2010). À partir des années 2010, les ERP traitent les données en temps réel et se familiariser avec les nouvelles technologies de l'information et de la communication [1] (APPANDAIRAJAN, N. Z. A. KHAN et MADIAJAGAN, 2012) ; (KIADEHI et MOHAMMADI, 2012). Aujourd'hui, à l'image de Forterro France, on se dirige vers l'adoption d'une plateforme ERP en nuage, vu les nombreux avantages qu'elle offre en matières de :

- réduction du capital car les dépenses en matériel sont réduites ;
- facilité d'édition et de développement de fonctionnalités avancées ;
- réduction des problèmes techniques dus aux processus de mise à jour ;
- accessibilité rapide et sécurisation des échanges.

2. **Gestion de la relation client (Customer Relationship Management (CRM))** : ce système est utilisé pour aider les organisations à gérer, améliorer et fidéliser de manière continue les relations clientèles en fournissant un ensemble de programmes qui coordonnent les processus adjacents aux clients, à savoir, la commercialisation, le marketing, le service après-vente ainsi que les tendances du marché (PEARLSON, SAUNDERS et GALLETTA, 2016). Ces systèmes permettent de développer les relations avec les clients en fournissant un meilleur service correspondant à leurs attentes et leurs besoins (CROTEAU et P. LI, 2003) en collectant et consolidant des données grâce à des programmes et technologies dédiés (MINAMI et DAWSON, 2008).
3. **Gestion de la chaîne logistique (Supply Chain Management (SCM))** : ces systèmes gèrent la chaîne logistique de l'organisation afin de faciliter les différents processus et tâches en relation avec leurs fournisseurs ou clients. L'objectif de ces systèmes est d'aider les parties prenantes dans la chaîne logistique, comme les fournisseurs, les clients, les distributeurs ou les transporteurs, à faire circuler les flux d'informations dans le but de conduire les processus qui amènent les différents produits et services de manière efficace et coordonnent les efforts de chaque partie de manière efficace et précise (PEARLSON, SAUNDERS et GALLETTA, 2016). Construire un tel réseau dans des environnements dynamiques est une tâche complexe et nécessite non seulement l'utilisation des nouvelles technologies de l'information et de la communication, mais aussi la prise en compte d'autres paramètres comme l'incertitude (LAURAS et al., 2021).
4. **Système de gestion des connaissances (Knowledge Management Systems (KMS))** : ces systèmes sont développés pour soutenir les activités et améliorer les performances d'une organisation en se basant sur la notion de connaissance qui peut être définie comme étant « *le fait de posséder la capacité requise dans une situation particulière pour traiter et résoudre des questions complexes de manière efficace* » (CHALMETA et GRANGEL, 2008). La gestion passe par la génération, la capture, la codification, le transfert et la mise à disposition des connaissances à tout moment et en tout lieu (K. C. LAUDON et J. P. LAUDON, 2018). Ces systèmes sont différents des autres types de systèmes d'information vu qu'ils permettent, en plus, de stocker la connaissance dans des documents ou des bases de données (ZAIM, MUHAMMED et TARIM, 2019), mais

1. Les technologies, méthodes, matériels, logiciels, réseaux et systèmes utilisés pour transmettre, stocker, traiter et sécuriser des informations et des données par des moyens électroniques.

aussi, permettent aux utilisateurs d'attribuer un sens, un contenu ou un contexte à la connaissance grâce à leurs expériences (B. GALLUPE, 2001).

D'autres applications d'entreprise existent aussi, à savoir :

— **Les systèmes de Gestion du cycle de vie du produit (Product Lifecycle Management (PLM))** : ces systèmes font référence au processus de transformation et de création automatisés de produit et qui comprennent les activités d'innovation, de développement, de gestion et de conception des produits en se basant sur un ensemble d'informations relatives au produit tout au long de son cycle de vie (SAAKSVUORI et IMMONEN, 2008).
— **Logiciel de pilotage de la production (Manufacturing Execution Systems (MES))** : ces systèmes fournissent une interface utilisateur qui permet une gestion des données commune pour offrir un suivi sur les processus en cours et optimiser d'autres activités de production à travers des communications bidirectionnelles (SAENZ DE UGARTE, ARTIBA et PELLERIN, 2009).

Chacune de ces applications d'entreprise intègre un ensemble connexe de fonctionnalités qui couvrent les niveaux de l'organisation ainsi que ces activités avec les clients, les fournisseurs et d'autres partenaires internes et externes. Ces applications sont construites sur un ensemble de processus métier et d'interactions connexes et utilisent des ressources et des technologies de l'information et de la communication. Ceci conduit alors à la multiplication et la diversité des environnements technologiques et des modèles de données et ce qui entraîne une complexité du point de vue de l'intégration des applications d'entreprise. Cette complexité a fait progressivement apparaître de nouveaux défis et champs de recherche autant théoriques que techniques. Dans la littérature, certaines études se sont orientées vers l'évaluation des succès et des échecs des SI et qui analysent les raisons pour lesquelles les SI répondent ou pas aux attentes et aux exigences des organisations (Y. K. DWIVEDI et al., 2015) ; (PETTER, DELONE et MCLEAN, 2012). L'objectif principal est de relever des défis et explorer de nouvelles perspectives et orientations de recherche afin de fournir de meilleurs supports à la conception et développement des SI en prenant en compte de multiples points de vue (technologique, sociale, socio-technique, etc.) (STRUIJK et al., 2022) ; (SUBAEKI et al., 2019) ; (N. WANG et al., 2016). Un de ces défis est l'interopérabilité (MEZGÁR et RAUSCHECKER, 2014) ; (RONOH, OMIENO et MUTUA, 2018) ; (SAKKA, 2012) qui est considérée comme un enjeu majeur favorisant le développement de diverses activités dans une organisation ou un réseau d'organisation et contribuant à pallier certaines lacunes des systèmes d'information et des applications d'entreprise qui nécessite une révision périodique face à la croissance rapide des technologies. La prochaine section détaillera donc la notion d'interopérabilité dans les systèmes d'information et les applications d'entreprise.

1.3 L'interopérabilité dans les systèmes d'information et applications d'entreprise

Les systèmes d'information jouent un rôle clé dans la gestion interne et externe d'une organisation. En effet, les organisations mènent leurs activités et processus dans un environnement dynamique et concurrentiel où ils doivent maintenir un niveau approprié d'excellence et de rendement afin de soutenir les prises de décision et atteindre leurs objectifs (EROL, SAUSER et MANSOURI, 2010). La donnée, l'information et la connaissance, qui proviennent de sources diverses et variées, notamment les systèmes d'information traditionnels (EIS, MIS, TPS, etc.) et les applications d'entreprise (ERP, CRM, SCM, etc.), sont donc au cœur des systèmes d'information qui fournissent un ensemble d'outils de gestion et de consolidation des flux d'information et de communication dont l'organisation a besoin pour fonctionner efficacement.

1.3.1 Besoin en interopérabilité

D'importantes mutations économiques, industrielles et technologiques continuent d'apparaître et d'évoluer sous l'ombre de l'industrie 4.0 (LASI et al., 2014) ; (L. D. XU, E. L. XU et L. LI, 2018) et récemment l'industrie 5.0 (X. XU et al., 2021).

L'**industrie 4.0** est un terme désignant la quatrième révolution industrielle qui révolutionne la façon dont les organisations fabriquent, améliorent et distribuent leurs produits ou services et qui s'appuie sur l'intégration de plusieurs technologies de l'information et de la communication se regroupant en quatre grandes catégories (CULOT et al., 2020) :

1. Les technologies d'interface physique-numérique qui englobent les technologies qui relient le monde physique au cyberespace et qui comprennent l'internet des objets (IoT pour Internet of Things) (OKANO, 2017), les systèmes cyberphysiques (CPS pour Cyber Physical Systems) (JAZDI, 2014) et les technologies de visualisation telle que la réalité virtuelle et augmentée (VR pour Virtual Reality et AR pour Augmented Reality) (DAMIANI et al., 2018) ou la réalité mixte (MR pour Mixed Reality) (BRUZZONE et al., 2019) ;
2. Les technologies de réseau qui offrent des fonctionnalités en ligne telle que l'informatique en nuage (Cloud Computing) (Y. LIU et X. XU, 2017) ou la chaîne de blocs (Blockchain) (BODKHE et al., 2020) ;
3. Les technologies de traitement des données pour le contrôle et la prise de décision comme l'intelligence artificielle (IA pour Artificial Intelligence) (JAVAID et al., 2022), l'analyse avancée des données (Data Analytics) (DUAN et DA XU, 2021), l'apprentissage automatique (ML pour Machine Learning) (ANGELOPOULOS et al., 2019) ou le jumeau numérique (Digital Twin) (PIRES et al., 2019) ;
4. Les technologies de processus numérique-physique qui regroupent des équipements ou du matériel connecté tels que l'impression 3D (3D Printing) (JANDYAL et al., 2022) ou la robotique avancée (Advanced Robotics) (GOEL et P. GUPTA, 2020).

Cette intégration a comme objectif principal d'interconnecter de façon *"intelligente"* le monde physique et virtuel en temps réel (SISODIA et JINDAL, 2021) et favoriser l'automatisation des processus en offrant une configuration flexible et un ajustement dynamique (DUAN et DA XU, 2021) pour se concentrer sur l'amélioration de la productivité et des performances (GOMES et al., 2020). L'industrie 4.0 est axée et dirigée dans un premier temps sur les technologies qui contribuent à la création, le stockage, la protection, à l'échange, le traitement, l'analyse et la visualisation des informations ou des données. Ensuite, dans un second temps, il est important de savoir tirer parti de ces informations et données pour créer des organisations intelligentes et autocontrôlées (RAUCH, 2020). Basée sur cette observation, l'industrie 5.0 est apparue pour aborder et traiter les principes d'équité sociale et de durabilité et ainsi compléter l'industrie 4.0 en fournissant un point de vue différent non axé exclusivement sur la numérisation et les technologies, mais sur l'importance de la recherche et de l'innovation pour soutenir l'industrie dans son service à long terme : vers une industrie centrée sur l'humain, durable et résiliente (BREQUE, DE NUL et PETRIDIS, 2021).

L'**industrie 5.0** n'est pas vouée à remplacer l'industrie 4.0 mais est une perspective pour encadrer la coexistence de l'industrie et les besoins sociétaux (L. D. XU, E. L. XU et L. LI, 2018). De ce fait, des technologies doivent donc être utilisées et améliorées pour s'aligner à ces évolutions. Plusieurs nouvelles technologies ont été identifiées pour soutenir ce changement (LENG et al., 2022) ; (MADDIKUNTA et al., 2022) :

— Les technologies d'interaction ou de collaboration homme-machine : comme la cobotique (Collaborative Robots) qui combinent l'innovation humaine et les capacités des machines (DOYLE-KENT, 2021).
— Les technologies de bio-inspiration et les matériaux intelligents qui permettent la production de matières premières à partir de déchets ainsi que l'intégration de matériaux vivants à l'image de la bionique (SACHSENMEIER, 2016).

— Les technologies de simulation et de modélisation (appelées aussi les métavers) représentent un système, un univers ou une situation dans laquelle le monde extérieur est construit et perçu par les utilisateurs (humains ou non humains) comme la réalité étendue (XR pour Extended Reality) qui utilise la réalité virtuelle, la réalité augmentée et réalité mixte (CÁRDENAS-ROBLEDO et al., 2022).
— Les technologies de gestion des données, d'intelligence artificielle et de communication : qui comprennent les technologies de transmission, de stockage, d'analyse, d'exploitation des données telles que l'informatique de périphérie (EC pour Edge Computing) (FRAGA-LAMAS, LOPES et FERNÁNDEZ-CARAMÉS, 2021), l'internet du tout (IoE pour Internet of Everything) (MURTUZA, 2022) ou les réseaux sans fil avancés (réseau 6G) (ZEB et al., 2022).

1.3.2 Analyse et conclusion

Les mutations citées précédemment favorisent et conjuguent les différentes relations organisationnelles (inter- et intra-) (GRILO et JARDIM-GONCALVES, 2010). En effet, les organisations doivent être de plus en plus compétitives, et de ce fait, elles repensent leurs stratégies d'évolution à travers non seulement la maîtrise des flux d'informations, mais aussi l'implication de l'humain pour faciliter l'échange de données, d'informations et de services (J.-Q. LI et al., 2017). Cette maîtrise passe à travers l'utilisation efficace et efficiente des technologies qui doivent d'un côté soutenir la mise en place de ces relations et d'un autre côté se faire en s'alignant aux objectifs et en assurant la cohérence avec les flux physiques de l'organisation et les exigences sociétales (PANETTO et MOLINA, 2008). Ainsi, le rôle des systèmes d'information et des applications d'entreprise se révèle encore plus important et central non seulement dans (1) la gestion interne et externe d'une organisation par l'établissement des relations organisationnelles, mais aussi (2) l'intégration des nouvelles technologies qu'ils doivent supporter. De plus, les données sont issues de différentes sources où l'incompatibilité culturelle, conceptuelle, organisationnelle, procédurale et technologique génère une hétérogénéité (D. CHEN et DOUMEINGTS, 2003a). En conséquence, elles sont analysées via ces technologies dans le but d'identifier des modèles et de développer des connaissances exploitables à mettre à la disposition de l'utilisateur (KRISHNAN, 2013) au travers de systèmes d'information. L'enjeu de la maîtrise des systèmes d'information est aussi une question considérable vu qu'elle aide les organisations à prospérer et à maintenir un rendement dans un environnement dynamique en étant capable de dissoudre en permanence l'incompatibilité.

Dans la pratique, les organisations utilisent différents systèmes d'information et applications d'entreprise basés sur différents modèles de données, normes, technologies et dispositifs pour gérer leurs opérations (POPPE et al., 2015). L'hétérogénéité affecte la performance et l'efficacité des relations et des interactions (MANSO et WACHOWICZ, 2009) ; (NAUDET et al., 2010). Par conséquent, la variété des entités susmentionnées, les technologies de l'information et de la communication, l'environnement dans lequel opèrent les organisations ou encore l'aspect humain et culturel produisent une complexité supplémentaire du point de vue de l'interopérabilité (KOTZÉ et NEAGA, 2010) ; (ZACHAREWICZ et al., 2017). Au final, trois situations d'interopérabilité peuvent alors être identifiées (BOZA et al., 2015) :

1. Le système d'information est un module gérant une fonction particulière d'une organisation, par exemple un domaine métier. Plusieurs modules peuvent donc exister et l'interopérabilité est produite en interne entre ces modules. Ces modules correspondent aux systèmes d'information traditionnels ;
2. L'interopérabilité se produit entre le système d'information ou l'application d'entreprise qui coexiste avec d'autres systèmes internes à l'organisation ;
3. Le système(s) d'information et/ou l'application(s) d'entreprise interagissent avec d'autres systèmes externes. L'interopérabilité se produit alors au sein d'un réseau de systèmes d'information.

Ces situations nous mènent à différencier deux notions souvent confondues. (1) l'intégration d'entreprise qui représente le fait de faire interagir des entités (applications, systèmes, machines, métier, etc.) dans le but d'atteindre des objectifs (F. VERNADAT, 1996). Cette notion reflète le sens de la coordination, la cohérence et l'uniformisation. (2) l'interopérabilité d'entreprise traduit le sens de coexistence, d'autonomie et d'environnement fédéré (D. CHEN, DOUMEINGTS et F. VERNADAT, 2008). Plus précisément, les entités intégrées sont capables de coopérer au sein d'un environnement homogène tandis que l'interopérabilité intervient quand on a des systèmes capables d'échanger et d'agir ensemble sous hétérogénéité et en autonomie (MEINADIER, 2003). (D. CHEN et DOUMEINGTS, 2003b) précisent que deux systèmes intégrés sont interopérables, mais deux systèmes interopérables ne sont pas nécessairement intégrés. La prochaine section présentera donc un état de l'art sur la notion d'interopérabilité et ses concepts basé sur la revue systématique de la littérature détaillée dans l'Annexe A.

1.4 Cadre pour l'interopérabilité

Le terme *interopérabilité* est souvent compris et interprété à l'aide de plusieurs notions comme la communication, l'échange, la coopération ou le partage. Le mot *interopérabilité* est dérivé du verbe *interopérer* qui découle du latin *inter* qui signifie *entre* et *operari* qui signifie *travailler*. L'interopérabilité, dans son sens le plus simple et le plus abstrait, est la capacité d'un ensemble d'entités à interopérer. À partir de ce postulat, l'interopérabilité est souvent décrite comme une propriété ou une capacité (ZDRAVKOVIĆ et al., 2017) qui permet de mettre un ensemble d'entités en relation (CARNEY, J. SMITH et PLACE, 2005). Autrement dit, l'interopération est un substantif qui désigne une ou plusieurs actions pour la réalisation pratique de l'interopérabilité (ÖHLUND, 2017).

Au fil du temps, l'interopérabilité s'est vue passer d'un simple besoin technique et technologique, où la principale occupation était les flux de données, à un sujet beaucoup plus large et vaste qui implique également des questions d'interaction humaine et institutionnelle (KOTZÉ et NEAGA, 2010) ; (ZACHAREWICZ et al., 2017). Les problèmes liés à l'interopérabilité relèvent autant de la culture que de la technologie et impliquent des individus, des organisations, des systèmes, etc. Les champs d'étude et la portée de l'interopérabilité s'est logiquement élargie pour être la pierre angulaire de plusieurs domaines d'application (le domaine militaire, l'informatique et sciences de l'ingénieur, gouvernement et services publiques, l'industrie et l'entreprise ou encore la santé), pour inclure plusieurs objectifs, types de relations et d'interaction, impliquer des personnes, des processus, des services et des organisations, ainsi que des technologies d'information et de communication.

Avec cette diversité, plusieurs aspects et concepts de l'interopérabilité sont apparus et ont été développés pour gérer la complexité croissante liée à sa réalisation. Ce constat a poussé certains auteurs à proposer des cadres théoriques pour encadrer les pratiques liées à l'interopérabilité et font donc référence à des tentatives d'encapsulation de l'état de l'art.

1.4.1 État de l'art : contributions à la théorie de l'interopérabilité

Dans le cadre de la feuille de route de la Commission Européenne pour développer l'interopérabilité des entreprises (CHARALABIDIS et al., 2008) ; (NAUDET et al., 2010) présentent une étude pour la construction d'une *base scientifique pour l'interopérabilité* en se basant sur le cadre d'interopérabilité des entreprises (D. CHEN et DACLIN, 2006) et l'ontologie de l'interopérabilité (NAUDET et al., 2006). Ils proposent une ontologie de l'interopérabilité [2]

2. Utilisée dans (GUÉDRIA et NAUDET, 2014) ; (WEICHHART, GUÉDRIA et NAUDET, 2016) ; (WEICHHART et NAUDET, 2014) ; (WEICHHART et STARY, 2015).

d'entreprise basée sur une approche systémique indépendante du domaine d'application. Cette ontologie fournit un cadre pour décrire les problèmes d'interopérabilité et les solutions à travers deux modèles, un modèle systémique qui décrit le système, ses éléments et ses relations, et un modèle décisionnel qui propose des solutions aux problèmes identifiés. (DOUMEINGTS, DUCQ et D. CHEN, 2009) ; (DUCQ, D. CHEN et DOUMEINGTS, 2012) étudient l'application de la théorie des systèmes aux systèmes de systèmes (composés de plusieurs systèmes d'information indépendants travaillant ensemble pour fournir de nouvelles capacités) afin de soutenir le développement de la base scientifique de l'interopérabilité durable des entreprises et répondre aux exigences associées. (NOF et al., 2008) ; (F. B. VERNADAT, 2009) présentent un cadre rassemblant les domaines, les normes, les technologies et les tendances futures liés aux modèles et architectures d'interopérabilité d'entreprises. (KOTZÉ et NEAGA, 2010) proposent un cadre conceptuel permettant d'étudier des composants à prendre en compte dans le développement d'un cadre global d'interopérabilité des entreprises pour des systèmes complexes qui se trouvent dans un réseau en identifiant les barrières non techniques et les solutions associées.

(CHARALABIDIS, R. J. GONÇALVES et POPPLEWELL, 2011) ; (MISSIKOFF, 2009) ; (POPPLEWELL, 2011) présentent les bases d'une *science de l'interopérabilité des entreprises* en tenant compte des concepts et des théories des sciences et des domaines scientifiques voisins, à l'image, des sciences sociales, sciences appliquées et les sciences formelles (mathématiques, systèmes et informatique). Dans la continuité, (LAMPATHAKI et al., 2012) renforcent cette idée en statuant sur le fait que, tout ancien ou nouveau domaine scientifique doit identifier ces relations (frontières, méthodologies partagées ou communes, conflits) avec les sciences ou disciplines scientifiques voisines. (JARDIM-GONCALVES et al., 2013) ; (PANETTO et al., 2016) proposent un ensemble de connaissances sur l'interopérabilité (IBoK pour Interoperability Body of Knowledge) en regroupant les travaux dans la littérature en modèles, théories (autrement dit, sciences voisines) et cadres d'interopérabilité. Les auteurs étudient la nécessité d'un nouveau cadre avec des modèles de référence, des méthodes formelles, une architecture standardisée, une plate-forme et des outils ouverts pour développer la prochaine génération de systèmes d'information d'entreprise. Dans le même contexte, (ZACHAREWICZ et al., 2017) mènent une étude de l'état de l'art en rassemblant plusieurs travaux sur le développement des systèmes d'information d'entreprise (TU, ZACHAREWICZ et D. CHEN, 2016) et statuent sur l'apport des approches basées sur les modèles à s'aligner sur les besoins industriels actuels et futurs et proposent un cadre conceptuel basé sur les modèles afin de surmonter les barrières qui empêchent l'établissement de l'interopérabilité. On retrouve ainsi l'interopérabilité dirigée par les modèles (MDI pour Model Driven Interoperability) basée sur l'utilisation des modèles et des transformations pour résoudre les problèmes d'interopérabilité des systèmes et applications d'entreprise en repartant de modèles de haut niveau d'abstraction jusqu'aux codes (BOUREY et al., 2007) ; (BOUREY et al., 2006).

(DIALLO, 2010) ; (DIALLO et al., 2011) fournissent des étapes vers une *théorie de l'interopérabilité* basée sur la construction d'un modèle de données à partir de la théorie des ensembles en définissant ses éléments, ses relations et ses dépendances. Les auteurs se concentrent sur l'interopérabilité des données, présentent et classent les différentes définitions de l'interopérabilité en deux catégories : les définitions qui considèrent l'interopérabilité comme une fonctionnalité inhérente à un système et les définitions qui considèrent l'interopérabilité uniquement pendant l'interaction. Les auteurs abordent également la question de la pratique de l'interopérabilité et identifient deux types d'utilisation, l'une basée sur des normes et des cadres et l'autre utilisant un modèle commun auquel les systèmes se conforment. La théorie des graphes est ensuite utilisée pour montrer la complexité de l'interopérabilité.

(KALB et al., 2013) mènent une étude basée sur la méthode de Delphes (OKOLI et PAWLOWSKI, 2004) qui est une série d'interrogations, au moyen de questionnaires, d'un groupe d'individus (experts) pour sonder et révéler les idées des experts dans ce domaine où les jugements sont d'intérêt. L'étude a pour objectif d'améliorer les efforts de recherche et de

développement en terme d'interopérabilité dans le domaine des bibliothèques numériques et identifie les limites, les besoins et les défis actuels et futures solutions de l'interopérabilité. (HENNING, 2018) propose un cadre théorique aidant à l'adoption des normes d'interopérabilité organisationnelle dans les réseaux d'information gouvernementaux. L'étude identifie les principaux facteurs dominants à prendre en compte pour établir une interopérabilité en les regroupant dans un cadre conceptuel comprenant : la gouvernance de l'interopérabilité, les caractéristiques du réseau, les résultats, les efforts d'adoption, les déterminants spécifiques à l'organisation, l'environnement réseau-externe et la caractérisation des normes d'interopérabilité.

(R. C. MOTTA, OLIVEIRA et TRAVASSOS, 2019) examinent l'interopérabilité dans le contexte des systèmes informatiques ubiquitaires. Les auteurs visent à caractériser l'interopérabilité en recueillant dans la littérature les différentes caractéristiques et définitions à travers une revue quasi systématique de la littérature analysée par la théorie ancrée. Les auteurs proposent une définition de l'interopérabilité qui reprend les concepts de : *"la capacité d'échange, la propriété du système, l'intégration, la coopération et la relation du système"*. Cette étude est menée pour guider la conception de systèmes ubiquitaires, leur évolution et propose un cadre théorique révélant des caractéristiques organisées en deux catégories :

1. Les caractéristiques structurelles qui permettent d'établir l'interopérabilité et qui comprennent :
   - l'*objectif* de l'établissement de l'interopérabilité ;
   - la *perspective*, à savoir, l'établissement de l'interopérabilité du point de vue des systèmes, organisations ou services ;
   - le *contexte* des informations ;
   - les *niveaux* pour lesquels l'interopérabilité est identifiée ;
   - les *attributs* liés aux systèmes.
2. Les caractéristiques comportementales qui permettent de mesurer, d'améliorer ou d'observer l'interopérabilité et qui comprennent :
   - la *méthode d'évaluation* de l'interopérabilité ;
   - les *défis* futurs ;
   - les *problèmes* signalés lorsqu'on aborde ou non l'interopérabilité ;
   - les *avantages* après l'établissement de l'interopérabilité.

Synthèse et analyse de l'état de l'art
Les cadres susmentionnés abordent des aspects théoriques, techniques et non techniques qui apparaissent parfois de manière disjointe et éparse. En effet, l'exploration de la littérature montre des ramifications dans les définitions, les concepts de l'interopérabilité et les cadres les organisant (HODAPP et HANELT, 2022). En outre, l'état de l'art actuel révèle plusieurs points de débat, l'importance de coordonner les efforts et les acteurs touche alors un horizon beaucoup plus large de sensibilisation et de compréhension (BAZZANELLA et TZITZIKAS, 2013).

De plus, les concepts sont d'une part proposés à un haut niveau d'abstraction théorique et d'autre part traitent de problèmes spécifiques liés séparément à l'interopérabilité et non de l'interopérabilité dans son ensemble. En fait, (ABUKWAIK et ROMBACH, 2017) ; (FOLMER et KRUKKERT, 2015) ; (GARLAPATI et BISWAS, 2012) ; (KALB et al., 2013) ; (KOTZÉ et NEAGA, 2010) ; (RILEY, 2020) ; (VALLE, GARCÉS et NAKAGAWA, 2021) abordent le sujet de l'interopérabilité dans la pratique et signalent l'absence d'architectures pour analyser, comprendre et guider la façon dont l'interopérabilité peut être traitée par les praticiens. En se basant sur des études de terrains sous différents angles, les auteurs proposent des processus qui peuvent aider et assister la mise en place de l'interopérabilité dans des domaines d'application spécifiques (compétition, administrations publiques, bibliothèque numérique, logiciels, etc.). Un cadre d'interopérabilité qui fournit un ensemble cohérent d'idées, de

notions et de connaissances organisées et structurées est alors nécessaire, enrichi par les fondements scientifiques de l'interopérabilité (JARDIM-GONCALVES et al., 2013) ; (PANETTO et al., 2016), tout en restant indépendant d'un domaine d'application (CHARALABIDIS, 2014). Ainsi, par la suite, nous allons présenter quatre principaux concepts de l'interopérabilité les plus étudiés dans la littérature.

1.4.2 Définitions de l'interopérabilité

La définition de l'interopérabilité a été une question récurrente au fil des ans. Alors que (FORD et al., 2007) répertorient trente-quatre définitions distinctes utilisées dans la recherche et la conception des normes, (GARLAPATI et BISWAS, 2012) présentent des définitions issues de la recherche académique et des praticiens de la santé. L'analyse documentaire soulève d'autres définitions apparues avec l'émergence de nouvelles technologies et disciplines.

L'interopérabilité est généralement définie par (IEEE, 1990) comme « *la capacité de deux ou plusieurs systèmes ou composants à échanger des informations et à utiliser les informations qui ont été échangées* » ou « *la capacité de deux systèmes à se comprendre et à utiliser les fonctionnalités de l'autre* » (D. CHEN, DOUMEINGTS et F. VERNADAT, 2008). Dans le cadre des applications d'entreprise, l'interopérabilité est « *la capacité d'un système ou d'un produit à fonctionner avec d'autres systèmes ou produits sans effort particulier de la part du client ou de l'utilisateur* » (KONSTANTAS et al., 2006) ou encore « *le niveau ultime de maturité collaborative (de l'organisation) adapté à l'intégration, qui peut être considéré comme le niveau ultime de collaboration (du réseau)* » (BÉNABEN et al., 2008b).

Dans le domaine des bibliothèques numériques où l'interopérabilité est un défi central (PAEPCKE et al., 1998), elle est définie comme « *la capacité d'une bibliothèque numérique à travailler en coopération avec d'autres bibliothèques numériques dans le but de fournir des services de meilleure qualité aux utilisateurs* » (SULEMAN, 2002). Les bibliothèques numériques visent à fournir un accès à une grande quantité d'informations numériques, y compris des textes, des images, des vidéos, de l'audio, et à gérer des composants provenant de différentes sources indépendantes.

Dans le domaine des réseaux d'énergie électrique, (GRIDWISE ARCHITECTURE COUNCIL, 2010) fait référence à l'interopérabilité par « *la connectivité transparente et de bout en bout du matériel et des logiciels, depuis les appareils jusqu'à la source d'énergie en passant par les systèmes de transmission et de distribution, améliorant ainsi la coordination des flux d'énergie avec des flux d'informations et d'analyses en temps réel* » ou encore « *la capacité de deux ou plusieurs réseaux, systèmes, dispositifs, applications ou composants à fonctionner ensemble, à échanger et à utiliser facilement des informations, de manière sûre, efficace et avec peu ou pas d'inconvénients pour l'utilisateur* » (CLEVELAND, SMALL et BRUNETTO, 2008). Outre le fait que l'interopérabilité facilite les flux de données en termes de collecte, transfert et analyse, elle participe à l'intégration et à l'interaction avec les autres parties du réseau (consommateurs, opérateurs et équipements) ainsi qu'à l'optimisation des coûts énergétiques.

Pour la communauté de la santé, (HEUBUSCH, 2006) fait référence à l'interopérabilité comme « *la capacité de différents systèmes informatiques et applications logicielles à communiquer, à échanger des données de manière précise, efficace et cohérente, et à utiliser les informations échangées* ». (JHA et al., 2008) associe l'interopérabilité à la notion de *d'échange d'informations sur la santé* (HIC pour Health Information Exchange) et définit comme « *l'échange ou le partage de données cliniques telles que les données médicales des patients, les notes des cliniciens ou d'autres informations médicales essentielles, d'une institution à l'autre* ». (STROETMANN et al., 2006) et (KIERKEGAARD, 2015) précisent que l'interopérabilité est « *la capacité d'échanger, de comprendre et d'agir sur les informations et les connaissances relatives aux patients et à d'autres aspects de la santé entre des cliniciens, des patients et d'autres acteurs linguistiquement et culturellement disparates, au sein d'une même juridiction ou entre plusieurs juridictions, dans un esprit de collaboration* ». La différenciation de l'interopérabilité interne qui traite des entités de la même organisation et l'interopérabilité externe qui implique des organisations externes est alors spécifiée (GAYNOR et al., 2014).

L'enjeu de l'interopérabilité reste très important en participant à l'amélioration de la qualité des services (MANTZANA, KOUMADITIS et THEMISTOCLEOUS, 2011), de l'aide à la décision clinique (BATES et SAMAL, 2018) et la sécurisation et la fiabilisation des flux d'informations (OYEYEMI et SCOTT, 2018).

Le cadre d'interopérabilité européen, présenté dans (EIF-PEGS, 2004) et discuté dans (MONDORF et WIMMER, 2016) ; (VAN OVEREEM, WITTERS et PERISTERAS, 2007) ; (F. B. VERNADAT, 2009), définit l'interopérabilité comme « *l'aptitude d'organisations*[3] *à interagir en vue de la réalisation d'objectifs communs mutuellement avantageux impliquant l'échange d'informations et de connaissances entre ces organisations via les processus métiers qu'elles prennent en charge, grâce à l'échange de données entre leurs systèmes informatiques* ». Au cours des deux dernières décennies, plusieurs travaux ont étudié l'interopérabilité dans le cadre d'initiatives gouvernementales et institutionnelles régionales, nationales et internationales impliquant des institutions et des organisations publiques, administratives et économiques (GUIJARRO, 2007) ; (GUIJARRO, 2009) ; (JIMENEZ, SOLANAS et FALCONE, 2014) ; (PERISTERAS, TARABANIS et GOUDOS, 2009) en tenant compte de plusieurs aspects de la gouvernance, à savoir, les aspects démocratiques, commerciales et gouvernementales (JAYASHREE et MARTHANDAN, 2010) pour le développement d'outils numériques (RAY, GULLA et DASH, 2007). Et à ce titre, l'interopérabilité joue un rôle très important (WIMMER, BONEVA et DI GIACOMO, 2018) puisqu'elle peut offrir plus d'efficacité, d'efficience et de réactivité (E. M. d. SANTOS et REINHARD, 2012).

Dans le cadre de l'*industrie 4.0* l'interopérabilité permet « *d'intégrer des services divers et distribués, des entreprises, des usines intelligentes, des dispositifs intelligents et des processus pour échanger des informations entre ces systèmes hétérogènes qui fonctionnent selon une grande variété de normes de communication* » (ENVIROTREC, 2020). Dans le cadre de l'*industrie 5.0* on définit l'interopérabilité comme « *la capacité de tous les composants, tels que les ressources humaines, les produits intelligents et toutes les technologies pertinentes, à se connecter, à communiquer et à fonctionner ensemble* » (PILLAI et al., 2021). Comme on l'a vu dans la Section 1.3.1, au regard des progrès technologiques, de multiples technologies ont émergé où l'interopérabilité reste une notion centrale (CULOT et al., 2020).

Dans le contexte de l'interopérabilité des entreprises et dans une perspective d'amélioration continue de leurs rendements pour soutenir les éventuelles opportunités de collaboration et de concurrence, les entreprises se tournent de plus en plus vers l'amélioration de leur système d'information aux moyens des technologies de l'information et de la communication. L'interopérabilité est, dans ce contexte, définie comme « *la capacité d'interaction (échange d'informations et de services) entre les systèmes de l'entreprise* » (D. CHEN, DOUMEINGTS et F. VERNADAT, 2008). L'interopérabilité des entreprises permet à (deux ou) plusieurs entités (interne ou externe) d'échanger ou de partager des informations et d'utiliser les fonctionnalités les unes des autres dans un environnement distribué et hétérogène (F. B. VERNADAT, 2010).

Synthèse des définitions de l'interopérabilité

À ce jour, comme nous l'avons vu, de multiples initiatives pour définir l'interopérabilité ont été proposées. Elles diffèrent d'un domaine d'application à un autre et au sein d'une même communauté. Dans certains cas, elles sont parfois associées ou complétées par les notions de compatibilité (P. ZHANG, PORTILLO et KEZUNOVIC, 2006), de portabilité (KAUR, SHARMA et KAHLON, 2017), de réutilisabilité et d'adaptabilité (D'AMBROGIO, GIANNI et IAZEOLLA, 2007), de flexibilité (CHARATSIS et al., 2005) ou encore d'intégration (D. CHEN, DOUMEINGTS et F. VERNADAT, 2008) ; (SCHOLL et KLISCHEWSKI, 2007) ; (F. VERNADAT, 2006). De plus, l'interopérabilité est un besoin pour de nombreux processus de gestion et traitements des données comme le partage des données (OTJACQUES, HITZELBERGER et FELTZ, 2007) ou encore la réutilisation des données (MEYSTRE et al., 2017). En somme, il n'existe pas de

3. Administration publique, institutions ou organes de l'Union européenne

définition commune et précise de l'interopérabilité. Il existe de nombreux points de vue sur ce qu'est l'interopérabilité. Néanmoins, même si les définitions sont souvent diverses et contextualisées, il y a plutôt une vue globale commune. Ainsi, on propose la définition globale suivante :

Définition 1.4.1 (Définition de l'interopérabilité)**.** Une capacité à créer des relations entre des entités par le biais d'actions prédéfinies pour un ou des objectifs dédiés dans un contexte en particulier :

- **Entités** : base de données, logiciels, systèmes, composants de systèmes, simulateurs, entreprise, organisation, etc.
- **Objectifs** : collaboration, amélioration, sécurisation, construction, etc.
- **Actions** : migrer, transférer, échanger, accéder, partager, connecter, interagir, utiliser, etc.
- **Contexte** : gouvernement, administrations, santé, réseau de collaboration, industrie, etc.

1.4.3 Établissement de l'interopérabilité

Au-delà de ces définitions, on peut se poser la question de savoir concrètement comment mettre en œuvre l'interopérabilité. Les entités impliquées dans le processus pouvant être hétérogènes, il est important de pouvoir détecter et identifier un certain nombre de caractéristiques. Ces caractéristiques sont essentielles, car la construction de l'interopérabilité entre entités est un problème multidimensionnel où différents types de défis sont rencontrés.

Comme nous l'avons vu dans la Section 1.4.2 concernant les définitions, il est également difficile d'avoir une compréhension commune de ces caractéristiques et de proposer une catégorisation puisqu'elles sont généralement spécifiques à un contexte (Lampathaki et al., 2012). De ce fait, plusieurs mots-clés peuvent être utilisés pour représenter ces caractéristiques, et nous trouvons, les niveaux (C4ISR et al., 1998) ; (Tolk et Muguira, 2003), les types (K. S. S. Santos, Pinheiro et Maciel, 2021), les aspects (Noran et Bernus, 2011), les dimensions (Shehzad et al., 2021), les vues (Chalmeta et Pazos, 2015), les conditions (Castelnovo et Simonetta, 2006), les obstacles (Noumeir, 2012) ou les questions (Rezaei, T. K. Chiew et S. P. Lee, 2014). Ces caractéristiques reflètent les problèmes d'incompatibilité qui peuvent bloquer la mise en place de l'interopérabilité comme un concept multidimensionnel.

Plusieurs travaux se sont alors orientés vers la structuration de cette notion à travers la mise en place de cadres, architectures, standards, normes ou modèles (Bazzanella et Tzitzikas, 2013) ; (Belchior et al., 2021) ; (Burns, Cosgrove et Doyle, 2019) ; (D. Chen, Doumeingts et F. Vernadat, 2008) ; (Deshmukh et al., 2021) ; (Eichelberg et al., 2005) ; (Elmhadhbi et al., 2020) ; (Iroju et al., 2013) ; (Kurilovas, 2009) ; (Lisboa et Soares, 2014) ; (Muller et al., 2019) ; (Mykkänen et Tuomainen, 2008) ; (Neiva et al., 2016) ; (Panetto, 2007) ; (Pardo, Nam et Burke, 2012) ; (Reegu, Daud et S. Alam, 2021) ; (Rezaei, T. K. Chiew et S. P. Lee, 2014) ; (Z. Zhang, C. Wu et Cheung, 2013), et nous citons par exemple :

- LISI pour Levels of Information Systems Interoperability (C4ISR et al., 1998),
- IDEAS pour Interoperability Development for Enterprise Application and Software (IDEAS, 2002-2003),
- AIF pour The ATHENA interoperability framework (ATHENA, 2004-2007) ; (A.-J. Berre et al., 2007) ; (D. Chen, Knothe et Zelm, 2004) ; (Guglielmina et A. Berre, 2005),

— INTEROP-NoE pour INTEROP Network of Excellence (BOURRIÈRES, 2006) ; (INTEROP, 2003-2007) ; (PANETTO, SCANNAPIECO et ZELM, 2004),
— EIF pour European Interoperability Framework (EIF-PEGS, 2004) ; (VAN OVEREEM, WITTERS et PERISTERAS, 2007),
— nEIF pour New European Interoperability Framework (KOUROUBALI et KATEHAKIS, 2019),
— LCIM pour The Levels of Conceptual Interoperability Model (TOLK et MUGUIRA, 2003),
— ICIF pour Inter-cloud interoperability framework (NODEHI et al., 2017),
— AAL pour Ambient Assisted Living (MEMON et al., 2014),
— RAMI 4.0 pour Reference Architectural Model Industrie 4.0 (ADOLPHS, 2015) ; (HANKEL et REXROTH, 2015),
— FIF pour Federated Interoperability Framework (TCHOFFA et al., 2021).

À l'instar des cadres théoriques présentés précédemment et le cadre qui sera détaillé dans ce chapitre, les cadres structurant les caractéristiques de l'interopérabilité identifient l'ensemble des types d'incompatibilité et les niveaux de l'interopérabilité (CHITUC, 2019). D'autres cadres se sont orientés vers l'ajout d'autres dimensions pour la proposition de solutions aux problèmes d'interopérabilité identifiés à l'image du cadre de l'interopérabilité d'entreprise.

Cadre de l'interopérabilité d'entreprise
(D. CHEN, DOUMEINGTS et F. VERNADAT, 2008) présentent une vue d'ensemble des architectures d'entreprise pertinentes en définissant et en clarifiant ses concepts de base pour les recherches passées sur l'intégration des entreprises et les avancées récentes sur l'interopérabilité des entreprises. De cette façon, les auteurs présentent le cadre de l'interopérabilité des entreprises illustré dans la Figure 1.1 qui est structuré en trois dimensions principales (D. CHEN, 2006) ; (D. CHEN, DOUMEINGTS et F. VERNADAT, 2008) [4] :

— **Dimension des barrières de l'interopérabilité** : représentent les problèmes qui bloquent le processus d'échange et de partage des informations. Ces obstacles doivent être identifiés et sont classés en trois grandes catégories :
  — **Barrières conceptuelles** : liées aux incompatibilités sémantiques et syntaxiques des informations échangées ou partagées, causées par la diversité des concepts qui peuvent être écrits et interprétés différemment d'un contexte à l'autre.
  — **Barrières technologiques** : concernant les problèmes liés à l'incompatibilité des plateformes, des architectures, des applications et des systèmes d'information qui gèrent les données.
  — **Barrières organisationnelles** : concernant les incompatibilités dans la structure de l'organisation et les techniques de gestion, à savoir la définition des rôles et la répartition des responsabilités.
— **Dimension des préoccupations** : il s'agit des points de vue à adopter au sein d'une organisation pour envisager et développer l'interopérabilité. Dans le cas où tous les niveaux sont atteints, on parle alors d'*interopérabilité complète* (MACIEL et al., 2017) mais qui n'est pas toujours faisable (BOUREY et al., 2007) ; (J. FERREIRA et al., 2011). Il existe quatre niveaux d'interopérabilité :
  — L'interopérabilité au **niveau de l'entreprise** : concerne la compréhension des aspects stratégiques, organisationnels et opérationnels de l'entreprise tels que les politiques décisionnelles, les visions, les législations, etc.
  — Interopérabilité au **niveau du processus** : où un processus se réfère à un système d'activités organisées qui, dans le cadre d'un réseau de collaboration, répondent à un besoin commun et qui sont décrites en utilisant différents langages.

4. Cadre élaboré sur la base des travaux (*ISO 14258 :* 1998), (EIF-PEGS, 2004). D'autres dimensions complémentaires de ce cadre sont présentés dans (D. CHEN, 2017).

  - L'interopérabilité au **niveau des services** : comprend les applications informatiques, les fonctions des sociétés et entreprises en réseau, qui doivent être identifiées, composées et exploitées ensemble, même si elles sont développées indépendamment.
  - L'interopérabilité au **niveau des données** : prend en compte l'échange, la compréhension et le traitement des informations contenues dans des structures de données hétérogènes, construites avec des langages différents et des règles de restriction.

- **Dimension des approches** : après avoir identifié les barrières et les niveaux auxquels ils peuvent être confrontés (l'espace de problèmes (ULLBERG, D. CHEN et P. JOHNSON, 2009)), il s'agit maintenant de définir les moyens de résoudre l'interopérabilité et d'assurer son développement, et à cette fin, il existe trois approches fondamentales pour atteindre ou établir l'interopérabilité :
  - **Approche intégrée** : les différentes parties utilisent un format commun pour les informations, accepté par tous et le processus d'intégration revient à utiliser le modèle partagé pour représenter tous les modèles.
  - **Approche unifiée** : les parties définissent un format commun et doivent ensuite traduire leurs données dans le modèle commun avant de les échanger. Elle est basée sur l'utilisation d'un métamodèle commun qui fournit un univers unifié pour une sémantique commune pour tous les concepts partagés par les différents modèles.
  - **Approche fédérée** : dans cette approche, il n'y a pas de format, de modèle, de langages ou de méthodes de travail communes. Pour établir l'interopérabilité, les parties doivent s'adapter à la volée et aucune partie n'impose un modèle ou une méthode de travail.

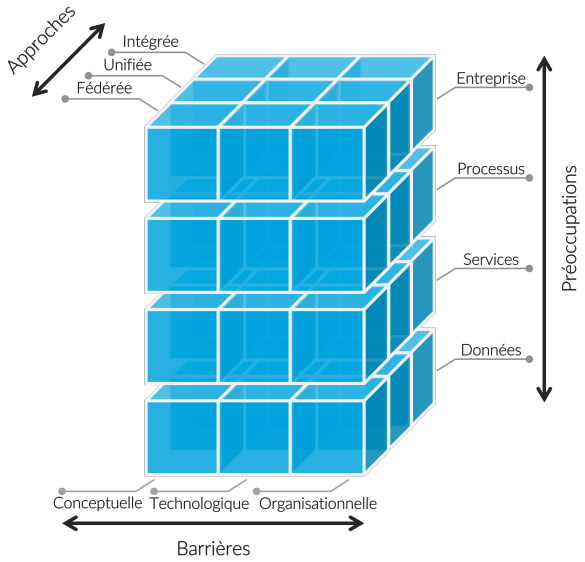


FIGURE 1.1 – Cadre de l'interopérabilité d'entreprise, (D. CHEN, DOUMEINGTS et F. VERNADAT, 2008).

1.4.4 Exigences de l'interopérabilité

Dans la continuité, une autre notion est étudiée, à savoir, les exigences d'interopérabilité, qui traduisent des besoins relatifs aux entités en termes d'interopérabilité.

Les exigences, également appelées propriétés (ROQUE et CHAPURLAT, 2009), attributs (PERAFÁN, CORREA et BUITRON, 2020) ; (ZDRAVKOVIĆ et al., 2015), caractéristiques (FERNANDES et al., 2020) ou encore capacités (ALMEIDA PRADO CESTARI et al., 2020), sont définies de manière générale comme « *une déclaration qui identifie une caractéristique ou une contrainte du système, du produit ou du processus, qui est sans ambiguïté, qui peut être vérifiée et qui est jugé nécessaire pour l'acceptabilité par les parties prenantes* » (HASKINS et al., 2006).

Cadre des exigences d'interopérabilité

(DACLIN et al., 2016) présentent le cadre des exigences d'interopérabilité [5] qui est une feuille de route pour définir, structurer et vérifier les exigences d'interopérabilité pour le processus d'interopérabilité.

Le cadre est présenté dans la Figure 1.2 et a été organisé sur la base des barrières et des préoccupations liées à l'interopérabilité. Une troisième dimension, celle des exigences, est donc ajoutée (DACLIN et MALLEK, 2014) :

- **Dimension des exigences d'interopérabilité** : l'exigence d'interopérabilité est définie par (DACLIN et al., 2016) comme « *une déclaration qui précise une fonction, une capacité ou une caractéristique, liée à l'aptitude d'un partenaire à assurer son partenariat en termes de compatibilité, d'interopérabilité, d'autonomie et de réversibilité, qu'il doit satisfaire* » :
  - Une **exigence de compatibilité** : est définie comme « *une déclaration qui spécifie une fonction, une capacité ou une caractéristique, considérée comme invariable tout au long de la collaboration et liée aux barrières de l'interopérabilité pour chaque niveau d'interopérabilité, et que les partenaires doivent satisfaire pour que la collaboration soit effective* ».
  - Une **exigence d'interopération** : est définie comme « *une déclaration qui spécifie une fonction, une capacité ou une caractéristique, considérée comme variable au cours de la collaboration, liée à la réalisation de l'interaction, et que chaque partenaire doit satisfaire* ».
  - Une **exigence d'autonomie** : est définie comme « *une déclaration qui spécifie une fonction, une aptitude ou une caractéristique liée à la capacité des partenaires à assurer leur propre gouvernance et à maintenir leur propre capacité opérationnelle pendant la collaboration, et que chaque partenaire doit satisfaire* ».
  - Une **exigence de réversibilité** est définie comme « *une déclaration qui spécifie une fonction, une capacité ou une caractéristique, liée à la capacité d'un partenaire à revenir à son état initial (en termes de performance) après la collaboration, et que chaque partenaire doit satisfaire* ».

Grâce à l'ingénierie des exigences, les besoins en termes d'interopérabilité sont alors formulés en exigences. Ces exigences sont ensuite les critères que les entités doivent respecter, répondre ou vérifier pour être interopérables ou complètement interopérables (MACIEL et al., 2017) tout au long du processus d'interopérabilité. Cependant, la définition d'exigences génériques reste difficile, car la nature des entités et le domaine d'application diffèrent d'un contexte à un autre, comme pour le Cloud Computing (PETCU, 2011), IoT (TAYUR et SUCHITHRA, 2017), informatique ubiquitaire (ROTH et al., 2018), les systèmes complexes (SZEJKA et al., 2014) ou encore systèmes à très large échelle (REZAEI et al., 2014b). Aussi, une relation peut lier les exigences entre elles (BOUKHARI, BELLATRECHE et JEAN, 2012), dans ce sens (G. S. LEAL, GUÉDRIA et PANETTO, 2020) proposent une approche afin d'identifier et formaliser les exigences et leur indépendance.

5. D'autres études antérieures sont présentées dans (CHAPURLAT et ROQUE, 2009) ; (DACLIN et MALLEK, 2014) ; (MALLEK, DACLIN et CHAPURLAT, 2011) ; (MALLEK, DACLIN et CHAPURLAT, 2012) ; (MALLEK et al., 2015)

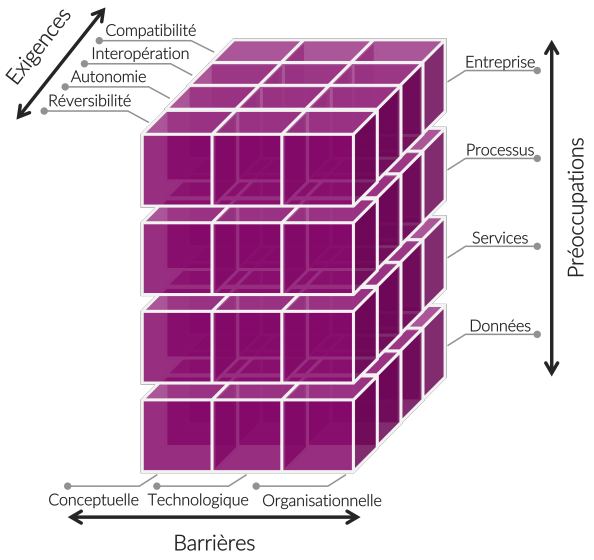


FIGURE 1.2 – Cadre des exigences d'interopérabilité, (DACLIN et al., 2016).

1.4.5 Évaluation et Amélioration de l'interopérabilité

Parallèlement aux exigences, les entités doivent pouvoir mesurer leur capacité à se comprendre (BENSON et GRIEVE, 2021) et à interopérer à travers des démarches d'évaluation et d'amélioration, et ce, afin d'amener les actions contribuant à améliorer le processus. Ainsi, après avoir défini l'interopérabilité, identifié ces caractéristiques, exigences et le type de stratégie pour l'établir, et dans un souci de durabilité, il faut pouvoir garantir une amélioration continue du processus en évaluant en plus l'interopérabilité, par exemple en termes de coût, de qualité ou de durée. Pour ce faire, plusieurs approches ont été proposées dans la littérature (BARUT, FAISST et KANET, 2002) ; (BIANCHINI et al., 2006) ; (M. CAMARA, DUCQ et DUPAS, 2010) ; (D. CHEN, VALLESPIR et DACLIN, 2008) ; (DACLIN, D. CHEN et VALLESPIR, 2006) ; (DUCQ et D. CHEN, 2008) ; (FORD et al., 2007) ; (GUÉDRIA, D. CHEN et NAUDET, 2009) ; (GUÉDRIA, NAUDET et D. CHEN, 2015) ; (MARGARITI et al., 2022) et sont classées de manière générale selon le contexte d'utilisation en trois types (G. d. S. S. LEAL, GUÉDRIA et PANETTO, 2019) ; (G. S. S. LEAL et al., 2016) :

1. **Évaluation du potentiel** : également appelée maturité, offre la possibilité d'évaluer le potentiel d'une entité à interopérer avec d'éventuelles entités non connues à l'avance. Cette mesure permet donc d'évaluer la capacité d'une entité à surmonter les obstacles futurs ;
2. **Évaluation de la compatibilité** : qui permet d'évaluer la compatibilité ou l'incompatibilité des entités connues avant ou après le processus d'interopérabilité. Cette mesure permet d'identifier les obstacles et les conflits qui se produisent ou peuvent se produire pendant le processus d'interopérabilité ;
3. **Évaluation de la performance** : qui est réalisée au cours du processus d'interopérabilité et permet de mesurer des indicateurs en termes de coûts ou de temps ;

Une deuxième classification concerne les mécanismes de mesure utilisés, qui se divisent en deux types :

1. Les mécanismes **qualitatifs** qui servent à attribuer un niveau de maturité à un niveau d'interopérabilité et permettent d'évaluer la qualité de l'interopérabilité à travers, par exemple, des mesures de capacité ;

2. Les mécanismes **quantitatifs** mesurent les propriétés de l'interopérabilité par le biais de modèles numériques, par exemple, en termes de temps ou de qualité de conformité, de fiabilité ou de connectivité ;

Plusieurs travaux décrivant et comparant des modèles d'évaluation de l'interopérabilité ont été publiés (ABUKWAIK et ROMBACH, 2017) ; (ALMEIDA PRADO CESTARI et al., 2020) ; (GUÉDRIA, NAUDET et D. CHEN, 2008) ; (JABIN, DIMYADI et AMOR, 2019) ; (G. d. S. S. LEAL, GUÉDRIA et PANETTO, 2019) ; (MARGARITI et al., 2022) ; (MARGARITI et al., 2020) ; (REZAEI et al., 2014b) ; (REZAEI, T.-k. CHIEW et S.-p. LEE, 2013). Ces études se concentrent sur des approches proposées dans les secteurs de l'entreprise, de l'administration, du gouvernement ou des logiciels (NOGUEIRA et al., 2016). Cependant, nous pouvons remarquer une difficulté à appliquer les mêmes mesures d'évaluation précédemment observées (REZAEI et al., 2014a). Un tel examen dans des secteurs tels que le Cloud Computing ou l'IoT est alors un défi ouvert (BELCHIOR et al., 2021), (GÜRDÜR et ASPLUND, 2018), (HAILE et ALTMANN, 2018).

D'autre part, il est également important de noter que le concept d'évaluation est un domaine qui n'est pas encore totalement exploré et il n'existe que peu de solutions (KRAMER, 2021). Les approches proposées ont certaines limites et fournissent une évaluation moins précise de l'interopérabilité en utilisant une grande quantité de mesures (G. d. S. S. LEAL, GUÉDRIA et PANETTO, 2019). Une utilisation hybride des mécanismes tout au long du processus d'interopérabilité en employant des méthodes d'optimisation mathématique, de la théorie des probabilités et de la théorie des graphes peut être utile en tenant compte de la possibilité d'utiliser des mécanismes d'amélioration automatiques et/ou itératifs (FORD et al., 2007).

Enfin, ces approches permettent d'analyser la situation actuelle et de faire un diagnostic afin d'améliorer l'interopérabilité (DACLIN, D. CHEN et VALLESPIR, 2016). Ainsi, les mesures d'interopérabilité doivent capturer les changements et, dans le meilleur des cas, s'adapter automatiquement à l'environnement dans lequel elles se trouvent (KRAMER, 2021). Ainsi, l'amélioration de l'interopérabilité a été étudiée dans la littérature dans certains travaux (BÉNABEN et al., 2008b) ; (FERNANDES et al., 2020) ; (FORTINEAU, PAVIOT et LAMOURI, 2013) ; (KOTZÉ et NEAGA, 2010) ; (REZAEI et al., 2014b) ; (ZACHAREWICZ et al., 2017) en considérant son impact sur le processus d'interopérabilité (M. S. CAMARA, DUCQ et DUPAS, 2014) et l'optimisation des efforts nécessaire pour y parvenir (ROUEN, 2013).

1.4.6 Le cadre général pour les concepts de l'interopérabilité

L'exploration de la littérature a naturellement conduit à la formalisation d'un cadre général de l'interopérabilité, et converge finalement vers une compréhension globale de l'interopérabilité comme une initiative générique avec des étapes claires qui requièrent chacune une orientation appropriée pour les faire aboutir. Ce cadre permet, d'une part, d'avoir une vue d'ensemble des principaux concepts d'interopérabilité et, d'autre part, d'assurer l'interopérabilité de la meilleure façon possible. Il est important de souligner que les ouvrages positionnés sont choisis (1) car ils apportent différentes informations et ont inspiré la construction du cadre général compte tenu de leur caractère de revue et d'état de l'art, et (2) ne constituent pas une liste complète et ne remettent pas en cause d'autres articles cités dans ce document ou que l'on a eu le malheur d'oublier ou de ne pas pouvoir trouver, néanmoins, ils peuvent être facilement placés. La Figure 1.3 illustre ce cadre. L'objectif principal est de contribuer à définir ce qu'est l'interopérabilité et de tracer un chemin pour y parvenir. En résumé, ce cadre comprend deux bases, une base pratique et une base scientifique :

1. La **base pratique** donne une vision plus précise de l'interopérabilité en offrant un support aidant à l'identification et au choix des solutions à mettre en œuvre en s'appuyant sur des expériences et les directives que peuvent donner les praticiens ;

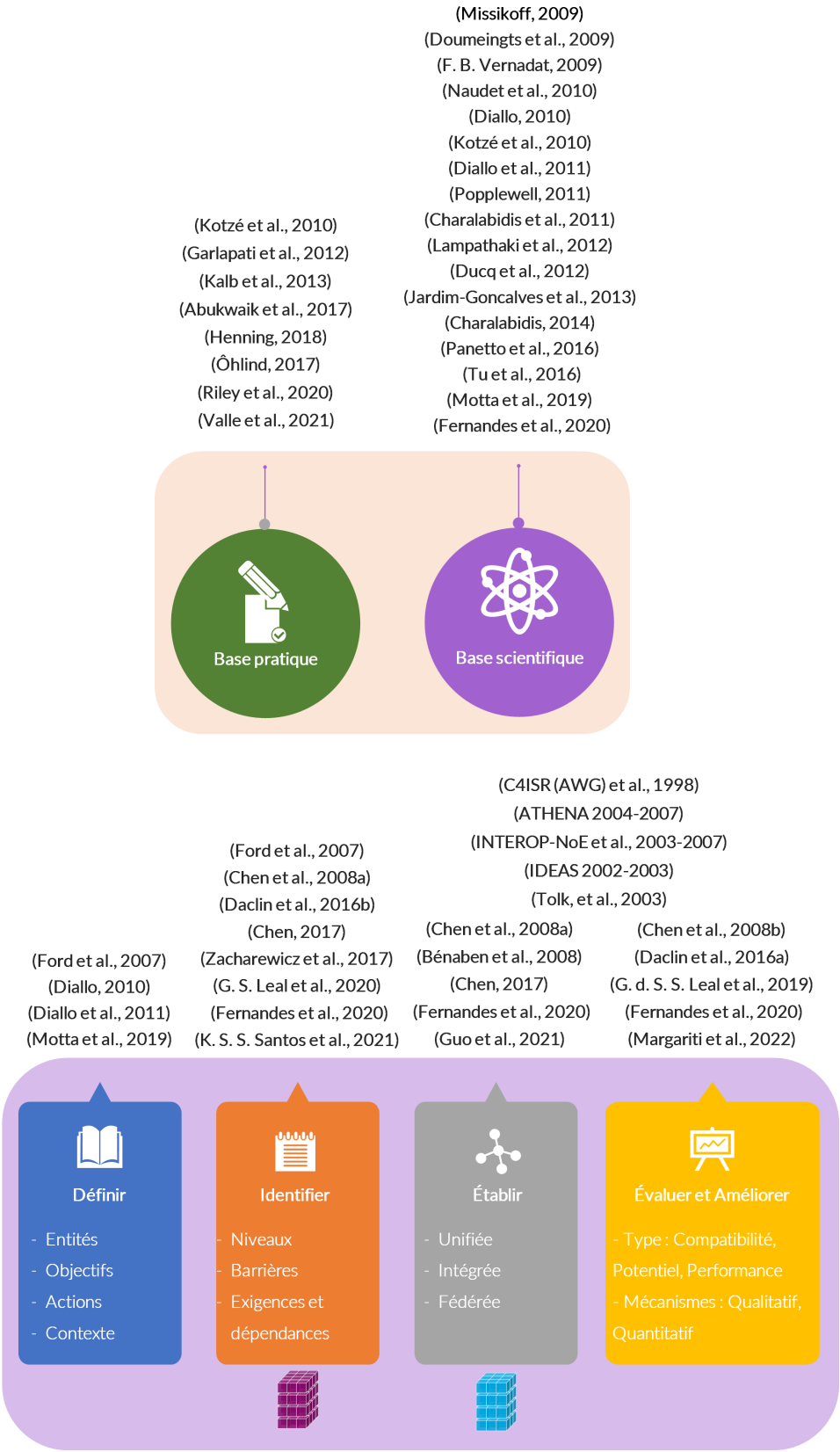


FIGURE 1.3 – Cadre général pour les concepts de l'interopérabilité.

2. La **base scientifique** contribue à donner les fondations théoriques et expérimentales pour l'identification et la compréhension des problématiques liées à l'interopérabilité puis, la conception, la réalisation, l'évaluation, la validation et la vérification des solutions proposées.

Ensuite, hors l'aspect théorique et pratique, plusieurs étapes et concepts doivent être pris en considération, de ce fait, nous proposons quatre étapes pour soutenir toute initiative de mise en œuvre ou d'amélioration de l'interopérabilité :

1. La **définition** du problème d'interopérabilité à traiter comprend les objectifs à atteindre, les entités impliquées, les actions à mener et le contexte général ;
2. L'**identification** des attentes d'une initiative d'interopérabilité en termes de niveaux ou l'interopérabilité doit être établie, les exigences et les différentes relations d'interdépendances, les obstacles ou problèmes d'incompatibilité ;
3. Le **type d'approches** à adopter pour résoudre le problème d'interopérabilité (unifiée, intégrée et/ou fédérée) ;
4. L'**évaluation** et l'**amélioration** de l'interopérabilité par l'utilisation de mesures qualitatives ou quantitatives et leurs analyses.

En somme, la Figure 1.3 établit l'état de l'art nécessaire qui justifie et documente chacune des étapes. Nous arguons que ce cadre peut être utilisé dans toute démarche d'interopérabilité et permet de manière rigoureuse de cadrer et répondre systématiquement de façon adéquate aux problématiques d'un projet d'interopérabilité en apportant un choix de plusieurs paradigmes basés, non seulement, sur des études scientifiques et des théories, mais aussi sur des expériences et expérimentations pratiques.

Ce cadre est également destiné à contribuer à la construction d'une théorie générale de l'interopérabilité en accord avec les infrastructures actuelles d'acquisition, de collecte et d'exploitation des données, que certains des auteurs mentionnés ci-dessus dans les bases scientifiques ont commencé à élaborer. Sans cette théorie, les points de débat autour de l'interopérabilité persisteront et il y aura autant de points de vue sur ce qu'est l'interopérabilité que sur la manière dont elle doit être réalisée dans des contextes spécifiques, et nous n'aurons pas de cadre stable. Ces points de débat sont dus au fait que l'interopérabilité est souvent résolue et étudiée au cas par cas, et aussi à l'énorme impact de l'interopérabilité, qui touche plusieurs domaines. En effet, les domaines où l'interopérabilité est un sujet d'étude sont divers, et les environnements industriels et technologiques évoluent et changent. La question d'interopérabilité et durabilité sont sujettes à études et sont deux aspects qui peuvent être liés (DASSISTI et al., 2013). Les solutions d'interopérabilité mises en œuvre doivent répondre à un cahier de charge couvrant non seulement les besoins actuels, mais aussi les besoins futurs (ZACHAREWICZ et al., 2017). Ainsi, la prochaine section va étudier le choix du type d'approche (unifiée, intégrée ou fédérée) à adapter pour mettre en œuvre une solution d'interopérabilité.

1.5 Le choix entre les approches de l'interopérabilité

(BÉNABEN et al., 2008b) a étudié le choix entre les trois approches dans le cadre de la mise en œuvre d'un système d'information de médiation supporté par une architecture orientée services qui gère les processus, les services et les données entre les entités impliquées dans un contexte de gestion de crise en utilisant l'approche unifiée. Les auteurs affirment que chacune des trois approches a sa situation optimale d'utilisation. (H. LIU, 2011) propose cinq critères de comparaison afin de distinguer les trois approches, à savoir, le champ d'application, la capacité d'adaptation aux changements, type de systèmes obtenus après interopération, nombre de connexions et nombre de traductions. L'auteur analyse les différences et les points communs entre ces approches dans le contexte des architectures des

systèmes d'entreprises et précise que l'utilisation de chaque approche dépend du stade de développement des systèmes d'information. Dans la même ligne, (FERNANDES et al., 2020) analysent comment ces approches affectent de différentes manières les caractéristiques des systèmes de systèmes d'information en comparant leurs impacts sur par exemple l'autonomie, l'interdépendance, la dynamique ou la connectivité et ajoute que le choix doit être fait selon les stratégies architecturales et les besoins commerciaux utilisés pour établir des liens d'interopérabilité entre les systèmes d'information en prenant en compte l'impact sur les caractéristiques. (H. GUO, Y. LIU et NAULT, 2021) proposent une étude qui analyse les trois approches appliquées aux problèmes d'interopérabilité des ressources dans la gestion des catastrophes et ceux afin d'évaluer l'impact économique résultant du choix entre les trois approches d'interopérabilité. (LEMRABET, 2012) étudient l'impact des niveaux de collaboration et les trois approches d'interopérabilité sur le développement des modèles et des plateformes collaboratifs. (KIOURTIS et al., 2017) proposent l'approche intégrée dans le contexte de la collaboration entre des entreprises dans le cas où il s'agit d'un nouveau réseau de collaboration et l'approche unifiée dans le cas où il est primordial d'utiliser les systèmes actuels, alors que (RUOKOLAINEN, 2009) proposent l'interopérabilité fédérée dans le cas d'une collaboration en fonction de ses caractéristiques à s'adapter au dynamisme et l'hétérogénéité du domaine.

(ELSHANI, WORTMANN et STAAB, 2020) souligne le besoin d'une interopérabilité fédérée basée sur des ontologies modulaires afin de partager une compréhension commune de l'information entre les disciplines. Dans la même ligne, (NORAN et ZDRAVKOVIĆ, 2014) statuent sur l'apport de l'interopérabilité fédérée dans la résilience. (ZDRAVKOVIĆ et al., 2015) discutent de l'interopérabilité en tant que propriété d'un système dans le contexte des organisations de gestion des catastrophes, en soulignant l'importance de la préservation de l'indépendance et de la résilience de ces organisations. (DESHMUKH et al., 2021) analysent les travaux proposés pour résoudre le problème de l'interopérabilité des plateformes de l'IoT qui suivent globalement l'une ou une combinaison des trois approches et soulignent l'apport de l'implémentation d'une interopérabilité fédérée flexible aux problèmes liés à ces plateformes hétérogènes opérantes dans des environnements dynamiques.

Au final, il est important de noter que le choix entre l'approche unifiée, intégrée ou fédérée dépend de leur efficacité à résoudre les problèmes d'interopérabilité en tenant compte des caractéristiques des entités participantes au processus et de leurs besoins ainsi que du contexte (par exemple, le temps disponible et ressources humaines et technologiques). Ainsi, en appliquant le cadre proposé, la prochaine section étudiera le type d'approche à adapter face aux défis actuels et futurs.

1.5.1 Application du cadre général pour l'interopérabilité aux défis actuels et futurs

Dans les défis actuels et futurs de l'IoT (DI MARTINO et al., 2018) et des domaines adjacents (Blockchains (BELCHIOR et al., 2021), CPS (GÜRDÜR et ASPLUND, 2018) ou le Cloud Computing (HAILE et ALTMANN, 2018)), l'interopérabilité est considérée comme l'un des défis centraux (AL-SAYED, HASSAN et OMARA, 2020). Ces paradigmes font référence à l'interconnexion entre des **entités hétérogènes** (des bases de données, des systèmes, des lieux, des environnements physiques, etc.) via des connexions réseaux et internet, en **recevant et transférant** des données. Ces formes de connexions permettent de rassembler de nouvelles masses de données et donc, de nouvelles connaissances et formes de savoirs. Ces données peuvent provenir de différentes sources issues de **plusieurs contextes**, à l'image du paradigme de la ville intelligente dont l'**objectif** est de créer des plateformes sur lesquelles des gouvernements, des entreprises et des citoyens peuvent communiquer et travailler ensemble (MEIJER et BOLÍVAR, 2016). Les interactions entre des entités hétérogènes sont l'une des principales caractéristiques de ces systèmes. Ainsi, les solutions déployées doivent

alors aborder plusieurs **niveaux** tels que l'interopérabilité technique qui englobe l'interopérabilité des appareils et des plateformes venant de domaines hétérogènes qui rendent aussi l'interopérabilité sémantique encore plus complexe. De plus, ces paradigmes sont sujets à des **changements technologiques** (sous l'ombre de l'industrie 5.0) et des **changements environnementaux** (une entité est amenée à quitter ou à rejoindre le cercle d'interaction). Cette complexité accroît l'hétérogénéité des données et accentue l'**incompatibilité** sémantique et syntaxique et ainsi, la difficulté d'avoir ou d'imposer un standard ou une norme commune pour toutes les entités.

Face à ces changements, les solutions d'interopérabilité unifiées ou intégrées ne sont pas toujours réalisables et efficaces en termes de coût ou de temps et où la négociation d'un modèle de données commun (intégration) ou un métamodèle commun (unification) semble être difficile à établir dû à la dynamicité de l'environnement qui est en développement et changement continu (DESHMUKH et al., 2021). En effet, l'intégration influence le fonctionnement individuel des entités en entravant par exemple leur **autonomie**. L'approche intégrée est un processus de fusion qui oblige les entités à se mettre d'accord sur un modèle unique afin de construire ou d'interpréter les informations actuelles et nouvelles en fonction de ce consensus. Cette approche assure l'homogénéité et la cohérence globale du processus. Tandis que l'unification réagit mal aux changements et à la dynamique de l'environnement, surtout si des concepts contradictoires sont présents ou des informations ne peuvent pas être prises en compte par le métamodèle. Un nouvel ajustement du modèle ou du métamodèle commun doit être effectué. L'approche unifiée, est mise en œuvre lorsqu'il s'agit d'élaborer un réseau (D. CHEN, 2017) où l'on établit un métamodèle pour lequel les autres participants au processus (et aussi les futurs participants), élaborent une mise en relation, en prenant en compte la possibilité d'une perte d'information puisque les entités ne sont pas représentées dans leur globalité. Cette approche est plus flexible que l'approche intégrée et il s'agit plutôt de trouver un compromis entre tous les participants. De plus, ces deux approches doivent aussi faire face à l'exigence de maintenir le processus d'interopérabilité avec d'autres entités (potentiellement **inconnus**) de manière autonome et en temps réel, chose qui n'est pas toujours évidente.

L'approche fédérée n'impose pas un modèle prédéfini. Plus on va vers l'approche fédérée, plus les entités sont interopérables, puisqu'elle facilite la communication (rapide) entre entités hétérogènes ainsi que la mise à jour dynamique de l'environnement (J. YOUSSEF, 2017) et préserve une sorte d'autonomie dans le fonctionnement indépendant des entités (FERNANDES et al., 2020). Ainsi, cette approche nous semble être la plus adéquate pour faire face à la dynamicité, la flexibilité et les changements environnementaux et technologiques où opèrent les entités. Elle doit permettre la résolution continue des problèmes d'interopérabilité (DUCQ, D. CHEN et DOUMEINGTS, 2012) et assurer au mieux la durabilité dans le temps (PALFREY et GASSER, 2012).

Au final, en parcourant l'ensemble du cadre proposé à travers ses bases et ses étapes, l'interopérabilité est approchée sous tous ces onglets et variantes. Ainsi, en confrontant ce cadre aux défis actuels et futurs, l'approche fédérée répond aux exigences des entités qui opèrent dans des environnements dynamiques et qui font face à de nombreux changements (AGOSTINHO et al., 2016) ; (KOTZÉ et NEAGA, 2010) ; (ROMERO et F. VERNADAT, 2016) ; (WEICHHART et STARY, 2017).

Cette approche est toujours considérée comme un défi de recherche majeur pour l'avenir (CHARALABIDIS et al., 2008) ; (ZACHAREWICZ et al., 2020). Dans la prochaine section, nous allons détailler cette approche, nous discuterons des solutions proposées adoptant une approche fédérée ainsi que l'importance de l'interopérabilité en général dans ce qu'on appelle les fédérations en s'appuyant sur la revue systématique de la littérature détaillée dans l'Annexe A.

1.6 Interopérabilité fédérée

Les approches fédérées apportent plus de flexibilité et offrent aux entités une autonomie qui leur permet d'être libres d'évoluer. Néanmoins, l'utilisation de ce type d'approche reste moins fréquente que les deux autres (l'intégration et l'unification).

1.6.1 État de l'art : approches d'interopérabilité fédérée

(ZACHAREWICZ, D. CHEN et VALLESPIR, 2009) présentent une approche fédérée pour l'interopérabilité d'applications d'entreprise. Cette approche traite trois aspects lors d'un processus d'interopérabilité : (1) l'aspect dynamique en termes de gestion du temps, (2) l'aspect d'interprétation des données et (3) la confidentialité des données. Pour ce faire, l'approche est basée sur l'utilisation du paradigme HLA (pour High Level Architecture) et de ce qu'on appelle *l'ontologie éphémère*. L'ontologie éphémère, comme son nom l'indique, est une ontologie qui peut être supprimée après utilisation. L'idée derrière cette ontologie est d'avoir l'interprétation des données à la demande et d'assurer une forme de confidentialité. De son côté, HLA permet en outre de créer un dialogue et un échange de données entre les entités (ZACHAREWICZ et al., 2009). L'objectif principal de HLA est de fournir une architecture ouverte offrant des services d'interopérabilité et de réutilisation des informations. Les participants sont appelés des fédérés. Les fédérés utilisent une infrastructure d'exécution commune (RTI pour Run-Time Infrastructure) pour communiquer. Des règles définissent le comportement requis des fédérés et des spécifications décrivent quels services peuvent être utilisés ou fournis. Enfin, le modèle objet (OMT pour Object Model Template) décrit le format de données impliquées par la définition de la sémantique et la syntaxe pour l'enregistrement des modèles, et pour lequel il y a une obligation de s'y conformer. D'autres utilisations du paradigme HLA pour soutenir l'interopérabilité fédérée peuvent être trouvés dans les travaux de (GORECKI et al., 2020) ; (TU, ZACHAREWICZ et D. CHEN, 2016) ; (J. R. YOUSSEF et al., 2016) ; (J. R. YOUSSEF et al., 2018a) ; (J. R. YOUSSEF et al., 2018b).

(USOV et al., 2010) présentent une approche pour la modélisation et la simulation d'infrastructures critiques[6]. Cette approche intègre un middleware DIESIS (Design of an Interoperable European Federate Simulation network for Critical InfrastructureS) qui permet de créer des groupes d'entités couplées de manière flexible. Les liens de couplage sont créés grâce à une base de connaissances basée sur des ontologies spécifiques au domaine (SINACI et ERTURKMEN, 2013). Ces ontologies permettent de modéliser des infrastructures hétérogènes et de représenter leurs dépendances. Contrairement à l'approche HLA qui se base sur un couplage central où les entités doivent adhérer à un format standard ou global pour l'échange de données en passant par l'infrastructure RTI, DIESIS considère un couplage latéral des groupes d'entités échangeant des informations. Concrètement, le couplage latéral crée des liens propres à un groupe de fédérées, mais qui peuvent être différents si on considère d'autres groupes. Ce type de couplage exige aussi la spécification des définitions syntaxiques et sémantiques (RAMAPANTULU, TEO et CHANG, 2017).

(WEICHHART, GUÉDRIA et NAUDET, 2016) étudient les bases et les principes pour le développement d'une infrastructure pour soutenir ce qu'on appelle une entreprise intelligente, sensible et durable. Cette infrastructure construit un réseau de systèmes d'information d'entreprise. Pour pouvoir prendre en charge les aspects distribués et dynamiques du réseau, les auteurs adoptent une approche fédérée pour l'interopérabilité en se basant sur un langage dédié (DSL pour Domain Specific Language) (NAUDET et al., 2010) et ajustable en fonction de la situation (WEICHHART et STARY, 2015). L'ontologie de l'interopérabilité des entreprises (OoEI pour The Ontology of Enterprise Interoperability) est adaptée avec des concepts liés à la théorie des systèmes complexes adaptatifs par le DSL pour modéliser les aspects dynamiques et est composée de deux types d'éléments, les concepts généraux de l'ontologie et les

6. Un actif qui est vital pour le fonctionnement d'une société ou de l'économie.

concepts des systèmes complexes adaptatifs et relié par des relations. L'aspect dynamique est alors capturé par des outils d'apprentissage sur des bases de connaissances.

(H. LIU, 2011) présentent une approche fédérée pour l'interopérabilité d'entreprise dans un contexte de collaboration en étudiant les dimensions conceptuelles et techniques. L'auteur présente un cadre général sur deux niveaux qui s'appuient sur deux ontologies. Un premier niveau où une méthode définit les propriétés de la collaboration, à savoir, tout ce qui concerne le processus général, les activités et les collaborateurs. Dans le second niveau, une architecture utilise l'Enterprise Service Bus (ESB) pour supporter les échanges et la gestion des informations. Une approche similaire est présentée dans (PIEST, IACOB et SINDEREN, 2020) qui proposent un concept d'architecture basée sur l'application de l'interopérabilité fédérée pour soutenir l'interopérabilité et la souveraineté des données dans les processus de partage et d'exploitation des données logistiques en temps réel dans un réseau de collaboration. L'architecture est composée de deux composants, un premier composant de connecteurs permettant à des entités hétérogènes d'échanger par le biais de connecteurs prédéfinis d'appariement et un deuxième composant de simulation permettant d'analyser et proposer des possibilités de collaboration. Cette architecture s'appuie sur plusieurs mécanismes, notamment, des ontologies, des mécanismes automatisés pour établir des appariements et les transformations associés (PIEST et al., 2020).

(DROUOT, GOLRA et CHAMPEAU, 2019) étudient l'interopérabilité des langages de la modélisation spécifique à un domaine (DSML pour Domain-Specific Modelling Language) pour analyser les cybermenaces en fédérant des données provenant de différents outils par la définition d'une sémantique partagée entre les différents langages qui les caractérisent. (FARIAS, ROXIN et NICOLLE, 2015) étudient l'intégration des données en proposant une architecture fédérée basée sur l'utilisation du langage de représentation des connaissances Web Ontology Language (OWL). (KLANN et al., 2016) présentent une plateforme i2b2 (Informatics for Integrating Biology and the Bedside) pour fédérer la conversion, l'échange et l'utilisation des données entre les réseaux médicaux moyennant une approche axée sur l'utilisation d'une ontologie commune. (ELSHANI, WORTMANN et STAAB, 2020) proposent un système fédéré basé sur l'utilisation des ontologies disciplinaires modulaires spécifique au modèle objet Bâtiments et Habitats (BHoM pour Buildings and Habitats object Model) (POINET, STEFANESCU et PAPADONIKOLAKI, 2020).

Synthèse et analyse de l'état de l'art

Les études présentées ci-dessus proposent des approches fédérées pour établir l'interopérabilité entre des entités hétérogènes à travers l'utilisation des ontologies et/ou des métamodèles. Sur la base des définitions de (GRUBER, 1993) et (BORST, 1999), (STUDER, BENJAMINS et FENSEL, 1998) considère l'ontologie comme « *une spécification formelle et explicite d'une conceptualisation commune. La conceptualisation désigne un modèle abstrait d'un phénomène identifiant les concepts qui lui sont pertinents. Formelle désigne le fait que l'ontologie doit être interprétable par une machine et explicite désigne le fait que les de concepts utilisés et leurs contraintes soient explicitement définis* ». Le métamodèle est présenté par (BÉZIVIN, 2005) comme un modèle qui définit un ensemble de concepts et de relations qui peuvent être utilisés pour créer d'autres modèles selon un point de vue donné. Ainsi, la structure d'une ontologie peut être décrite par un métamodèle et ces concepts peuvent être utilisés de façon complémentaire (BÉNABEN et al., 2008a) ; (LAURAS et al., 2014) ; (MONTARNAL, 2015). Ces concepts sont statiques ou dynamiques, c.-à-d. qui évoluent suivant les entités impliquées et les changements de leur environnement. Les approches présentées utilisent des concepts prédéfinis en définissant les principaux éléments et relations (DROUOT, GOLRA et CHAMPEAU, 2019) ; (KLANN et al., 2016) ou des concepts liés aux domaines (ELSHANI, WORTMANN et STAAB, 2020) ; (USOV et al., 2010) ; (WEICHHART, GUÉDRIA et NAUDET, 2016). D'autres approches mettent à jour ces concepts au fur et à mesure que des changements opèrent et qui sont non couverts (PIEST, IACOB et SINDEREN, 2020) ; (ZACHAREWICZ, D. CHEN et VALLESPIR, 2009) ou les créés à partir des informations récoltées à chaque changement (FARIAS, ROXIN et NICOLLE,

2015) ; (WEICHHART, GUÉDRIA et NAUDET, 2016). Ces concepts apportent donc une capacité d'interprétation et de compréhension des échanges en élaborant la syntaxe, la sémantique et les relations portées par leurs éléments. Le dynamisme de l'interopérabilité fédérée est donc soutenu par la capacité d'adaptation et de coordination des entités.

Par ailleurs, dans des environnements sujets à divers changements, les entités qui interopèrent créent ce qu'on appelle une *fédération*. Le terme de fédération est souvent utilisé pour décrire des entités hétérogènes et autonomes qui travaillent ensemble, comme celles présentées auparavant dans (ZACHAREWICZ et al., 2009). L'hétérogénéité constitue une source de problèmes qui réduit l'efficacité et les bénéfices de la fédération. En effet, la fédération implique parfois des entités conçues indépendamment et différentes en termes de matériel, de système d'exploitation, de langages, de schémas, de modèle de données, de processus, etc. Pour assurer l'interopérabilité entre ces entités, il faut identifier et résoudre les problèmes d'incompatibilité. L'interopérabilité apparaît donc, comme l'une des exigences les plus importantes. Dans la prochaine section, nous allons présenter deux types de fédérations qui opèrent dans des environnements dynamiques et nous allons voir l'apport que peut avoir une approche basée sur l'interopérabilité fédérée pour faire face aux changements.

1.6.2 L'interopérabilité et les fédérations

Fédération dans le Cloud Computing : dans le domaine du Cloud Computing, le Cloud Federation désigne une approche apparue au cours des deux dernières décennies (GROZEV et BUYYA, 2014) ; (TOOSI, CALHEIROS et BUYYA, 2014) où de multiples ressources, indépendantes, hétérogènes, privées et publiques, internes et/ou externes, provenant de fournisseurs de services sont partagés suivant une norme ouverte afin de fournir un environnement de calcul pour améliorer les services des uns et des autres (KURZE et al., 2011) ; (ROCHWERGER et al., 2009) et de surmonter certaines limites telles que la qualité des services et le manque d'interopérabilité (ASSIS et BITTENCOURT, 2016). Les problèmes d'interopérabilité ont été abordés dans la littérature en tenant compte des niveaux d'incompatibilité (données, application, matériel, réseau, ressources virtuelles, stockage et systèmes d'exploitation) (HAILE, ALTMANN et al., 2015) et de son impact (HAILE et ALTMANN, 2018). Des ontologies prédéfinies ou construites par la fusion d'autres ontologies sont utilisées pour structurer et annoter sémantiquement les données (AGARWAL et al., 2016) ; (DESHMUKH et al., 2021) ; (SERRANO et al., 2017) mais qui doivent être reconfigurées pour supporter de nouvelles connexions.

Fédération dans l'apprentissage : un autre concept de fédération qui a significativement émergé au cours des cinq dernières années est l'apprentissage fédéré (où Federated Learning) qui est un domaine de l'intelligence artificielle et l'apprentissage où l'entraînement des modèles se fait de manière simultanée et distribuée sur des sites ou des appareils appelés propriétaires de données, qui peuvent être des smartphones, des organisations ou des objets connectés. Ces sites viennent ensuite enrichir périodiquement un modèle global présent dans un serveur central, en communiquant des mises à jour locales (Q. YANG et al., 2019). Puis ce modèle global consolide les mises à jour reçues et envoie les nouveaux paramètres aux modèles locaux (T. LI et al., 2020). Ce concept a trouvé sa motivation suite à la prise de conscience de la sécurité des données des utilisateurs et de la confidentialité de leur vie privée (MOTHUKURI et al., 2021) ; (YIN, ZHU et J. HU, 2021). En effet, l'apprentissage fédéré permet d'exploiter les paramètres des modèles appris tout en conservant les données au sein des propriétaires des données, évitant ainsi l'échange et le stockage de données en dehors de ces derniers (L. LI et al., 2020). L'hétérogénéité dans ces structures vient de la dimension du réseau, les contraintes fonctionnelles et organisationnelles des dispositifs, la variété et la diversité des sources de données et des modèles d'apprentissage, la nature et les caractéristiques des données (J. XU et al., 2021). L'orientation des recherches propose des solutions où le serveur central est évité et où chaque site échanges avec les autres directement (LALITHA et al., 2019) ; (ROY et al., 2019). Cependant, un accord commun pour l'ensemble du réseau sur la procédure de formation des échanges doit être établi (BOUACIDA

et MOHAPATRA, 2021) ce qui reste une tâche très difficile étant donné la taille du réseau et l'hétérogénéité des sites.

Synthèse et discussion
Dans les différents aspects où l'on cherche à construire une fédération, la notion d'interopérabilité n'est pas toujours directement prise en compte lors du développement de la fédération, mais plutôt considérée après la mise en œuvre. Par conséquent, il est important de noter que dans ces architectures fédérées, l'interopérabilité est l'une des exigences à satisfaire.

Au vu des changements technologiques et des enjeux industriels actuels et futurs, la construction d'une interopérabilité fédérée est une approche prometteuse et une solution à long terme pour répondre aux exigences des fédérations en raison de sa nature générique d'adaptation à la volée à un environnement dynamique et évolutif. Les approches proposées pour l'interopérabilité fédérée se basent sur le partage d'ontologie (et/ou de métamodèle). Cependant, la création de concepts à partager et les négociations ne sont pas toujours faisables (NORAN et ZDRAVKOVIĆ, 2014). De plus, le manque de connaissance du web sémantique [7] par ces plateformes (DESHMUKH et al., 2021) et la lenteur de son adoption dans les entreprises et l'industrie (LYTRAS et GARCÍA, 2008) rendent l'utilisation des concepts à partager difficile. Ceci accentue la nécessité d'utiliser une approche pour l'interopérabilité fédérée qui ne tient pas compte seulement de l'utilisation des ontologies ou des métamodèles, et c'est ce que nous allons essayer de proposer dans la suite du manuscrit.

1.7 Conclusion

Dans ce chapitre, nous avons présenté une vue d'ensemble de la littérature sur l'interopérabilité, ses concepts et les domaines sous-jacents. Un cadre général pour l'interopérabilité est proposé pour servir de ligne directrice et soutenir tout processus d'interopérabilité. En même temps, la compilation de la littérature pour obtenir ce cadre a fait prendre conscience de l'éparpillement des connaissances actuelles sur l'interopérabilité. Ce chapitre vise donc à contribuer à la clarté et à la réunification des points de vue théoriques et pratiques sur l'interopérabilité et faciliter les travaux futurs visant à donner un sens, une précision, une profondeur et une base solide à la notion d'interopérabilité. En somme, il a donc permis de donner une meilleure vision de l'état de l'art non fermée à un domaine.

Nous avons présenté les défis et enjeux de la recherche autour de l'interopérabilité et ces différents aspects, ces points de débats, et son importance dans divers domaines d'application. En appliquant le cadre général renforcé par une analyse de l'état de l'art, l'approche fédérée a été retenue comme une méthode d'établissement de l'interopérabilité en vue de ces caractéristiques et qui permet donc de répondre aux besoins actuels et futurs qui sont principalement liés aux données. Dans la suite du manuscrit, nous allons nous recentrer sur l'interopérabilité des données qui est une déclinaison et un aspect très important de l'interopérabilité et plus profond techniquement. Le prochain chapitre présentera alors les bases sur lesquelles nos travaux se sont appuyés pour contribuer à proposer une approche implémentant une interopérabilité fédérée dans le contexte d'une migration de base de données, en proposant une approche répondant au mieux aux exigences actuelles et prenant en compte divers aspects.

7. Ce paradigme propose des standards pour exploitation des données et interpréter le contenu sémantique des pages Web. Des langages comme RDF (Resource Description Framework), OWL (Ontology Web Language), et XML (eXtensible Markup Language) sont utilisés pour collecter et structurer les données.

2

Approche fédérée pour l'interopérabilité des données : outils et concepts de base

2.1 Introduction

Au-delà de la question de définir l'interopérabilité et ses concepts, le cadre présenté dans le chapitre précédent a permis de confronter les défis actuels et futurs de l'interopérabilité et de mettre en avant l'apport de l'interopérabilité fédérée pour répondre à ces derniers ainsi que les lacunes que présentent certaines approches de fédération. Ces défis sont une source et un témoin de la croissance rapide du nombre de données hétérogènes qui accompagnent la perception, la communication et le traitement des informations. Ainsi, des ensembles de données massifs sont analysés dans le but d'identifier des modèles et de développer des connaissances exploitables à la disposition de l'utilisateur (KRISHNAN, 2013). Les deux concepts de *données* et d'*interopérabilité* sont donc étroitement liés et d'après le cadre présenté dans (D. CHEN, DOUMEINGTS et F. VERNADAT, 2008), l'interopérabilité des données est une déclinaison du problème global de l'interopérabilité et un aspect très important et très

central (L. LIU et al., 2020). Ainsi, la Figure 2.1 présente le positionnement de la thèse dans le cadre de l'interopérabilité décrit dans la Section 1.4.3 du Chapitre 1. L'approche est fédérée et vise à établir l'interopérabilité des données face aux barrières conceptuelles.

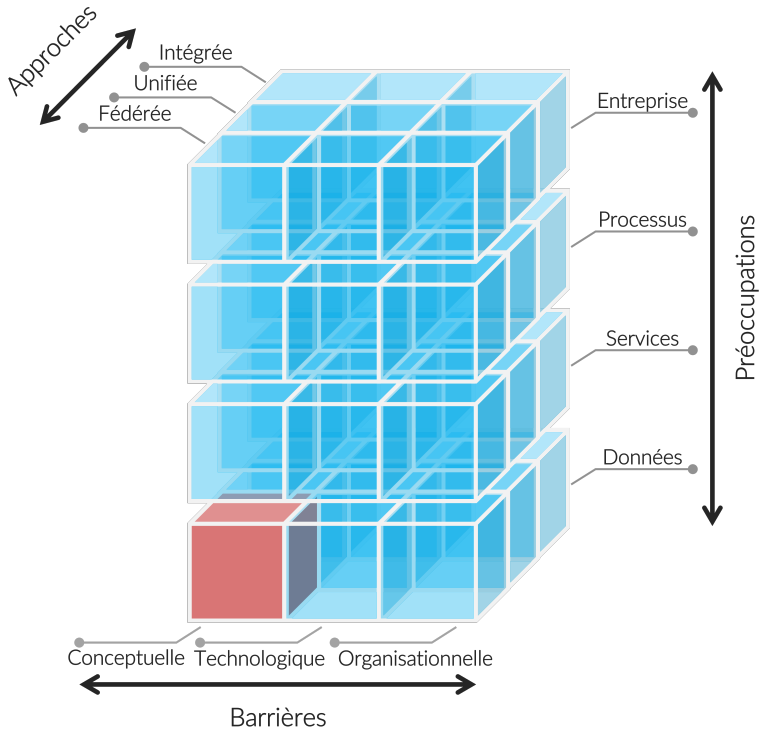


FIGURE 2.1 – Positionnement de la problématique de la thèse.

Pour adresser l'interopérabilité des données, nous nous appuyions sur les conclusions du Chapitre 1 pour proposer une approche fédérée à travers la modélisation et la résolution des problèmes d'appariement. Ainsi, nous allons voir que la modélisation par l'utilisation de la théorie des graphes et la résolution par l'utilisation des modèles d'optimisation soutenue par les techniques issues du traitement automatique des langues sont des voies intéressantes pour implémenter une interopérabilité fédérée.

L'organisation du chapitre sera comme suit, tout d'abord, la Section 2.2 et la Section 2.3 introduirons les principes de la théorie des graphes et des modèles d'optimisation. La Section 2.4 donnera une description générale des bases de données relationnelles. Ensuite, la Section 2.5 décrira l'interopérabilité des données et la Section 2.6 présentera la notion du traitement automatique des langues. La Section 2.7 présentera les problèmes d'appariement et leurs caractéristiques avec un état de l'art sur les travaux proposés pour résoudre ces problèmes. Un récapitulatif des caractéristiques de ces travaux sera effectué et qui conclura sur des éléments de réponses aux besoins d'une approche flexible, globale et générique pour contribuer à la résolution des problèmes d'appariement. Enfin, la Section 2.8 conclura le chapitre.

2.2 Principes et bases de la théorie des graphes

La théorie des graphes a été utilisée pour modéliser, analyser et appliquer des calculs dans diverses problématiques issues de plusieurs domaines d'application (BOUSSAÏD, SIARRY et

AHMED-NACER, 2017) ; (CAZABET, AMBLARD et HANACHI, 2010) ; (DI-JORIO, LAURENT et TEISSEIRE, 2009) ; (NABİYEV et al., 2016) ; (REBHI et al., 2017) et est considérée comme un outil robuste permettant de découvrir et construire des relations entre des structures. En effet, comme présenté précédemment, l'hétérogénéité est présente dans le contenu et la structure des systèmes modélisés à plusieurs niveaux d'abstraction et afin de pouvoir modéliser et manipuler les données, la théorie des graphes est un concept générique.

On note généralement $G = (V, E)$ un *graphe simple non orienté* qui est défini comme suit :

Définition 2.2.1. Un *graphe simple* non orienté G est défini par une paire $G = (V, E)$ telle que :

— V un ensemble fini de *nœuds* où $V = \{v_1, v_2, \cdots, v_n\}$,
— E un ensemble fini d'*arêtes* avec $E = \{e_1, e_2, \cdots, e_m\}$ où chaque arête représente une relation entre une paire de nœuds.

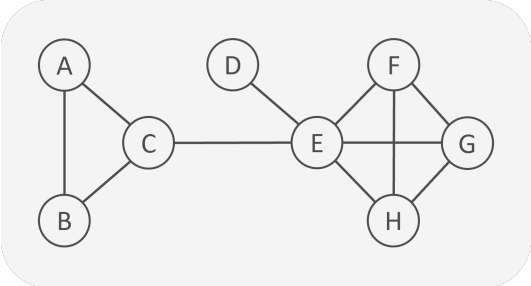


FIGURE 2.2 – Graphe simple non orienté $G = (V, E)$.

Définition 2.2.2. On dit que $G = (V_1, E_1)$ est un *sous-graphe* de $H = (V_2, E_2)$ noté $G \subseteq H$ si $V_1 \subseteq V_2$ et $E_1 \subseteq E_2$.

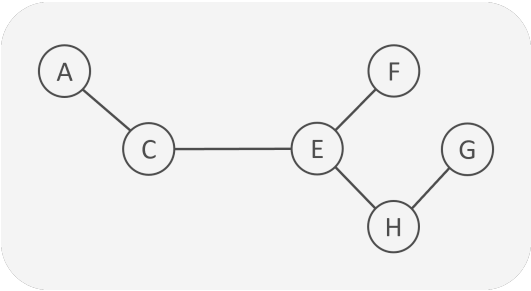


FIGURE 2.3 – G un sous-graphe de H.

Définition 2.2.3. Un *sous-graphe induit* de G est un graphe G' ayant pour nœuds un sous-ensemble S des nœuds de G et pour arêtes seulement celles joignant les nœuds de S et on écrit : $G' = (V', E')$ où $V' \subset V$ et $E' = \{(u, v) \in E \mid u \in V' \text{ et } v \in V'\}$.

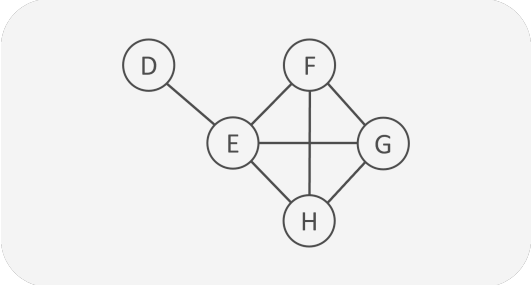


FIGURE 2.4 – G' un sous-graphe induit de G.

Définition 2.2.4. Un *graphe partiel* de G est un graphe G'' ayant pour nœuds l'ensemble des nœuds V de G et pour arêtes un sous-ensemble de E et on note : $G'' = (V, E'')$ où $E'' = \{(u, v) \in E'' \mid u \in V \text{ et } v \in V\}$.

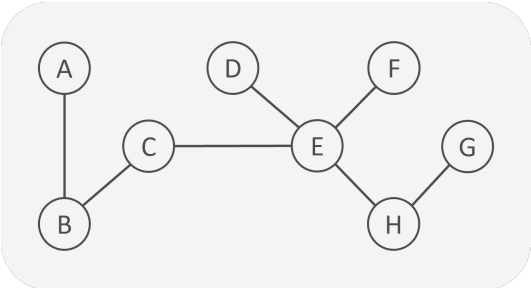


FIGURE 2.5 – G'' un graphe partiel de G.

D'autres caractéristiques peuvent être associées au graphe : l'aspect orienté des arêtes est utilisé dans le cas où des relations de dépendance existent entre les éléments d'un graphe, par exemple, pour modéliser des contraintes de dépendance comme le montre Figure 2.6. Les arêtes sont alors appelées *arcs*.

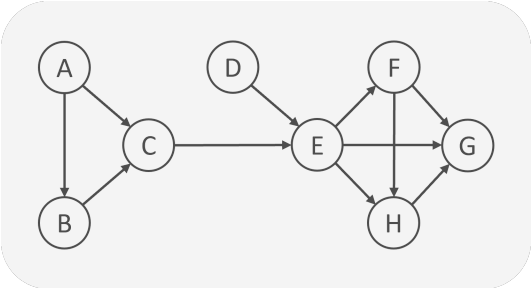


FIGURE 2.6 – Graphe simple orienté $G = (V, A)$.

Définition 2.2.5. Soit $G = (V, E)$ un graphe simple, on dit que $P = \{v_1, v_2, \cdots, v_k\}$ une suite de nœuds adjacents est un *chemin* avec v_1 et v_k les extrémités du chemin où aucune

arête n'apparaît plus d'une fois et $\forall i, (v_i, v_{i+1}) \in E$. La longueur du chemin P est le nombre d'arêtes qui le constituent.

Le nombre de nœuds ou le nombre d'arêtes (ou arcs) et leurs caractéristiques et disposition forment des graphes structurellement différents :

Définition 2.2.6. On dit que $T = (V, E)$ est un *arbre* si T est graphe acyclique et connecté.

La connexité est une propriété d'un graphe où il est possible d'aller d'un nœud à un autre en empruntant les arêtes ou les arcs. Un graphe acyclique est un graphe dans lequel il n'existe pas un chemin fermé (le nœud de départ est le nœud d'arrivée). Par exemple, la Figure 2.7 représente un arbre où il n'existe aucun chemin fermé reliant deux nœuds et où on peut trouver une succession d'arêtes pour aller de n'importe quel nœud vers un autre.

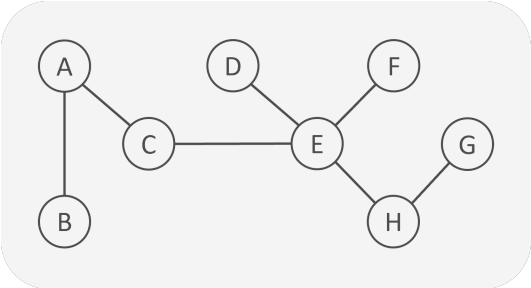


FIGURE 2.7 – Arbre $T = (V, E)$.

Définition 2.2.7. On dit que $G = (V, E)$ est un *graphe biparti* si l'ensemble des nœuds V peut être partitionné de telle sorte que $E \subseteq V_1 \times V_2$.

La Figure 2.8 montre un graphe biparti avec deux sous-ensembles V_1 et V_2 où pour une arête (u, v) chaque extrémité est dans un des sous-ensembles de V. La notion de partition des graphes bipartis est souvent utilisée pour modéliser les problèmes d'appariement par paire où chaque partition de nœuds représente les éléments d'un schéma et où chaque nœud d'une partition est relié à tous les autres nœuds de l'autre partition (graphe biparti complet). Résoudre le problème d'appariement revient à trouver un sous-ensemble d'arêtes sous certaines contraintes.

Définition 2.2.8. Soit $G = (V, E)$ un graphe non orienté. Un nœud u est *voisin* d'un nœud v s'il existe une arête $(u, v) \in E$. Pour un graphe orienté $G = (V, A)$, un nœud u est un *voisin* du nœud v si l'arc $(u, v) \in A$ ou l'arc $(v, u) \in A$. $N_G(v)$ est l'ensemble des *voisins* d'un nœud $v \in V$ tel que pour tout $u \in N_G(v)$, $(u, v) \in E$. Pour un graphe orienté, le voisinage $N_G^+(v)$ ($N_G^-(v)$) d'un nœud $v \in V$ est l'ensemble des nœuds u tel que pour tout $u \in V$, $(v, u) \in A$ ($(u, v) \in A$) où u est un successeur (un prédécesseur) de v.

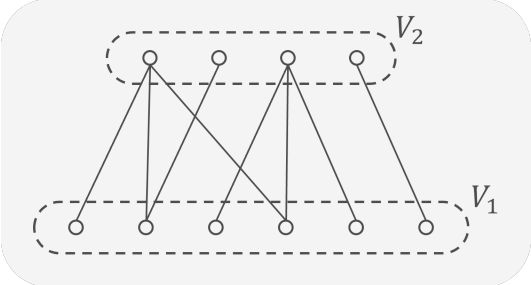


FIGURE 2.8 – Graphe biparti.

Définition 2.2.9. Dans un graphe simple non orienté, le *degré* d'un nœud v noté $deg(v)$ est le nombre des nœuds voisins. Pour un graphe simple orienté, le degré d'un nœud $deg(v) = d^+(v) + d^-(v)$ est la somme du degré sortant $d^+(v)$ (nombre d'arcs du nœud v) et du degré entrant $d^-(v)$ (nombre d'arcs vers le nœud v).

Moyennant des fonctions, des attributs (ou étiquettes) sont ajoutés aux nœuds et/ou aux arêtes du graphe afin de décrire leurs propriétés. Un graphe *attribué* ou *étiqueté* $G = (V, E)$ est un graphe auquel ont associé une fonction $\lambda : V \rightarrow L$ qui pour chaque nœud $v \in V$ associe un attribut qui décrit par exemple une tâche, un évènement ou une valeur. La structure est alors un graphe attribué $G = (V, E, \lambda)$. Une fonction peut aussi être associée aux arêtes, on note alors, $G = (V, E, \lambda, \omega)$ graphe attribué, telle que, $\omega : E \rightarrow F$.

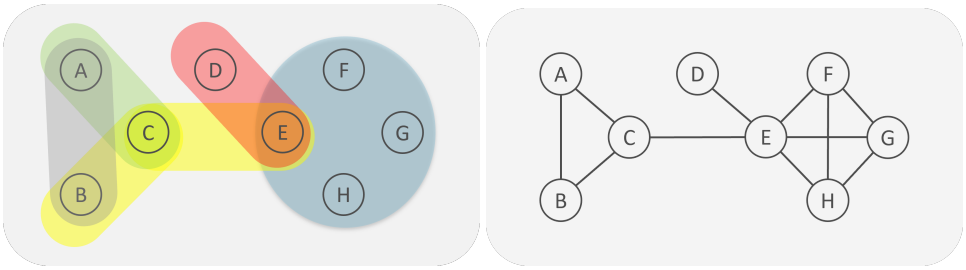
D'autres types de graphes sont utilisés pour faciliter la représentation de l'information et pour ajouter plus de détails. Par exemple, les hypergraphes généralisent la notion de graphe en définissant des (hyper)arêtes qui peuvent relier plus de deux nœuds, alors que les arêtes (ou arcs) des graphes simples ne relient que deux nœuds. D'un point de vue théorique, les hypergraphes généralisent et utilisent les mêmes concepts et notions que les graphes (BRETTO, 2013). Cette généralisation permet dans certains cas de mieux modéliser certains types de relations (J. LEE, CHO et K. M. LEE, 2011) et on définit généralement un *hypergraphe* comme suit (BERGE, 1984) :

Définition 2.2.10. Un *hypergraphe* HG est un couple (V, HE) où $V = \{v_1, v_2, \cdots, v_n\}$ est un ensemble non vide (généralement fini) de nœuds et $HE = \{he_1, he_2, \cdots, he_m\}$ est une famille de parties de V avec $he_i \neq \emptyset, i \in \{1, 2, \cdots, m\}$ et $\cup_{(i=1,2,\cdots,m)} he_i = V$, où he_i l'*(hyper)arêtes i* de l'hypergraphe H.

À partir de cette définition, les (hyper)arêtes sont représentées par des sous-ensembles de nœuds constituant des composantes liées de l'hypergraphe. Ces (hyper)arêtes traduisent ce qu'on appelle une information d'*ordre élevé* (BICK et al., 2021). En fait, les nœuds représentent une information de *premier ordre* et les arêtes ou les arcs représentent une information de *second ordre* (H. XU et al., 2020). Les hypergraphes offrent donc un niveau de formalisation plus élevé et permettent une meilleure représentation de certaines relations complexes. En effet, les relations ne sont pas toujours par paires, mais peuvent reliées plus de deux éléments.

En pratique, pour illustrer, nous proposons l'exemple d'un graphe simple $G = (V, E)$ illustré dans la Figure 2.9b composé d'un ensemble de nœuds $V = \{A, B, C, D, E, F, G, H\}$ où des

relations existent entre eux. Ces relations sont exprimées par des arêtes définies en termes de couple $E = \{(A,B),(A,C),(B,C),(C,E),(D,E),(E,F),(E,H),(E,G),(F,G),(H,G),(F,H)\}$ dans le graphe simple G. La Figure 2.9a est un exemple d'une projection du graphe G en hypergraphe $HG = (V, HE)$ où les relations sont représentées par des (hyper)arêtes définies en termes d'ensembles colorés $HE = \{(A,B)(A,C),(B,C,E),(D,E),(E,F,G,H)\}$. Comme on le voit, les hypergraphes sont utiles pour capturer les relations d'ordre élevé et permettent une meilleure représentation.



(a) Hypergraphe HG **(b)** Graphe simple G

FIGURE 2.9 – Modélisation d'un graphe et l'hypergraphe équivalent.

2.3 Modèle d'optimisation

Les problèmes d'optimisation traduisent des problématiques recherchées dans les graphes et consistent à trouver la meilleure solution à un problème modélisé en optimisant une fonction sur un ensemble. On écrit généralement (S. BOYD, S. P. BOYD et VANDENBERGHE, 2004) :

$$
\begin{array}{rlll}
 & \text{minimiser} & f(x) & \\
\text{(PL)} & \text{sous contraintes} & f_i(x) \leq 0, \quad \forall i = 1,\dots,m & (2.1)\\
 & & h_j(x) = 0, \quad \forall j = 1,\dots,p & (2.2)\\
 & & x \in \mathcal{R}^n &
\end{array}
$$

Résoudre le problème revient à trouver les valeurs de x qui minimisent la fonction $f(x)$ et qui satisferont les deux contraintes $f_i(x) \leq 0$ et $h_j(x) = 0$. Et de manière générale (LORCA, 2014) :

- La variable x est appelée *variable de décision* qui est l'inconnue du problème et qui décrit (décide) des valeurs à déterminer. Elle régit la situation à modéliser et peut être réelle, entière ou binaire.
- La fonction $f(x)$ est la *fonction objectif* qui sert de critère pour déterminer la meilleure solution au problème en lui associant une valeur à travers les variables de décision.
- L'inéquation $f_i(x) \leq 0$ et l'équation $h_j(x) = 0$ sont appelées les *contraintes* du problème exprimé en fonction des variables de décision et que les variables doivent vérifier.

Le problème (PL) est réalisable s'il existe au moins une assignation d'une valeur à x qui vérifie les contraintes du problème et infaisable dans le cas contraire. La valeur optimale $p^\star$ du problème (PL) est définie comme : $p^\star = inf\{f(x)|f_i(x) \leq 0, i = 1,\cdots,m, h_j(x) = 0, j = 1,\cdots,p\}$. Autrement dit, il n'existe pas une autre valeur q qui soit inférieure à $p^\star$. Les problèmes linéaires sont les problèmes d'optimisation où la fonction objectif et les contraintes sont linéaires.

D'autres catégories de problèmes d'optimisation peuvent être identifiées ou dérivées :

— Les *problèmes d'optimisation continue* où les variables prennent des valeurs réelles (continues) et qui sont en général simples à résoudre.
— Les *problèmes d'optimisation discrète* où les variables de décision prennent des valeurs entières (discrètes) ou binaires. Si une partie des variables prend des valeurs réelles, le problème d'optimisation est dit *mixte en nombres entiers*, sinon, le problème d'optimisation est dit *en nombres entiers*. Ce genre de problème est plus difficile à résoudre que les problèmes continus.
— Les *problèmes d'optimisation combinatoire* sont les problèmes d'optimisation où les ensembles réalisables sont finis, mais qui augmentent exponentiellement en fonction de la taille du problème.
— Les *problèmes d'optimisation multiobjectif* sont les problèmes d'optimisation où un compromis doit être trouvé entre plusieurs objectifs.
— Les *problèmes d'optimisation stochastique* sont les problèmes d'optimisation où l'incertitude peut être associée aux variables.

2.4 Base de données

De manière générale, une base de données est une collection organisée de données stockées dans des fichiers et potentiellement hébergés dans des serveurs en ligne. Pour une organisation ou une entreprise, les *données* proviennent des systèmes d'information, des applications d'entreprise ou d'autres sources externes. Les données peuvent être quantitatives (numériques) ou qualitatives (descriptives). Elles sont généralement brutes, sans contexte. Une fois qu'une donnée est contextualisée, la donnée présente alors une *information*. Les informations pertinentes sont ensuite analysées et utilisées pour produire une *connaissance*. Les bases de données sont gérées par ce qu'on appelle des *systèmes de gestion de base de données* (SGBD) qui permettent par exemple d'écrire, de manipuler, de partager ou encore de sécuriser les données grâce notamment à des requêtes écrites dans des langages spécifiques, par exemple, SQL (Structured Query Language). Un SGBD gère des bases de données structurées comme les bases de données *relationnelles* ou bien des bases de données peu structurées comme les bases de données NoSQL (Not only SQL) qui ont reçu une attention particulière avec le développement de l'IoT (AZAD et al., 2020).

Une base de données relationnelle est le type de base de données le plus courant. Elle utilise un schéma, qui est un modèle pour dicter la structure des données stockées dans la base. Plusieurs notions sont associées aux données relationnelles illustrées dans la Figure 2.10 et la Figure 2.11 et présentées ci-dessous :

— les données sont organisées dans une ou plusieurs *tables*,
— une table représente une *relation*,
— chaque *ligne* de la table est appelée *tuple* ou *enregistrement*,
— une table est composée de plusieurs *colonnes* et chaque colonne est nommée par un *attribut* qui est donc un identificateur décrivant une information,
— il n'y a pas d'ordre particulier entre les colonnes ou les lignes d'une table,
— en plus d'un attribut, chaque colonne possède un *type* de données. Un *domaine* est l'ensemble des valeurs valides (possibles) d'un attribut, ou ce qu'on appelle l'intégrité de domaine,
— le *schéma d'une relation* (table) est la description d'une relation (d'une table) par ses attributs, leurs domaines et les clés,
— l'ensemble des schémas des relations qui la composent constituent le *schéma d'une base de données relationnelle*,
— une *clé primaire* est un ensemble non vide d'attributs qui permet d'identifier de manière unique un tuple.

— une *clé unique* est similaire à la clé primaire, mais contrairement à cette dernière, une clé unique autorise des valeurs nulles,
— une *clé étrangère* est une clé qui fait référence à une clé primaire ou unique appartenant à une autre table.

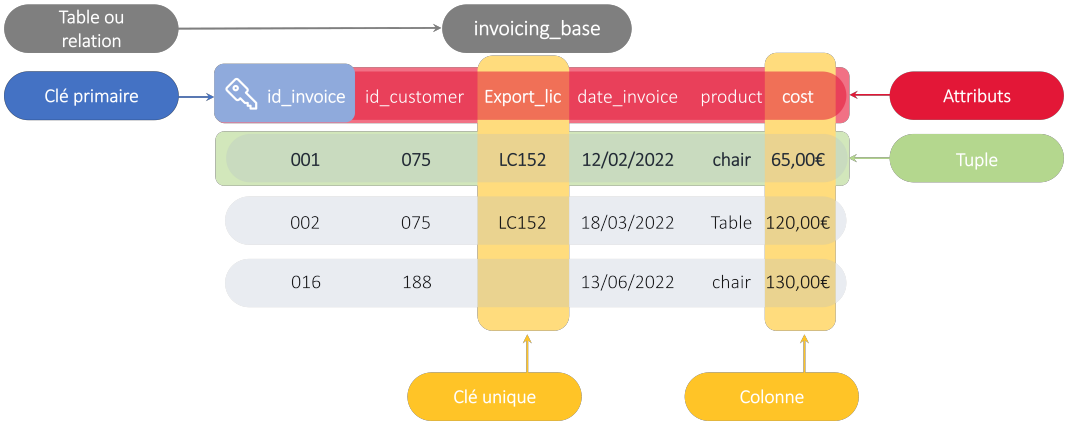


FIGURE 2.10 – Exemple d'une table dans une base de données relationnelle.

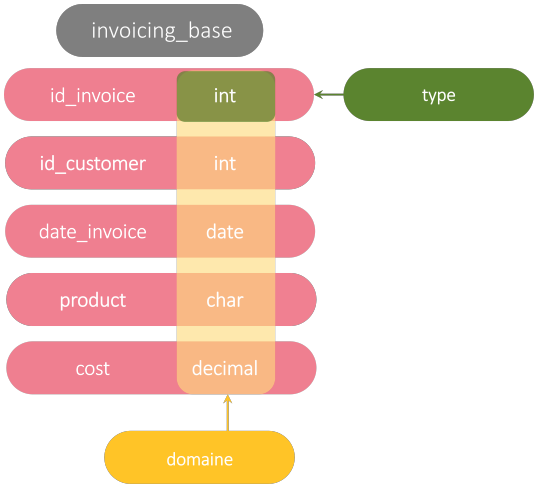


FIGURE 2.11 – Le domaine et les types de données associés à la table dans la Figure 2.10.

De manière plus formelle, dans ce qu'on appelle l'algèbre relationnelle, soit un ensemble d'attributs $\{A_1, A_2, \cdots, A_n\}$ et les domaines associés $\{D_1, D_2, \cdots, D_n\}$ (BELTRAN, JAUDOIN et PIVERT, 2015) :

Définition 2.4.1.

— Une *relation* de degré n sur les domaines $\{D_1, D_2, \cdots, D_n\}$ est un sous-ensemble fini du produit cartésien $D_1 \times D_2 \times \cdots \times D_n$.
— Un schéma de relation s'écrit $R(A_1 : D_1, A_2 : D_2, \cdots, A_n : D_n)$, R est le nom de la relation et A_i les noms d'attributs deux à deux distincts.
— Un élément de R est un *n-uplet* $(a_1, a_2, \cdots, a_n)$ avec a_i la valeur de l'attribut A_i.

Définition 2.4.2. Soit S un sous-ensemble d'attributs du schéma de relation R, et A un attribut de R. On dit que A dépend fonctionnellement de S, et on écrit $S \rightarrow A$, si, et seulement si, des valeurs identiques de S impliquent des valeurs identiques de A.

Définition 2.4.3. Une clé d'une relation R est un sous-ensemble minimal K des attributs, tel que tout attribut de R dépend fonctionnellement de K.

2.5 Interopérabilité des données

Les données sont à la base de toute information, mais sont souvent dispersées et stockées dans des silos et formats différents. Les entités génèrent des données précieuses, mais ne sont pas toujours interprétables ni toujours utilisées par une autre entité. Le cycle de vie des données fait référence aux étapes que traversent les données depuis la création jusqu'à la suppression. Il comprend plusieurs étapes telles que les processus de collecte, de stockage, de transformation, et d'analyse des données. L'interopérabilité intervient alors dans de nombreuses étapes pendant lesquelles des mécanismes doivent être réalisés. Parmi eux, le couplage de dossiers, l'échange de données, l'intégration ou la migration de données (MAHANTI, 2021). Ces mécanismes agissent en général sur trois types d'objets autour des données (A. A. A. ALGERGAWY, 2010) :

1. Le modèle de données concerne la structure utilisée pour représenter les données (tableaux, objets ou fichiers) ;
2. Le schéma de données concerne la structure qui décrit ou supporte les données (deux données peuvent représenter la même information, mais utilisent des types ou des noms différents) ;
3. L'instance de données concerne les tuples des données (différentes données provenant de différentes sources représentent la même entité).

L'objectif de l'intégration des données est de créer une vue pour joindre des informations provenant de différentes sources indépendantes suivant un schéma de données global (DOAN, A. HALEVY et IVES, 2012). Additionnellement à l'intégration des données, la migration des données est un autre domaine de recherche important qui consiste à transférer des données en changeant le système de stockage, de schémas ou modèles de données (MORRIS, 2012). Il est important de noter que souvent l'intégration et la migration des données sont des processus difficiles à réaliser et parfois confondus (MAATUK, ALI et ROSSITER, 2008), en raison de la complexité et de l'ambiguïté des caractéristiques des données (ANTHES, 2010). La migration est parfois considérée comme un processus complémentaire ou une sous étape à l'intégration (SCHULTZ et WISNESKY, 2017) ; (SHRESTHA et al., 2019) et vice et versa (DRUMM et al., 2007).

Ces deux mécanismes s'appuient sur diverses tâches ou étapes communes, on en cite deux des plus importants [1] :

— Résolution d'entité (ER pour Entity Resolution) (CHRISTOPHIDES et al., 2019) : cette tâche consiste à identifier et fusionner des enregistrements ou des tuples (une instance de données peut être divisée en tuple) qui correspondent à la même entité, mais qui proviennent de différentes sources de données (GETOOR et MACHANAVAJJHALA, 2013). Différents processus peuvent être réalisés : la détection des duplications (NAUMANN et HERSCHEL, 2010) ou la correspondance d'entités (FAN et al., 2009).

1. (SCHULTZ, SPIVAK et WISNESKY, 2016) présentent d'autres mécanismes

— Appariement [2] des schémas (SM pour Schema Matching (RAHM et BERNSTEIN, 2001) ou Schema Mapping (KOLAITIS, 2005)) [3] : qui gèrent l'identification des correspondances entre les éléments des schémas de données, autrement dit les schémas qui décrivent la structure générale qui moule les données en utilisant leurs informations et propriétés, à savoir, la structure, les données et d'autres sources d'information.

L'appariement des schémas joue un rôle central et important non seulement dans l'intégration ou la migration des données, mais aussi dans de nombreux autres mécanismes (DRUMM, 2008). Cet axe de recherche a reçu une attention particulière au cours des dernières années (PARUNDEKAR, KNOBLOCK et AMBITE, 2014). En mettant en œuvre un processus d'appariement, cela identifie des relations entre des systèmes représentés par des schémas de données en format XML (JEONG et al., 2008), des bases de données (CASANOVA et al., 2007), des métamodèles (LAFI, FEKI et HAMMOUDI, 2014) ou des ontologies (SHVAIKO et EUZENAT, 2011), etc. En d'autres termes, les systèmes sont présentés comme modèle(s) source(s) et cible(s) et évoquent des données hétérogènes générées indépendamment dans différents contextes avec des modèles, des schémas, des niveaux d'abstraction, de compréhension et de représentation différents. L'objectif est alors de créer les connexions entre ces systèmes (KAPIL, AGRAWAL et R. KHAN, 2016). Pour assister le processus d'appariement, la discipline du traitement du langage naturel offre un panel de méthodes et de techniques pour la compréhension, la manipulation et la génération des informations textuelles.

2.6 Traitement du langage naturel

Avant d'aborder les problèmes d'appariement, nous allons présenter les principes du traitement du langage naturel (NLP pour Natural Language Processing) souvent utilisés dans ce domaine. Ce champ de recherche est à l'intersection du domaine linguistique, l'informatique, les mathématiques et l'intelligence artificielle. Plusieurs champs et applications sont étudiés et sont scindés en plusieurs catégories, on en cite par exemple :

— Chargement des sources de données par exemple en utilisant le Web Scraping (B. ZHAO, 2017).

— L'analyse et le traitement du texte ou des mots :

  — Tokenisation permet de diviser un mot ou une chaîne de caractères en sous-chaînes (MICHELBACHER, 2013) ;

  — Racinisation syntaxique permet de retrouver la racine d'un mot après avoir supprimé les préfixes et les suffixes textuels (SINGH et V. GUPTA, 2016) ;

  — Lemmatisation permet de produire des racines (forme de base ou lemme) grâce notamment à l'analyse morphologique ou l'utilisation des bases de données lexicales comme WordNet (G. A. MILLER, 1995) ;

  — Reconnaissance d'entités nommées permet d'extraire des mots ou un groupe de mots et de les classer dans des classes ou des entités (COCHE et al., 2020) ;

  — Le calcul de similarité permet de quantifier à quel point des mots (des phrases, des documents ou des séquences numériques) sont similaires du point de vue de la syntaxe (la forme) et de la sémantique (le sens) (GOMAA, FAHMY et al., 2013) ; (HAKAK et al., 2019) ; (PRAKOSO, ABDI et AMRIT, 2021) ; (J. WANG et Y. DONG, 2020) ;

2. Identifier dans la littérature par la conciliation ou la réconciliation, la mise en correspondance, l'alignement, la liaison.

3. Les notions de Matching et Mapping sont parfois confondues, mais généralement dans la littérature, le processus du Mapping vient après le processus de Matching. Le matching vise à trouver les correspondances entre les éléments des schémas (appariement) alors que le processus de mapping crée les expressions dans un format spécifique permettant de traduire les correspondances précédemment trouver (BELLAHSENE et al., 2011) ; (BONIFATI et al., 2010).

  - Représentation des mots par des vecteurs de réels (représentation distribuée (MIKOLOV et al., 2017), représentation dynamique (Y. WANG et al., 2020), représentation par plongement lexical (F. ALMEIDA et XEXÉO, 2019)).
- Développement de système de :
  - Génération de texte par exemple par la synthèse ou le résumé de texte (ALLAHYARI et al., 2017) ;
  - Recommandation (TAGHAVI et al., 2018).
- Visualisation des informations, par exemple les graphes de connaissance (HOGAN et al., 2021).

Les techniques sont diverses et leurs utilisations a pour objectif principal de traiter et analyser des données pour assister des systèmes à les comprendre et les interpréter. Pour le prochain chapitre, nous allons aborder le concept du calcul de la similarité et de la représentation des mots par plongement lexical.

2.7 Problèmes d'appariement

Dans la littérature, différents mécanismes font référence ou viennent compléter le processus d'appariement, on cite :

- La réconciliation des schémas définie comme un processus d'appariement (FOLINO et al., 2012) ou un processus complémentaire d'examen, de validation et de correction des correspondances générées (QUOC VIET NGUYEN et al., 2013).
- L'alignement des schémas est défini comme le processus d'aligner les mêmes concepts, autrement dit, apparier les mêmes éléments présents dans des schémas différents, mais homogènes (SUCHANEK, ABITEBOUL et SENELLART, 2011).
- La transformation des schémas définie comme le processus d'appariement en deux étapes, la définition des règles de transformation et leur exécution (FOERSTER et al., 2010).
- L'intégration des schémas consiste à trouver un schéma de données unifié, appelé schéma intégré (CHITICARIU et al., 2007).
- La liaison des schémas est un autre processus qui identifie des relations entre des données textuelles et les éléments d'une base de données (B. WANG et al., 2019).

En général, deux grands axes de recherche sont exploités, le premier axe concerne les recherches qui traitent deux schémas en entrée, un schéma source et un schéma cible, qu'on appelle les problèmes d'appariement *par paire* (SHVAIKO et EUZENAT, 2011), et le second axe concerne les recherches qui traitent plus de deux schémas en entrée qu'on appelle *appariement holistique* (NGO et BELLAHSENE, 2016) qui a reçu une attention particulière ces dernières années (VOIGT, 2011). Dans le cadre de nos travaux, on s'intéresse à l'appariement par paire. Pour le résoudre, trois types d'informations sont utilisés (ALWAN et al., 2017) :

1. Informations disponibles au niveau des schémas : qui concernent les informations sur les éléments et les contraintes qui constituent le schéma, autrement dit le modèle du schéma (NIKOLOV, UREN et E. MOTTA, 2010).
2. Informations disponibles dans les instances de données : qui concernent leurs valeurs ou contenus des données (DORNELES, R. GONÇALVES et SANTOS MELLO, 2011).
3. Informations disponibles dans des sources auxiliaires : qui concernent les informations supplémentaires pouvant être contenues dans des sources externes (Y. LIU et al., 2015) ; (PORTISCH, HLADIK et PAULHEIM, 2021).

En exploitant ces informations, résoudre le problème d'appariement revient à trouver les correspondances entre les éléments des schémas source et cible. Plus formellement, le *problème de l'appariement* est généralement défini comme suit (BERLIN et MOTRO, 2002) ; (Z. ZHANG et al., 2008) :

Définition 2.7.1. Un *schéma* S est défini comme étant un ensemble fini d'attributs $S = \{a_1, a_2, \cdots, a_n\}$. Étant donné deux schémas S_1 et S_2 avec n et m attributs respectivement. Résoudre le problème d'appariement des schémas revient à trouver un *ensemble de correspondances* entre les deux schémas S_1 et S_2. Une *correspondance* est définie comme une relation entre un (ou plusieurs) élément d'un schéma et un (ou plusieurs) élément d'un autre schéma, avec notamment un certain degré de *similarité*.

Les attributs d'un schéma peuvent être (1) simples, c.-à-d. qu'ils représentent un seul élément du schéma ou (2) complexes (composés de plusieurs attributs) qui représentent un sous-ensemble d'attributs, mais ne doivent pas nécessairement être disjoints (GAL, 2006). Les valeurs des attributs d'un même ensemble peuvent dépendre les unes des autres (CHRISTEN et VATSALAN, 2013).

Par exemple, Figure 2.12 illustre un exemple de deux schémas S_1 et S_2, chacun contenant une seule table représentant des informations sur des factures. Les deux tables se composent respectivement de 9 et 8 attributs. Le problème d'appariement consiste à trouver quel attribut de la table du schéma S_1 correspond à quel autre attribut de la table du schéma S_2. L'ensemble des correspondances possibles est $\mathcal{S} = S_1 \times S_2$ de taille $|\mathcal{S}| = 72$.

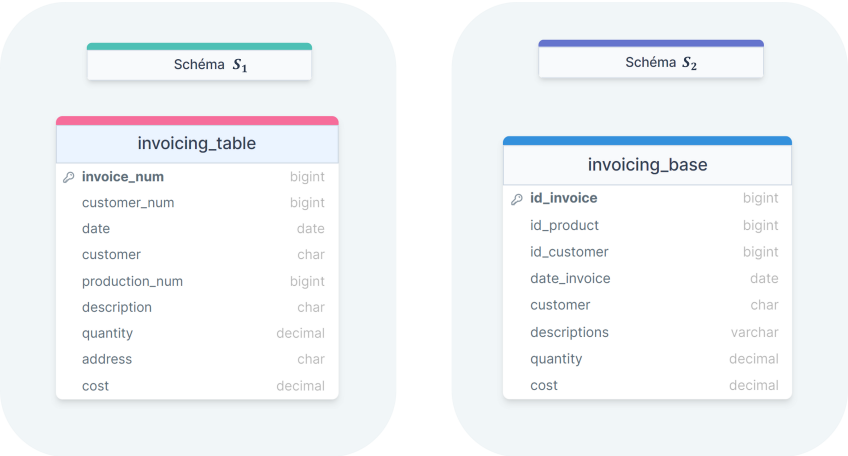


FIGURE 2.12 – Exemple de deux schémas S_1 et S_2 composés d'une table chacun.

2.7.1 Processus général d'appariement

Le processus d'appariement présenté dans la Figure 2.13 (réadapté de (RAHM, 2011)) est généralement complexe et est constitué de deux grandes phases, appariement de première ligne et appariement de seconde ligne (GAL et SAGI, 2010). Souvent, ces deux phases sont précédées par une phase de prétraitement (mais pas toujours) où deux tâches sont effectuées : une tâche de transformation qui permet d'importer et de traduire les schémas en entrées à d'autres modèles de données et une tâche d'identification qui permet d'établir (si possible) et identifier les premières relations (correspondances) entre les éléments des schémas et de les utiliser ensuite dans les principales phases du processus. D'autres tâches de suppression, d'ajout ou de modification d'éléments peuvent être aussi envisagées.

2.7.1.1 Appariement de première ligne

Des mesures de similarité sont appliquées directement aux schémas (source(s) et cible(s)) en entrée. La mesure de similarité est généralement exprimée sous la forme d'une valeur

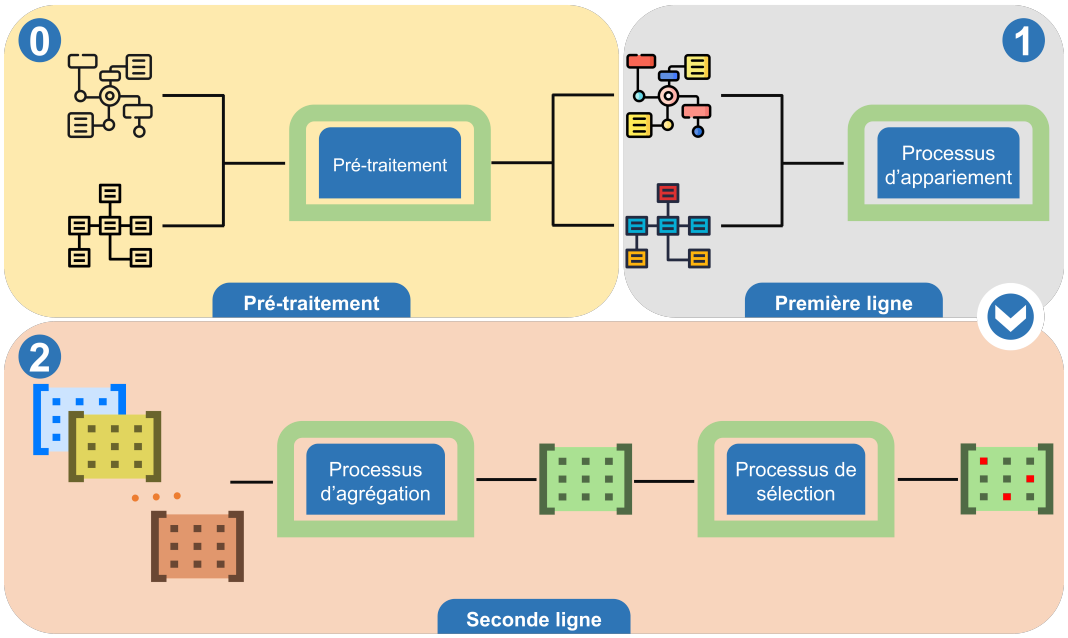


FIGURE 2.13 – Processus général d'appariement.

numérique comprise entre 0 et 1. Il est important de noter la différence entre les termes similarité et distance (BATET et SÁNCHEZ, 2015) ; (HADJ TAIEB, ZESCH et BEN AOUICHA, 2020). La similarité estime la similitude, par exemple linguistique, entre deux mots, tandis que la distance estime ou quantifie à quel point les deux mots sont différents, autrement dit, la distance est la transformation linéaire de la similarité. Ainsi, une distance égale à 1 (resp. une similarité égale à 1) signifie que les deux éléments comparés sont complètement dissemblables et n'ont aucune caractéristique commune (resp. signifie que les deux éléments comparés sont identiques ou parfaitement similaires). Enfin, les techniques, métriques ou mesures de similarité utilisées exploitent les informations des schémas et permettent de quantifier ou de qualifier le degré de similarité. L'utilisation d'une mesure dépend du contexte d'application, les éléments disponibles ainsi que le type et l'hétérogénéité des schémas sources et cible.

Pour les deux schémas S_1 et S_2, $\mathcal{S} = S_1 \times S_2$ l'ensemble des correspondances d'attributs possibles entre S_1 et S_2. Soit M la matrice de taille $n \times m$ qui représente donc les valeurs de similarité entre chaque paire d'élément de S_1 et S_2 avec $n = |S_1|$ lignes, $m = |S_2|$ colonnes et $m_{ij} \in [0,1]$ qui quantifie le degré de similarité entre les attributs $i \in S_1$ et $j \in S_2$.

2.7.1.2 Appariement de seconde ligne

Des algorithmes qui prennent une ou plusieurs matrices de similarité sont utilisés pour générer une seule matrice binaire notée $\mathcal{M}$ avec $m_{ij} \in \{0,1\}$ qui signifie que si $m_{ij} = 1$, l'attribut $i \in S_1$ correspond à l'attribut $j \in S_2$, sinon si $m_{ij} = 0$, l'attribut i ne correspond pas à l'attribut j. Trois processus sont alors appliqués :

1. Le processus d'agrégation : appliqué aux matrices, permet d'agréger plusieurs matrices de similarité issues de l'appariement de première ligne en une seule matrice en utilisant des stratégies de maximum, minimum, pondération ou encore la moyenne et l'optimisation (d'autres stratégies sont présentées par (ELSHWIMY et al., 2014)).
2. Le processus de sélection : appliqué à la matrice résultante de l'agrégation qui permet de déterminer la meilleure correspondance pour un élément d'un schéma en utilisant la notion de seuil, du maximum ou encore une sélection des k meilleures correspondances (GAL et al., 2012) (d'autres stratégies sont expérimentées par (GAL et SAGI, 2010)).

3. Le processus de composition est optionnel. Cette étape utilise des stratégies de moyennes ou de ratio en prenant le nombre d'éléments appariés et est appliquée (au besoin) pour donner une unique valeur évaluant le process d'appariement (des techniques sont présentées par (GALI et al., 2019)).

Soit $\Sigma = 2^{\mathcal{S}}$ l'ensemble des correspondances possibles et, soit $\Gamma : \Sigma \rightarrow \{0, 1\}$ une fonction booléenne qui indique si une application de contraintes spécifiques [4] a été considérée ou non (GAL et SAGI, 2010). À titre d'exemple, on peut considérer ou non les contraintes de cardinalité liées aux problèmes d'appariement qui forcent le choix d'une ou plusieurs correspondances, comme dans le cas où un attribut $a \in S_1$ correspond à deux attributs b et c tel que $b, c \in S_2$. L'ensemble $\Sigma_\Gamma = \{\sigma \in \Sigma | \Gamma(\sigma) = 1\}$ est alors l'ensemble des correspondances valides accepté sous la fonction Γ.

2.7.2 Contraintes dans les problèmes d'appariement

Des conditions et des restrictions sont également prises en compte pour guider le processus d'appariement par le biais des contraintes qui peuvent être :

— Contraintes de cardinalité : c'est-à-dire que la correspondance peut mettre en relation un ou plusieurs éléments d'un schéma avec un ou plusieurs éléments d'un autre schéma (ou d'autres schémas). Quatre cas sont déduits : un-à-un *(1 : 1)*, un-à-plusieurs *(1 : m)*, plusieurs-à-un *(n : 1)*, plusieurs-à-plusieurs *(n : m)* (GAL, 2005).
— Contraintes de dépendance : par exemple spécifique à un modèle de données qui exprime une relation hiérarchique dans un XML schéma (YU et JAGADISH, 2006) ou une relation de dépendance fonctionnelle entre des attributs d'une base de données relationnelles (FAN, Y. WU et J. XU, 2016) ou une relation d'interdépendance basée sur les instances (KANG et NAUGHTON, 2008).

2.7.3 Hétérogénéité dans les problèmes d'appariement

Les informations exploitées par les techniques d'appariement présentent plusieurs formes d'hétérogénéité qui peuvent être présentées en trois niveaux (RAM et PARK, 2004) :

1. Au niveau de la structure : le format d'écriture ou de stockage des données et des schémas (XML, TXT, CSV, etc.).
2. Au niveau des données :
   — Valeur : différentes interprétations de la valeur de la donnée (un nombre peut représenter une quantité ou un prix) ;
   — Représentation : type ou format de représentation de la donnée (représenter une catégorie de produit par une lettre ou un nombre) ;
   — Unité : utilisation de différentes unités de mesure (par exemple *cm* et *m*) ;
   — Précision : différentes échelles ou de précision (nombre de chiffres utilisés pour exprimer une valeur).
3. Au niveau des schémas :
   — Nom : les étiquettes ou noms des éléments du schéma ;
   — Identifiants : utilisation de différents identifiants (par exemple différentes clés primaires) pour les mêmes concepts ou attributs ;
   — Définition : utilisation de différents ensembles d'attributs pour définir la même entité (différents attributs pour une même table) ;
   — Structuration : différentes entités pour deux schémas représentant la même information.

4. Plus de détails sur les types de contraintes et leurs caractéristiques peuvent être trouvés dans (DOAN, DOMINGOS et A. Y. HALEVY, 2001).

2.7.4 Méthodes et techniques de base d'appariement

L'utilisation des informations aux niveaux des schémas, des instances de données et des informations auxiliaires, revient au contexte, la disponibilité et l'hétérogénéité de ces derniers. Plusieurs travaux se sont tournés vers la classification des méthodes d'appariement utilisant ces informations (BERNSTEIN, MADHAVAN et RAHM, 2011) ; (EUZENAT, SHVAIKO et al., 2007) ; (EVERMANN, 2008) ; (X. LIU et al., 2021) ; (OCHIENG et KYANDA, 2018) ; (OTERO-CERDEIRA, RODRÍGUEZ-MARTÍNEZ et GÓMEZ-RODRÍGUEZ, 2015) ; (RAHM, 2011) ; (SHVAIKO et EUZENAT, 2011) ; (SUTANTA et al., 2016) ; (THIÉBLIN et al., 2020).

(RAHM et BERNSTEIN, 2001) proposent l'une des classifications les plus utilisées à ce jour, présentée dans la Figure 2.14 et fournissent une base pour des travaux ultérieurs. Les auteurs présentent une taxonomie à plusieurs niveaux :

- Approche individuelle : qui se basent sur une méthode où on considère :
  - *schema-only based* : qui sont les méthodes qui utilisent les informations contenues dans les schémas, à savoir, *element-level* qui couvrent les noms, les descriptions, les contraintes, etc. ou *structure-level* qui couvrent la structure ;
  - *instance/contents-based* : qui utilisent la valeur des données contenues dans les instances.
- Approches combinées : approches qui se basent sur la combinaison de plusieurs méthodes :
  - *hybrid* : qui combinent plusieurs méthodes afin de trouver des candidats mieux adaptés ;
  - *composite* : qui combinent des résultats obtenus séparément manuellement ou automatiquement.

Un dernier niveau concerne l'utilisation des informations *linguistic* (les propriétés sémantiques et syntaxiques) ou l'utilisation des *constraints*. Dans une version améliorée, (DO, 2006) ; (DO et RAHM, 2002) ajoutent une autre catégorie, à savoir, *reuse-oriented* qui utilise des informations supplémentaires telles que des dictionnaires de synonymes, des répertoires structurés ou des résultats déterminés précédemment.

La complexité du problème d'appariement, les contraintes et l'hétérogénéité qui réside à plusieurs niveaux, ont motivé le développement d'approches combinées au cours des trois dernières décennies (RAHM, 2011) ; (RAHM et BERNSTEIN, 2001) ; (THIÉBLIN et al., 2020). Cette motivation vient de l'hypothèse que l'amélioration des résultats des problèmes d'appariement vient par la combinaison de plusieurs méthodes exploitant différentes informations ou plusieurs méthodes exploitant le même type d'information. Dans le cadre de nos travaux, on s'intéresse aux approches combinées. La prochaine section s'intéressera davantage aux approches combinées qui utilisent la représentation en forme de graphe pour modéliser les schémas, les modèles d'optimisation pour la recherche des correspondances, la combinaison de plusieurs stratégies pour la sélection ou encore la combinaison de mesures de similarité.

2.7.5 État de l'art : approches générales d'appariement

Une des premières approches les plus utilisées est le COMA pour COmbining MAtchers, avec une première version pour le problème d'appariement des schémas (DO et RAHM, 2002) puis une amélioration qui supporte l'appariement d'ontologies avec une interface utilisateur (GUI pour Graphical User Interface) (AUMUELLER et al., 2005) et enfin une dernière mise à jour pour gérer d'autres problèmes d'appariement plus complexes (MASSMANN et al., 2011). Le système représente les schémas sous forme de graphe orienté acyclique et possède une bibliothèque de méthodes d'appariement et de stratégies de combinaison. COMA exploite les instances de données, les noms et les types des éléments, les informations structurelles ainsi que les précédents résultats d'appariement.

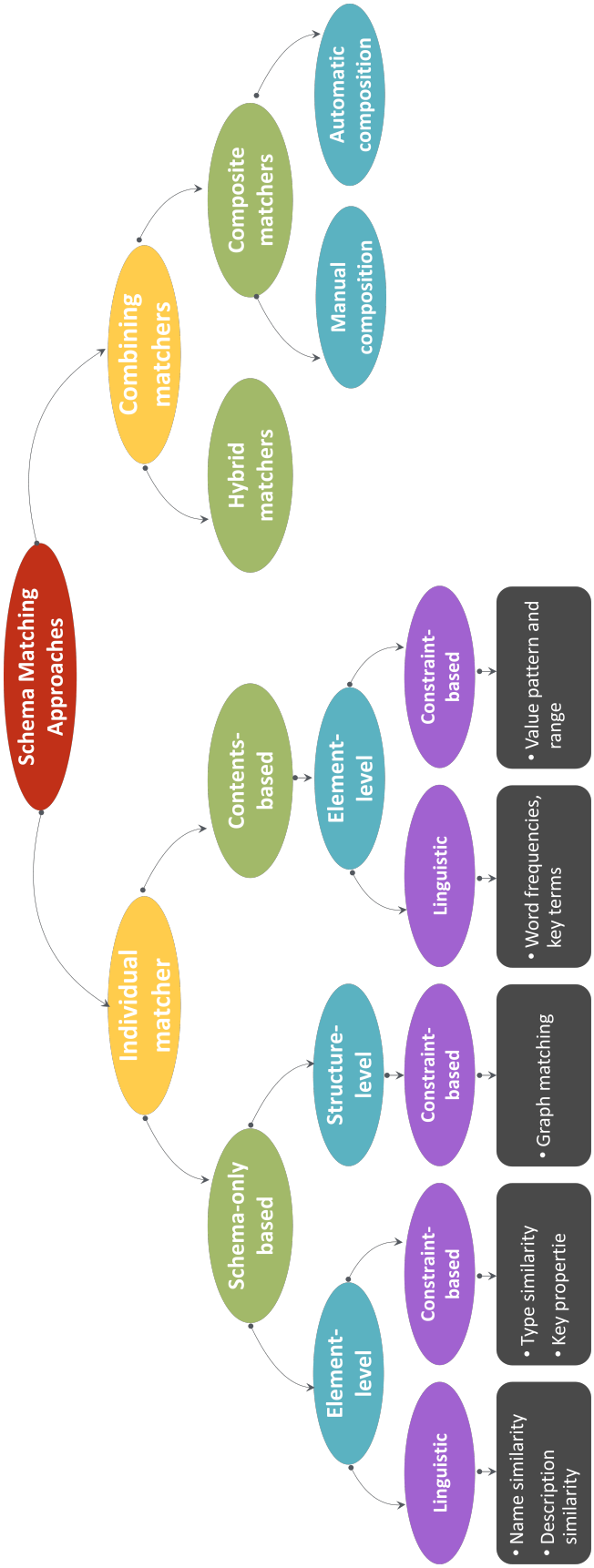


FIGURE 2.14 – Classification basique des méthodes d'appariement des schémas, (RAHM et BERNSTEIN, 2001).

Une autre approche connue dans la littérature est Similarity Flooding (MELNIK, GARCIA-MOLINA et RAHM, 2002). Cette approche utilise la représentation des schémas par des graphes orientés où les nœuds ont des attributs. Cette approche s'appuie sur le principe que deux nœuds qui sont susceptibles d'être appariés, appartiennent au même voisinage. Cette approche calcule la similarité basée sur les attributs et applique une opération de propagation dans les voisinages des nœuds jusqu'un point fixe afin d'obtenir des correspondances raffinées. Une dernière étape sélectionne les paires de nœuds satisfaisant un seuil de similarité. L'approche prend donc un ensemble de valeurs de similarité et les propage à travers la structure du graphe. D'autres aspects d'amélioration de la propagation sont proposés dans la littérature (MELNIK, 2004) ; (PESCHL et DEL FABRO, 2015) ; (Q. WANG et X. WEN, 2015) et ces travaux statuent sur l'importance de propager des informations certaines afin de réduire la marge d'erreur (DURAND et al., 2020).

(CRUZ, ANTONELLI et STROE, 2009) présentent AgreementMaker, un système itératif conçu pour l'appariement d'ontologies sur trois niveaux. Un premier niveau qui calcule les matrices de similarité linguistiques avec des comparateurs de chaînes de caractères et des bases de données lexicales comme source d'information auxiliaire. Un deuxième niveau exploite les informations concernant les propriétés et les relations des éléments formant les ontologies. Le troisième niveau combine les résultats des deux précédents niveaux en se basant sur les matrices de similarité construites. Ensuite, grâce à l'utilisation d'un seuil déterminé manuellement et un modèle d'optimisation, les correspondances finales sont données. Ce système a connu de nombreuses améliorations pour inclure d'autres mesures de similarités, des mécanismes d'évaluation et d'amélioration de la qualité des résultats (FARIA et al., 2013) ; (FARIA et al., 2019).

(UNAL et AFSARMANESH, 2010) proposent SASMINT pour Semi-Automatic Schema Matching and INTegration, un système semi-automatique pour l'appariement des schémas et l'intégration (création des règles sous forme d'un langage spécifique) de bases de données relationnelles sur la base d'informations linguistiques et structurelles. Le système opère selon quatre étapes, une première étape où des poids sont déterminés manuellement ou moyennant une fonction de calcul [5]. Une deuxième étape concerne le choix de la stratégie de sélection, par exemple, par l'utilisation du maximum ou un seuil. Une troisième étape consiste à transformer les schémas en graphes orientés acycliques en modélisant les tables, colonnes et les clés étrangères. La dernière étape consiste à sélectionner les potentielles correspondances. L'utilisateur peut intervenir pour valider ou modifier les résultats, ensuite, les schémas sont intégrés moyennant des règles d'intégration.

(SELLAMI, 2009) propose une méthodologie pour l'appariement de schémas structurés XML. La méthodologie est sur trois étapes, une première étape lors de laquelle des sous schémas sont générés en se basant sur la structure et les informations linguistiques. Ensuite, des mesures de similarités sont appliquées aux schémas dérivés deux-à-deux. Dans la dernière étape, l'agrégation est effectuée avec une stratégie de pondération entre la similarité structurelle et sémantique. La valeur de similarité finale entre deux schémas est normalisée par le nombre des schémas dérivés. (VOIGT, 2011) étudie l'apport de la théorie des graphes pour améliorer la qualité des solutions du problème d'appariement des métamodèles. L'approche proposée utilise les graphes planaires pour représenter les métamodèles qui sont des graphes où il n'y a pas de croisement d'arêtes (autrement dit, le graphe est plongé dans un plan en donnant des coordonnées 2D à chaque nœud) et des techniques de fouille des graphes (Graph Mining) pour extraire un ensemble de sous-graphes qui partagent les mêmes structures entre deux graphes. L'approche utilise des mesures de similarité syntaxiques et structurelles en exploitant des informations présentes sur des arêtes.

(BERRO, MEGDICHE et TESTE, 2015) étudient le problème d'appariement holistique en le modélisant comme un modèle d'optimisation linéaire où les schémas sont représentés par

5. Voir (NGO et BELLAHSENE, 2016) pour d'autres alternatives.

des graphes orientés attribués. Les contraintes du modèle englobent les contraintes de cardinalités *(1 : 1)*, contraintes de seuils et contraintes structurelles propres à la hiérarchie des schémas. Les mesures de similarité appliquées concernent les mesures de similarité linguistiques et la fonction d'agrégation maximum est sélectionnée. (HÄTTASCH et al., 2022) proposent une approche pour l'appariement des schémas de bases de données basée sur l'utilisation du plongement lexical pour représenter les noms des tables, des attributs et des instances de données sous forme de vecteur de nombres réels. Cette représentation remplace un mot ou une chaîne de caractères par un vecteur de telle sorte que deux mots issus du même contexte et exprimant le même sens (ou ont un sens proche) sont représentés par deux vecteurs proches en s'appuyant sur des données précédemment entrainées, par exemple, les vecteurs des termes *roi* et *reine* devraient être proches. Le processus d'appariement est réalisé en deux étapes, la première étape concerne l'appariement des tables et la seconde étape concerne l'appariement des attributs. Pour la première étape, deux approches sont proposées : la première utilise les vecteurs des noms des tables et des attributs et la deuxième exploite le contenu des tables, à savoir, les instances de données en regroupant toutes ou une partie des attributs et en comparant les fréquences d'occurrences des données. Les expériences sont menées sur un problème d'appariement de base de données et les deux approches peuvent être utilisées de manière complémentaire. Dans la seconde étape, uniquement les noms des attributs ou les instances de données sont utilisés moyennant différentes techniques de plongements simples et contextualisés où les expériences sont menées sur des ontologies.

(KANG et NAUGHTON, 2003) présentent une approche se basant sur les instances de données pour détecter les correspondances en présence de noms d'entités opaques. L'approche se déroule en deux étapes. Dans une première étape des corrélations entre les éléments des schémas sont calculées et un graphe de dépendance est construit. Ensuite, dans une deuxième étape, un algorithme d'appariement de graphes est appliqué afin de trouver les paires correspondantes aux deux schémas. (MAZILU et al., 2022) propose une méthodologie d'appariement des schémas où les relations entre les éléments ne sont pas déclarées. Les auteurs proposent un algorithme de programmation dynamique qui exploite les informations dans les schémas traduit en graphe simple orienté et des sources auxiliaires pour produire des caractérisations statistiques. Des règles sont utilisées afin de propager ces statistiques, à savoir, l'identification des clés et les dépendances fonctionnelles. À la fin plusieurs candidats potentiels sont sélectionnés grâce à un modèle d'optimisation en considérant deux types de mesures de similarité.

2.7.6 Synthèse et analyse de l'état de l'art

Il est important de noter la qualité des efforts qui ont été et qui sont toujours menés dans le but d'améliorer et contribuer à la résolution des problèmes d'appariement. Plus particulièrement, les approches présentées ci-dessus, varient dans leurs processus qui peuvent être **itératifs** (POULIOT et al., 2018) ou par étapes, où un certain nombre de **tâches** sont définies (SORRENTINO et al., 2009). Les schémas en entrées (source(s) et cible(s)) sont **homogènes** ou **hétérogènes** à plusieurs niveaux, selon ce qu'ils représentent (base de données, base de connaissances, bibliothèque numérique, etc.) ou par nature (schéma, ontologie ou métamodèle (IVANOV et VOIGT, 2011)), le modèle de données (XML, SQL, CSV, RDF, OWL, etc.), le schéma de données (syntaxe, structure) ou l'instance des données (syntaxe, sémantique). Ces schémas sont pris par paire ou dans un réseau en considérant des **contraintes** internes spécifiques aux schémas ou à l'appariement.

Pour évaluer la **similarité**, des **mesures** sont généralement appliquées. Des techniques issues du **traitement du langage naturel** sont utilisées pour les propriétés linguistiques des éléments des schémas représentés par des caractères (PRADHAN, GYANCHANDANI et WADHVANI, 2015) ou par des vecteurs en utilisant le plongement lexical (CAPPUZZO, PAPOTTI et THIRUMURUGANATHAN, 2020). D'autres techniques sont employées pour les

instances de données afin d'améliorer la précision du calcul de similarité comme l'**analyse** (J. WANG et al., 2004) ou les **statistiques** (DING, H. DONG et G. WANG, 2012). Des techniques d'apprentissage sont utilisées pour former d'autres modèles et fournir une autre **interprétation des informations** (MAHMOOD, FAROOQ et FERZUND, 2017).

Pour la **modélisation** des schémas, la **théorie des graphes** est utilisée de différentes manières et est suggérée dans plusieurs travaux (GAL et al., 2005) ; (HUNG et al., 2019) ; (SAHA, STANOI et CLARKSON, 2010). Les techniques de partitionnement sont appliquées pour réduire et limiter l'espace de recherche des correspondances (SAHAY, MEHTA et JADON, 2020) et les probabilités sont utilisées pour rendre le processus d'appariement aussi indépendant que possible des paramètres et des interventions humaines (C. J. ZHANG et al., 2013).

Différents paramètres peuvent être fixés, par des algorithmes génétiques (GULIĆ, VRDOLJAK et PTIČEK, 2018). Selon leur disponibilité, les précédents **résultats** d'appariement sont **réutilisés** (HEYVAERT et al., 2017) et des **stratégies** d'agrégation, sélection ou de combinaison sont employées comme les réseaux de neurones (Y. LI, D.-B. LIU et W.-M. ZHANG, 2005) ou les techniques d'optimisation (KIM et al., 2011) ; (SMILJANIĆ, KEULEN et JONKER, 2005).

Des **outils** sont développés pour les différentes phases du problème d'appariement où le champ de recherche reste ouvert (J. WANG, B. GUO et L. CHEN, 2022). Ces outils sont automatiques (DRUMM et al., 2007) ou semi-automatiques (VOIGT, IVANOV et RUMMLER, 2010), avec interface utilisateur graphique (GUI) pour aider l'utilisateur à choisir les métriques à utiliser dans le processus, à valider les résultats ou parfois assistés par des évaluations (BELLAHSENE et DUCHATEAU, 2011) ; (KOUTRAS et al., 2021). Des mesures du temps, de la qualité des résultats, de l'impact des techniques utilisées et le degré de l'effort humain sont ainsi mesurés (DO, MELNIK et RAHM, 2002) ; (SHRAGA et GAL, 2022).

2.7.7 Discussion

Les variations citées précédemment : (1) rendent le processus d'évaluation des approches et outils d'appariement difficile à évaluer. Des schémas de référence accompagnés de l'ensemble des correspondances attendues sont parfois proposés (BELLAHSENE et al., 2011). On retrouve ainsi des benchmarks spécifiques pour évaluer des outils d'appariement (CRESCENZI et al., 2021), des schémas XML (DUCHATEAU, BELLAHSENE et HUNT, 2007), des ontologies (OAEI, 2022) ou encore des schémas de base de données (C. GUO et al., 2013) ; (2) accentuent le besoin d'avoir une méthodologie flexible (PEUKERT, EBERIUS et RAHM, 2011), globale (PATEL, DEBNATH et BHUSHAN, 2022) et générique qui peut être adaptée à autant de points de variation que possible mentionnés ci-dessus (A. ALGERGAWY, NAYAK et SAAKE, 2010) ; (MADHAVAN, BERNSTEIN et RAHM, 2001). **Ainsi, la méthodologie qui sera présentée dans la suite du manuscrit, aura l'ambition de contribuer à ces besoins et ceci en se basant sur l'utilisation des principes de la théorie des graphes, des modèles d'optimisation et des techniques de traitement automatique des langues.**

Les approches basées sur l'apprentissage ne peuvent pas toujours être appliquées, car les modèles entrainés sont très spécifiques et les données d'entrainement ne sont pas toujours disponibles en quantité suffisante, pour des cas industriels où les informations en termes de sémantique et de syntaxe (sens et forme) sont difficiles à interpréter. Dans d'autres cas, un ensemble bien structuré de termes et de concepts, à l'image des bases de connaissances, les ontologies ou les métamodèles représentant la signification des connaissances, sont construits.

En s'appuyant sur les conclusions du Chapitre 1 et dans la vision d'une méthode qui s'adapte à la volée, l'utilisation de ces concepts reste limitée à des domaines ou des contextes précis, même si des ontologies ou des métamodèles dynamiques peuvent être générique, négociés et créés, un effort humain est parfois nécessaire pour reconnaître et interpréter des informations. En effet, la création de ces concepts est généralement basée sur trois étapes (CRISTANI et

CUEL, 2005) : (1) le recueil, (2) la modélisation et (3) la représentation des connaissances et se fait sur des informations qui ne sont pas toujours interprétables à la volée. Ces tâches restent difficiles à entièrement automatiser, surtout quand il s'agit de connaissances liées au domaine industriel ou l'IoT, où l'automatisation de l'interprétation ou même l'accès à certaines connaissances sont restreints. Adopter une approche fondée sur un accord, ou une approche basée sur des médiateurs, ou une approche fédérée basée uniquement sur une ontologie ou un métamodèle ne semble pas répondre à toutes les exigences (HUNG et al., 2019), surtout dans des environnements changeants. Il est alors nécessaire de disposer (1) d'une approche permettant d'intégrer diverses sources d'interprétation pour pallier l'absence ou l'impossibilité d'utiliser des sources d'information (en plus des bases de connaissances), (2) disposer d'une autre abstraction et modélisation des connaissances. Ces deux objectifs peuvent être approchés par la théorie des graphes et les modèles d'optimisation.

Les approches qui utilisent des graphes traduisent les schémas souvent structurés en graphe simple où une relation (arête ou arc) ne relie qu'une paire d'éléments. Une relation d'un hypergraphe peut relier des groupes de plus de deux éléments. Ainsi, nous pouvons représenter par exemple les schémas des bases de données comme un hypergraphe dans lequel nous avons plusieurs groupes d'éléments partageant la même information (NGUYEN, 2014) ; (SAHA, STANOI et CLARKSON, 2010). En effet, les hypergraphes ont l'avantage de réduire le degré de complexité et d'augmenter le niveau d'abstraction des relations entre les éléments des schémas.

La théorie des graphes, au-delà de ces avantages en termes de représentations visuelles, est aussi un outil de modélisation, d'intégration et de découvertes d'interactions complexes entre des éléments. Ainsi, les graphes offrent cette possibilité d'**adaptation** à plusieurs types de schémas **à la volée** à condition de savoir traduire les informations en entrée (MEGDICHE, TESTE et TROJAHN, 2016). La résolution du problème d'appariement des schémas est alors réduite à la résolution du problème d'appariement des graphes. La traduction mathématique de ce problème revient à résoudre un modèle d'optimisation ou des variables de décisions et des contraintes adaptées sont identifiées. Ainsi, le modèle d'optimisation accueillera les différentes *contraintes* et paramètres du problème d'appariement : **flexibilité**. Il permettra de s'adapter aux problèmes d'appariement par paire et se chargera du processus de **sélection** des correspondances grâce aux *variables de décisions* : **globalité**. Enfin, il sera extensible pour inclure divers types de mesures de similarité par le biais de la *fonction objectif* : **généricité**.

2.8 Conclusion

Dans ce chapitre, nous avons présenté le problème d'interopérabilité des données en général et plus particulièrement sa déclinaison en un problème d'appariement des schémas. Résoudre les problèmes d'appariement des schémas permet d'établir des relations entre des schémas source(s) et cible(s), ainsi, la réalisation de cette tâche permettra de mettre en œuvre l'interopérabilité des données entre ces schémas. Aussi, le constat de l'état de l'art présenté se traduit par des besoins en termes de flexibilité, globalité et de généricité. Ainsi, ces verrous sont adressés par l'utilisation du traitement automatique des langues pour le calcul des similarités linguistiques, l'optimisation pour les processus de sélection des correspondances et la théorie des graphes pour la modélisation et la compréhension du problème d'appariement. Ces concepts, outils et méthodes offrent alors des leviers d'automatisation, d'adaptabilité et de transformation à la volée. L'approche que nous proposons n'est pas exclusive à l'appariement des schémas, mais à toutes problématiques dont l'objectif est de modéliser et retrouver des relations entre des concepts et d'optimiser le processus de mise en relation. Dans le chapitre suivant, nous présenterons les étapes de mise en œuvre de cette approche et ses principales contributions face aux besoins relevés dans le Chapitre 1 et nous démontrerons les motivations et l'apport de la théorie des graphes et l'optimisation comme concepts à utiliser pour établir l'interopérabilité fédérée et assurer ses caractéristiques. Enfin,

moyennant en plus le traitement automatique des langues, nous montrerons l'apport de l'approche aux besoins des problèmes d'appariements relevés dans ce chapitre.

3

Contribution à la résolution du problème d'appariement : approche flexible, globale et générique

3.1 Introduction

Le problème d'appariement est l'une des principales tâches dans plusieurs processus traitant des structures de données, par exemple les bases de connaissances, les bases de données, les entrepôts de données ou les comptoirs de données dans lesquels des mécanismes de migration sont nécessaires. En effet, l'appariement implique la recherche des liens entre des schémas hétérogènes. Cette hétérogénéité réside à plusieurs niveaux, il est alors nécessaire de disposer d'une méthodologie capable de s'adapter au plus grand nombre de points de variation.

Comme analysé dans le Chapitre 2, les approches qui traitent les problèmes d'appariement suivent les mêmes phases, mais chaque phase peut avoir ses propres étapes où plusieurs techniques et stratégies diversifiées sont utilisées. L'objectif de ce chapitre n'est pas de présenter une énième méthode d'appariement des schémas. L'objectif est de proposer un unique environnement homogène qui peut accueillir les différentes phases (modélisation, agrégation et sélection), les techniques (mesures de similarité), et les contraintes (de seuil, de cardinalité ou de structure) des problèmes d'appariement par l'utilisation de la théorie des graphes et l'optimisation et adresser un verrou vers une automatisation des outils. À cette fin, la méthodologie exploite les informations disponibles dans les deux schémas (source et cible) dans le cas d'un processus de migration des bases de données avec l'hypothèse que **les instances de données dans le schéma cible ne sont pas disponibles**.

Le chapitre sera organisé en deux parties, la première partie est présentée par la Section 3.2 qui expliquera la modélisation des schémas des bases de données par les hypergraphes et la réduction du problème d'appariement des schémas au problème d'appariement d'hypergraphes. La seconde partie est présentée par la Section 3.3 décrira l'approche générale. Dans un premier temps, les étapes principales seront données, ensuite, la Section 3.3.1 présentera les différentes mesures de similarité utilisées pour calculer les matrices de similarité, puis Section 3.3.2 donnera les stratégies d'agrégation et de composition utilisées pour agréger et composer les matrices de similarité. La Section 3.3.3 présentera le modèle d'optimisation général qui sera utilisé pour la sélection des correspondances et inclura les variations et contraintes d'appariement. Enfin, la Section 3.4 conclura le chapitre.

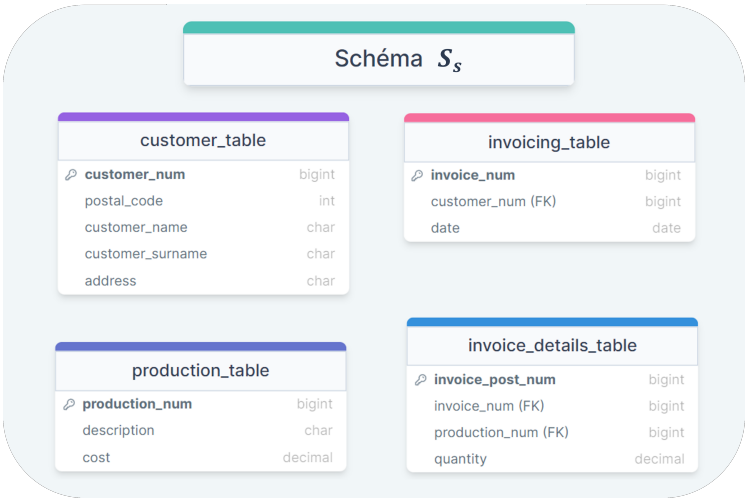
3.2 Appariement des schémas par l'appariement d'hypergraphes

Dans cette section nous traitons le problème d'appariement par paire par une approche qui combine la théorie des graphes et les modèles d'optimisation. En effet, raisonner sur les graphes permet non seulement de construire des relations entre des schémas modélisés, mais aussi d'effectuer des transformations, optimiser les calculs de similarités et les procédures de sélection ou d'agrégation.

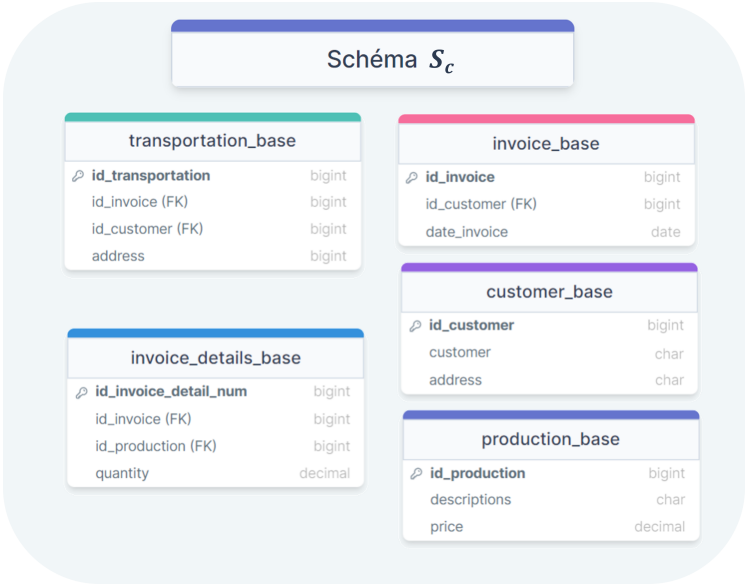
3.2.1 Modélisation des schémas par les hypergraphes

> Pour mieux illustrer, nous proposons deux schémas de base de données simples S_s et S_c présentés dans la Figure 3.1 qui décrivent les structures de deux bases de données relationnelles représentant des informations de facturation.

Les schémas sont composés de plusieurs tables caractérisées par un nom et chaque table est constituée d'un ensemble d'attributs caractérisés par un nom et un type. Les attributs marqués avec l'icône clé représentent des clés primaires (PK pour Primary Key). Cette clé identifie chaque tuple ou ligne de manière unique. Ainsi, deux lignes d'une table ne peuvent avoir la même valeur de la clé primaire. Les attributs contenant la mention FK correspondent à des clés étrangères (FK pour Foreign Key) font référence (par définition) à des clés primaires dans d'autres tables. Chaque attribut a un type de données et l'ensemble des types de données constituent le domaine. Les notions de clé primaire, clé étrangère et le domaine sont importants pour la cohérence des données et des informations que véhicule la base de données (ATENCIA, DAVID et SCHARFFE, 2012). Le choix et l'identification de ce type de contraintes préserve l'intégrité des données (CODD, 2007). Cependant, dans de nombreux cas réels, des ensembles de données sont structurés en tables, mais sont construites sans définition explicite de ces notions.



(a) Schéma S_s d'une base de données source



(b) Schéma S_c d'une base de données cible

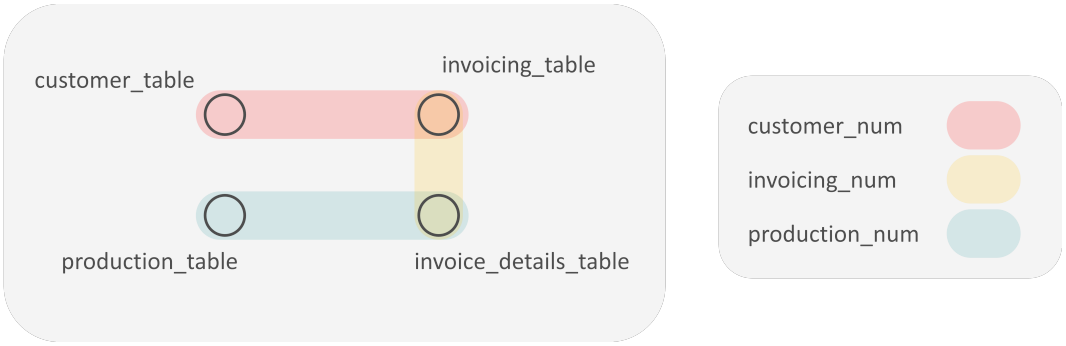
FIGURE 3.1 – Schémas S_s et S_c représentant deux bases de données relationnelles.

La modélisation des schémas par les hypergraphes se basent notamment sur la notion de relation, offerte dans le cas des bases de données relationnelles par la notion de clé. L'absence de schéma ou le manque de relation dans les bases de données non-relationnelles peuvent être traités par des techniques de profilage (ABEDJAN, GOLAB et NAUMANN, 2015), des méthodes de rétro-ingénierie (F. SHI et al., 2018) ou des techniques d'extraction des schémas (CASTELLTORT et LAURENT, 2017). Ne constituant pas l'objectif principal de cette thèse, nous n'allons pas aborder en détails ces techniques, néanmoins, il demeure important dans les problèmes d'appariement d'être capable de pallier l'absence d'informations, de relation, de données.

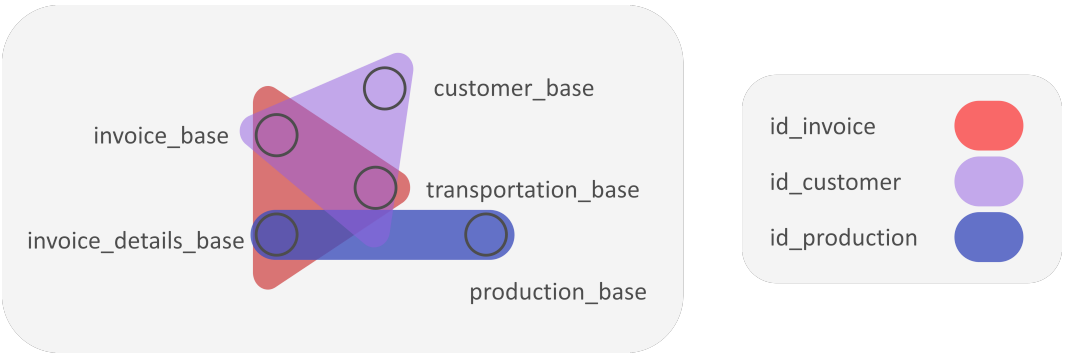
Les hypergraphes sont utilisés dans plusieurs applications relatives à la gestion des données comme l'intégration (MASMOUDI et al., 2021), la détection des fréquences ou de structures récurrentes (M. ALAM et al., 2021) ; (D. LI, LAURENT et TEISSEIRE, 2007). Pour les bases de

données relationnelles, l'utilisation des hypergraphes facilite non seulement la représentation visuelle (BRETTO, 2013) mais aussi la modélisation et la résolution de certains problèmes en les transformant en problèmes connus dans les hypergraphes (GHALEB et al., 2020). Pour les problèmes d'appariement, les hypergraphes apportent deux principaux avantages, (1) une modélisation plus précise, et à un haut niveau formalisation permettant, une meilleure représentation de certaines relations complexes et (2) une approche de résolution optimisant les résultats globaux en réduisant le problème d'appariement des schémas en un problème d'appariement d'hypergraphes.

En retournant à l'exemple de la Figure 3.1, nous traduisons les deux schémas des bases de données relationnelles en deux hypergraphes attribués : l'hypergraphe source $HG_s = (V_s, HE_s, \lambda_s)$ et l'hypergraphe cible $HG_c = (V_c, HE_c, \lambda_c)$ illustrés dans la Figure 3.2 où les nœuds de V_s et V_c représentent les tables et contiennent les informations relatives aux tables et à leurs attributs qui sont portés par les fonctions λ_s et λ_c. Les propriétés des hypergraphes sont résumées dans le Tableau 3.1 où les (hyper)arêtes simplifiées correspondent aux noms des clés.



(a) Hypergraphe source HG_s et les (hyper)arêtes associées



(b) Hypergraphe cible HG_c et les (hyper)arêtes associées

FIGURE 3.2 – Hypergraphes HG_s et HG_c représentant les deux schémas S_s et S_c.

L'hypergraphe HG_s est composé quatre nœuds et de trois (hyper)arêtes. Chaque (hyper)arête correspond à une clé étrangère qui regroupe deux nœuds. L'hypergraphe HG_c est composé cinq nœuds et de trois (hyper)arêtes. Chaque (hyper)arête correspond à une clé étrangère et regroupe deux ou trois nœuds. Une (hyper)arête est construite d'une manière à ce qu'elle contient la table de la clé primaire (dans certains cas la clé unique ou UK pour Unique Key) et les tables de la clé étrangère associée. Les nœuds sont donc reliés par les ensembles d'(hyper)arêtes HE_s et HE_c. À noter que pour des raisons de simplification, nous avons

Hypergraphes	$HG_s = (V_s, HE_s, \lambda_s)$	♯	$HG_c = (V_c, HE_c, \lambda_c)$	♯
Nœuds	{invoicing_table ; customer_table ; production_table ; invoice_details_table}	**4**	{invoice_base ; customer_base ; production_base ; invoice_details_base ; transportation_base}	**5**
(Hyper)arêtes	{[invoicing_table, customer_table] ; [invoicing_table, invoice_details_table] ; [production_table, invoice_details_table]}	**3**	{[invoice_base, customer_base, transportation_base] ; [invoice_base, invoice_details_base, transportation_base] ; [production_base, invoice_details_base]}	**3**
Clés	{[customer_num] ; [invoicing_num] ; [production_num]}	**3**	{[id_customer] ; [id_invoice] ; [id_production]}	**3**

TABLEAU 3.1 – Caractéristiques des hypergraphes HG_s et HG_c

gardé le même nom des clés primaires dans les autres tables où elles apparaissent comme clés étrangères. Chaque nœud contiendra ensuite le nom de la table, le nom et type de données de ces attributs ainsi que d'autres types d'information ou de contraintes. Par exemple, la Figure 3.3 représente les détails que porte le nœud *customer_table* de l'hypergraphe HG_s.

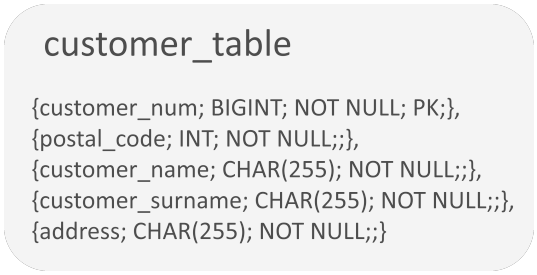


FIGURE 3.3 – Nœud *customer_table* de l'hypergraphe HG_s.

Les deux hypergraphes sont hétérogènes en termes de structure c'est-à-dire nombre de nœuds, nombre d'(hyper)arêtes et informations sur les tables et les attributs données par les fonctions λ_s et λ_c. Les hypergraphes ne sont pas typés : les nœuds où les (hyper)arêtes n'ont pas de type, mais seulement des attributs.

Dans une modélisation plus détaillée, on peut considérer les dépendances fonctionnelles entre les attributs d'une table. Ces dépendances sont utilisées dans le cadre de la normalisation (restructuration) des bases de données relationnelles afin d'assurer une intégrité des données. Ces dépendances peuvent être modélisées sous forme d'un graphe simple orienté où le nombre de nœuds est le nombre d'attributs et le nombre d'arcs correspond aux nombres de dépendances fonctionnelles.

Par exemple, en reprenant le nœud *customer_table* de l'hypergraphe HG_s, les dépendances fonctionnelles peuvent être représentées dans la Figure 3.4.

Dans cette table, l'attribut *customer_num* est la clé primaire et les autres attributs (*postal_code*, *customer_name*, *customer_surname* et *address*) sont des attributs non clés. Les dépendances fonctionnelles s'écrivent : $\{customer\_num \rightarrow postal\_code; customer\_num \rightarrow customer\_name; customer\_num \rightarrow customer\_surname; customer\_num \rightarrow address\}$.

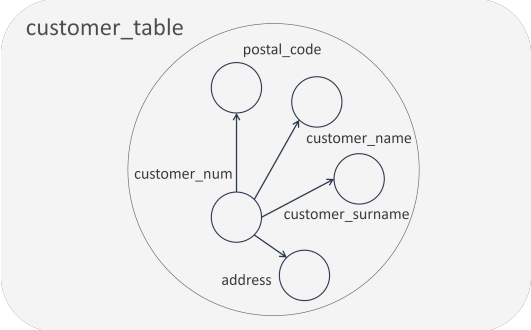


FIGURE 3.4 – Nœud *customer_table* de l'hypergraphe HG_s avec les dépendances fonctionnelles.

L'attribut gauche de la relation, détermine de manière unique la valeur de l'attribut droit de la relation. Pour trouver ces dépendances fonctionnelles, des algorithmes sont proposés (PAPENBROCK et al., 2015).

En revenant au problème d'appariement des schémas modélisés par les hypergraphes, ce dernier consiste alors à trouver les meilleures correspondances possibles entre les deux hypergraphes. Une correspondance entre les tables des schémas S_s et S_c revient à trouver les correspondances entre les nœuds des hypergraphes HG_s et HG_c.

3.2.2 Appariement des schémas par l'appariement d'hypergraphes

Comme nous l'avons vu dans Section 2.7.5, différents types de graphes (graphe simple, hypergraphe, arbre, graphe ou hypergraphe attribué, etc.) ont été proposés et utilisés dans la littérature afin de modéliser des schémas. Dans ce qui suit, nous allons décrire les caractéristiques de l'appariement d'hypergraphes. Les propriétés et approches décrites dans la littérature pour cette catégorie de graphes restent pour la plupart une généralisation ou une extension de celles issues de la recherche sur les graphes simples et, en général, pour résoudre le problème de l'appariement d'hypergraphes, les techniques utilisées dans l'appariement de graphes sont adaptées (LIAO, Y. XU et LING, 2021). Il faut noter que l'appariement dans les hypergraphes est une autre discipline où l'objectif est de trouver un sous-ensemble de nœuds dans l'hypergraphe qui satisfait à certaines contraintes, telles que la taille ou la connectivité des nœuds dans le sous-ensemble (KEEVASH et MYCROFT, 2014).

En général, l'appariement de graphes est le problème qui consiste à trouver les correspondances entre deux ensembles de nœuds tout en préservant les relations qui peuvent exister entre les nœuds. Sans perte de généralité, pour les hypergraphes, le problème d'appariement d'hypergraphes consiste à trouver des correspondances de nœuds entre deux hypergraphes (ZASS et SHASHUA, 2008).

Ainsi, pour les deux hypergraphes HG_s et HG_c de la Figure 3.2, la modélisation du problème d'appariement se fait généralement à l'aide d'un graphe biparti $G = (V, E)$ où $V = V_s \cup V_c$ représente l'ensemble des attributs de HG_s et HG_c (deux partitions) et une arête $(u, v) \in E$ avec $u \in HG_s$ et $v \in HG_c$ représente une correspondance possible entre les deux nœuds u et v. Une solution au problème d'appariement renvoie un sous-ensemble d'arêtes $E' \subseteq E$.

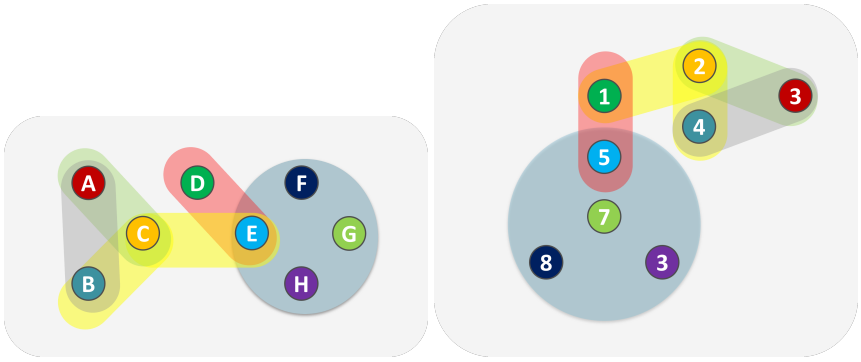
Deux aspects de l'appariement d'hypergraphes sont connus, l'*appariement exact*, comme son nom l'indique, vise à trouver si deux hypergraphes (ou une partie des hypergraphes) sont similaires ou non, tandis que l'*appariement inexact* propose une évaluation sur le degré de similarité ou de dissimilarité entre deux hypergraphes (ou une partie des hypergraphes).

Appariement exact : l'appariement exact est caractérisé par le fait qu'il préserve les propriétés des hypergraphes.

Pour mieux illustrer, nous repartons de l'hypergraphe $HG = (V, HE)$ de la Figure 2.9a et ajoutons aux nœuds des attributs indiqués par des couleurs. L'hypergraphe est alors un hypergraphe attribué $HG = (V, HE, \lambda)$. Considérons A une matrice d'incidence de taille $n \times m = 8 \times 5$ associée à l'hypergraphe HG où a_{ij} égale à 1 si le nœud i appartient à l'(hyper)arête j et égale à 0 sinon avec $\mid V \mid = n = 8$, $\mid HE \mid = m = 5$.

$$A = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{array}{c} \begin{array}{ccccc} \color{gray}{\bullet} & \color{lightgreen}{\bullet} & \color{yellow}{\bullet} & \color{lightcoral}{\bullet} & \color{lightblue}{\bullet} \end{array} \\ \begin{pmatrix} he_1 & he_2 & he_3 & he_4 & he_5 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Pour un même hypergraphe, cette matrice n'est pas unique, les nœuds peuvent être ordonnés de $n!$ ou $m!$ façons, on peut donc obtenir plusieurs matrices d'incidence. À cet effet, pour comparer un autre hypergraphe HH avec HG, il est nécessaire de définir une fonction qui permet de trouver les correspondances exactes non seulement en termes de structure, ce qu'on appelle *edge-preserving*, mais aussi en termes d'attributs, ce qu'on appelle *attributes-coherence* (Livi et Rizzi, 2013). Ensuite, si deux nœuds sont reliés par une (hyper)arête dans le premier hypergraphe, ils seront mis en correspondance avec deux autres nœuds également reliés par une (hyper)arête dans le second hypergraphe et cette correspondance bijective utilise alors la définition d'*isomorphisme d'hypergraphe*. Un exemple d'isomorphisme est illustré dans la Figure 3.5.



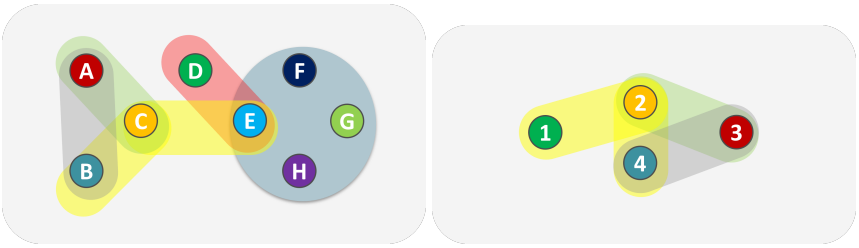
(a) Hypergraphe HG **(b)** Hypergraphe HH

Figure 3.5 – Hypergraphe HG isomorphe à l'hypergraphe HH.

Définition 3.2.1. Deux hypergraphes attribués $HG = (V, HE, \lambda)$ et $HH = (U, HX, \gamma)$ avec $\mid V \mid = \mid U \mid$ sont isomorphes, et on écrit $HG \simeq HH$ si étant donné une fonction bijective d'isomorphisme d'hypergraphe $f : HG \rightarrow HH$ qui satisfait les propriétés suivantes (L. Wang, Egorova et Mokryakov, 2018) :

— $\forall v \in V : \exists u = f(v) \in U$,
— $\forall he = (v, \cdots, w) \in HE, \exists hx \in HX : hx = (f(v), \cdots, f(w))$,
— $\forall hx = (u, \cdots, z) \in HX, \exists he \in HE : he = (f^{-1}(u), \cdots, f^{-1}(z))$,
— $\lambda(u) = \gamma(f(u))$.

L'isomorphisme des graphes établit alors des correspondances entre chaque nœud des deux hypergraphes, c.-à-d., cela revient à trouver des correspondances *(1 : 1)*. La complexité du problème d'isomorphisme des graphes reste une question ouverte (R. MILLER, 2013). Le problème est généralement considéré dans la classe $\mathcal{NP}$ (GAREY, 1979) : il n'existe pas un algorithme polynomial pour le résoudre, sauf dans des cas précis avec un certain type de graphe (ARVIND et KÖBLER, 2006). Pour les hypergraphes, la question est, elle aussi, ouverte : une généralisation du problème d'isomorphisme des graphes est déduite par la réduction en temps polynomial d'un hypergraphe en un graphe biparti où le premier ensemble contient les nœuds et le deuxième les (hyper)arêtes. Une autre forme moins restreinte consiste à trouver un isomorphisme de sous-hypergraphe (CORDELLA et al., 2004) illustré dans la Figure 3.6.



(a) Hypergraphe HG **(b)** Sous-hypergraphe HG'

FIGURE 3.6 – Sous-hypergraphe HG' isomorphe à une partie de l'hypergraphe HG.

Définition 3.2.2. Soit $HG = (V, HE, \lambda)$ un hypergraphe et $HG' = (V', HE', \lambda')$ un sous-hypergraphe de HG avec $V' \subseteq V$ et $HE' \subseteq HE$. Il existe une fonction injective $f : HE \to HE'$ s'il existe un hypergraphe $HG' \subseteq HG$ tel que f est un isomorphisme de sous-hypergraphe entre HG and HG'.

Dans des cas plus restreints du monde réel, l'appariement exact n'est pas toujours possible (RIESEN, X. JIANG et BUNKE, 2010). En effet, la recherche de correspondances ne se limite pas seulement à la structure, mais également aux informations contenues et supportées par les nœuds et notamment les (hyper)arêtes, et ces structures et informations sont le plus souvent hétérogènes, il est donc nécessaire de permettre des correspondances approximatives évaluées par une mesure avec une certaine tolérance aux erreurs. Dans notre cas, on parle de mesure de similarité (BUNKE, X. JIANG et KANDEL, 2000).

Appariement inexact : le problème de l'appariement inexact des hypergraphes est plus flexible (et réaliste) que l'appariement exact. Une fonction de coût ou une fonction d'évaluation est employée afin de quantifier le degré de similarité. Ainsi, les algorithmes utilisés pour trouver les correspondances tentent d'optimiser cette fonction. La distance d'édition est la mesure de base et l'une des plus utilisées (SANFELIU et FU, 1983) qui consiste à trouver le nombre minimum d'opérations élémentaires (suppression, insertion ou substitution) de

nœuds et/ou d'(hyper)arêtes afin de transformer un hypergraphe en un autre (GAO et al., 2010). En général, pour les différents types de graphe, des méthodes et des algorithmes sont proposés dans la littérature et sont examinés dans (AGGARWAL, H. WANG et al., 2010) ; (CONTE et al., 2004) ; (S. P. DWIVEDI, 2020) ; (FOGGIA, PERCANNELLA et VENTO, 2014) ; (GALLAGHER, 2006) ; (GAO et al., 2010) ; (LAURA, WESARG et SAKAS, 2022) ; (LIVI et RIZZI, 2013) ; (MADI, 2016) ; (SUN, ZHOU et FEI, 2020). Ces algorithmes sont souvent basés sur l'optimisation par le biais de résolution de modèles d'optimisation par des méthodes exactes ou d'heuristiques, spécifiques et efficaces dans le domaine d'application où elles ont été conçues (GAO et al., 2010) et sont adaptables en fonction de la structure des graphes choisie. Par conséquent, ces algorithmes utilisent divers paramètres qui offrent une certaine flexibilité. Un défi majeur pour ces algorithmes est la complexité pour traiter des graphes ou des hypergraphes à grande échelle (LIVI et RIZZI, 2013), néanmoins, les avancées en termes de théorie des graphes et d'informatique ne cessent d'évoluer et stimuler les travaux de recherche dans ce domaine pour pallier certains de ces inconvénients (GHANSHYAMBHAI et GHODASARA, 2020). Enfin, l'aspect incertitude est également étudié dans certains travaux (YAN et al., 2016) et est utilisé au moyen de probabilités ou de fonctions. En outre, l'incertitude est souvent présente en raison de l'hétérogénéité des données et la précision est susceptible d'être faussée, par conséquent, la prise en compte de ce paramètre offre un mécanisme d'évaluation qui aide à identifier les correspondances potentielles.

3.3 Conception et architecture de l'approche proposée

Nous allons décrire l'approche que nous proposons, illustrée dans la Figure 3.7, pour l'appariement des schémas dans le cas d'une migration de données. Nous considérons la situation où nous avons deux bases de données, c'est-à-dire un schéma source et un schéma cible. Comme nous l'avons vu précédemment, les différentes utilisations de la théorie des graphes et de l'optimisation offrent des capacités à s'adapter à différentes problématiques et à accueillir plusieurs paramètres et contraintes spécifiques. Par notre approche, nous voulons offrir un environnement homogène permettant d'optimiser l'évaluation générale. Dans cette section, ces deux concepts (la théorie des graphes et l'optimisation) seront les deux premières briques de notre approche.

Dans la première étape, l'approche propose tout d'abord d'extraire les informations des deux schémas (par exemple deux schémas en format *SQL*) et de les modéliser en hypergraphes en traduisant les informations extraites. À cette fin, les nœuds des hypergraphes représentent les tables et les attributs d'un nœud représentent le nom de la table et ses informations. Les (hyper)arêtes représentent la notion de clés étrangères et construisent donc la structure des hypergraphes. Les deux hypergraphes sont disjoints, c'est-à-dire, qu'il n'y a aucune relation entre leurs nœuds. L'objectif est donc de construire ces relations.

Ensuite, dans la deuxième étape et afin d'alimenter le modèle d'optimisation, une matrice résumant la similarité des différentes dimensions prises en compte doit être calculée. Les dimensions concernent la sémantique, la syntaxe, les types de données et la structure. Avec l'hypothèse de l'indisponibilité des instances de données dans la base de données cible, la similarité entre les instances de données n'est pas prise en compte. En effet, dans certains cas de migration, les instances de données ne sont pas toujours présentes ou accessibles dans le schéma cible. Ceci crée alors une forme de déséquilibre quand on procède au calcul de la similarité basée sur les instances de données.

Les méthodes du traitement du langage naturel seront la troisième brique de notre approche. Le calcul de la similarité se fait sur trois niveaux.

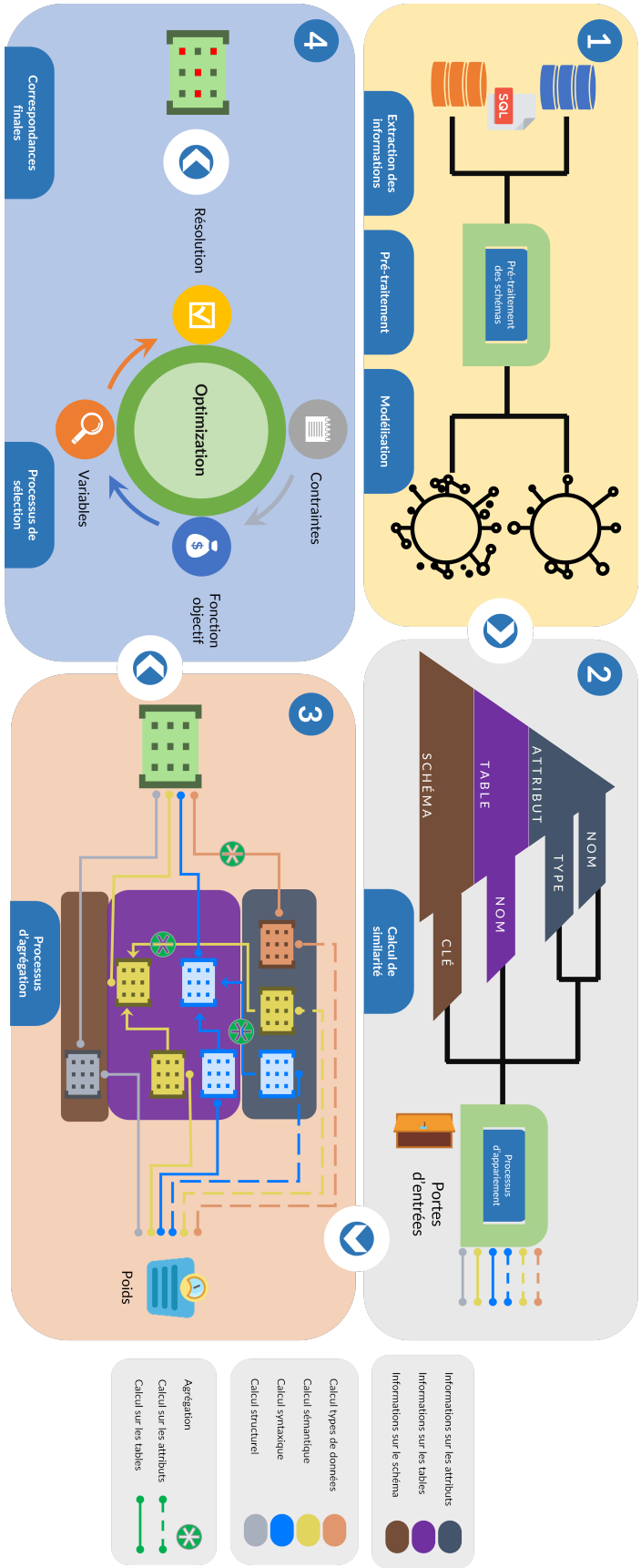


FIGURE 3.7 – Vue générale de l'approche proposée.

1. **Niveau des attributs** : (informations contenues dans les nœuds des hypergraphes) en utilisant la sémantique, la syntaxique et les types de données des attributs. Ainsi, pour chaque paire de nœuds entre les deux hypergraphes, trois matrices de similarité sont calculées et la taille des matrices est fonction du nombre d'attributs de chaque nœud. Ensuite, chaque matrice est agrégée pour donner une seule valeur en utilisant des algorithmes d'agrégation. En somme, pour chaque paire de nœuds (paire de tables), on aura trois valeurs de similarité basées uniquement sur les attributs.
2. **Niveau des tables** : où nous utilisons les informations sémantiques et syntaxiques des noms des tables pour calculer deux matrices de similarité. Ici, la taille des matrices est fonction du nombre d'attributs. Ensuite, pour chaque paire de nœuds, une somme pondérée est utilisée pour combiner les valeurs sémantiques et syntaxiques obtenues au niveau précédent avec celles obtenues au niveau actuel.
3. **Niveau des schémas** : où nous utilisons les clés étrangères pour construire la structure des hypergraphes. Ensuite, nous calculons la matrice de similarité structurelle.

On obtient quatre matrices de similarité correspondant aux quatre dimensions (structurelle, sémantique, syntaxique et types de données) qui sont utilisées pour obtenir la matrice finale servant d'entrée au modèle d'optimisation. L'agrégation des matrices est donc la troisième étape de l'approche.

Pour la quatrième et dernière étape, le problème d'appariement d'hypergraphes est traduit en un modèle d'optimisation où l'on maximise une fonction objectif exprimée sous forme d'une somme pondérée, sous contraintes de cardinalités et contraintes logiques. Les contraintes de cardinalités sont celles qui agissent sur les cardinalités des problèmes d'appariement *((1 : 1), (1 : m), (n : 1) et (n : m))*. Les contraintes logiques sont les contraintes qui permettent aux variables de décision de prendre des valeurs entières ou réelles, autrement dit le domaine des variables. L'objectif étant de découvrir des relations entre les hypergraphes, et pour ce faire, on crée ce que nous appelons un ensemble d'*arêtes virtuelles* qui représentent l'ensemble des correspondances potentielles et où chaque nœud de l'arête appartient à un hypergraphe différent. Avec les valeurs des variables de décisions, l'ensemble des correspondances correctes est trouvé. Dans le cas d'un problème d'appariement avec une cardinalité *(1 : 1)*, la résolution du problème consiste à choisir pour chaque nœud au maximum un nœud correspondant, autrement dit une seule arête. Pour la cardinalité *(1 : m)* (resp. *(n : 1)*) l'objectif revient à trouver, pour chaque nœud (resp. pour un ensemble de nœuds n), un ensemble de nœuds m (resp. un nœud) qui lui sont similaire(s). Enfin, pour la cardinalité *(n : m)*, la résolution du problème consiste à trouver, pour chaque ensemble de nœuds n, un ensemble de nœuds m. La Figure 3.8 illustre ces cardinalités.

Par la suite, nous allons donc décrire les mesures de similarité utilisées, les stratégies d'agrégation et de composition et enfin le modèle d'optimisation.

3.3.1 Mesures de similarité

Les mesures de similarité sont des techniques utilisées pour déterminer la similarité entre deux ou plusieurs objets. Ces objets peuvent être par exemple des séquences d'ADN, des documents, des textes, des chaines de caractères, des graphes ou des valeurs numériques.

Les méthodes qui mesurent la similarité sont diverses et variées et sont largement décrites dans la littérature (GALI et al., 2019) ; (GOMAA, FAHMY et al., 2013) ; (HAKAK et al., 2019) ; (PRADHAN, GYANCHANDANI et WADHVANI, 2015) ; (PRAKOSO, ABDI et AMRIT, 2021) ; (J. WANG et Y. DONG, 2020). Le choix d'une mesure dépend du problème, du contexte d'application, des informations disponibles et de leurs caractéristiques. La généricité de notre approche nous permet de prendre en compte divers types de mesures et les intégrer grâce aux sommes pondérées.

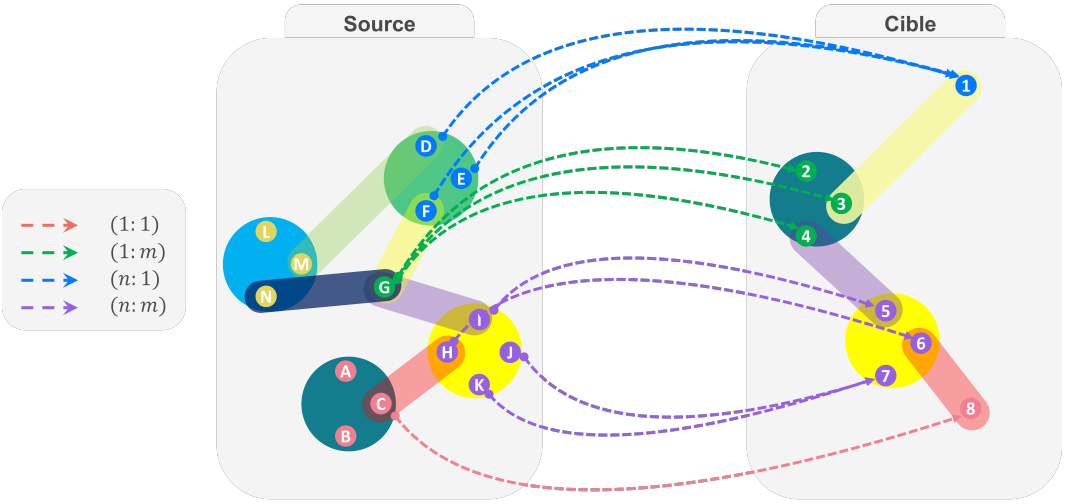


FIGURE 3.8 – Cardinalités du problème d'appariement traduites en appariement d'hypergraphes.

3.3.1.1 Mesure de similarité structurelle

Pour mesurer la similarité (niveau des schémas), nous utilisons les informations structurelles. Par information structurelle, nous entendons les informations qui traduisent la topologie générale des hypergraphes, c'est-à-dire les relations entre les nœuds construites par la notion de clé étrangère. L'Algorithme 1 que nous proposons est une procédure qui peut être appliquée pour différents types de graphes, à savoir les hypergraphes ou les graphes simples, les arbres ou les graphes acycliques avec ou sans racine. L'algorithme prend en compte la notion de degré et de voisinage.

Définition 3.3.1. Dans un hypergraphe $HG = (V, HE)$, le *degré* d'un nœud $v \in V$ noté $deg(v)$ est le nombre d'(hyper)arêtes qui lui sont incidentes, autrement dit, auxquelles il appartient. Soit $HE(v) = \{he_j \mid v \in he_j\}$ avec $he_j \neq \emptyset$ pour $j \in \{1, \cdots, m\}$ l'ensemble des (hyper)arêtes où v apparait. Le degré de v est alors $deg(v) =\mid HE(v) \mid$.

Par exemple, pour l'hypergraphe HG_s, $deg(invoicing\_table) = 2$, $deg(customer\_table) = 1$, $deg(production\_table) = 1$ et $deg(invoice\_details\_table) = 2$. Tandis que pour l'hypergraphe HG_c, $deg(invoice\_base) = 2$, $deg(customer\_base) = 1$, $deg(transportation\_base) = 2$, $deg(production\_base) = 1$ et $deg(invoice\_details\_base) = 2$.

Définition 3.3.2. Soit l'hypergraphe $HG = (V, HE)$ avec $HE = \{he_1, he_2, \cdots, he_m\}$ l'ensemble des (hyper)arêtes et V l'ensemble des nœuds. $N(v)$ le *voisinage* du nœud v est l'ensemble des nœuds w tel que $w \in \cap_{j \in \{1,\cdots,m\}} he_j$.

Pour le nœud, *invoice_details_base* de l'hypergraphe HG_c est $N(v) = \{transportation\_base, invoice\_base, production\_base\}$.

L'utilisation des principes du voisinage et de degré permettent d'évaluer localement la similarité de deux nœuds dans deux hypergraphes hétérogènes dont la structure globale peut être très différente. D'autres mesures existent, à l'image de la densité et le diamètre.

Définition 3.3.3. La *densité* d'un hypergraphe est une mesure du nombre d'(hyper)arêtes de l'hypergraphe par rapport au nombre de nœuds dans l'hypergraphe. Une densité de 0 correspond à un hypergraphe où tous les nœuds sont isolés.

Définition 3.3.4. Un *chemin* ou *(hyper)chemin* HP dans un hypergraphe $HG = (V, HE)$ entre deux nœuds v_1 et v_{k+1} est une séquence $v_1, he_1, v_2, he_2, \cdots, he_k, v_{k+1}$ tel que $\{v_i, v_{i+1}\} \subseteq he_i$ pour $1 \leq i \leq k$ et $v_i \neq v_j$, $he_i \neq he_j$ et k la longueur de l'(hyper)chemin HP. La distance notée $dist(v_i, v_j)$ entre les nœuds v_i et v_j est la longueur minimale d'un chemin qui les relie (c.-à-d. la longueur du plus court chemin entre v_i et v_j). Le diamètre noté $diam(HG)$ est défini par $diam(HG) = \max\{dist(v_i, v_j) \mid v_i, v_j \in V, v_i \neq v_j\}$.

Ces deux concepts (ainsi que d'autres) sont des propriétés structurelles spécifiques qui prennent en compte les caractéristiques globales des hypergraphes (nombre de nœuds, nombre d'(hyper)arêtes, longueur des chemins, etc.) (BANERJEE et PARUI, 2022). Donc, en présence de différences structurelles, ces mesures peuvent donner des évaluations de précisions différentes.

Maintenant, nous donnons la définition de la similarité structurelle basée sur la notion de voisin :

Définition 3.3.5. Deux nœuds u, v dans un hypergraphe $HG = (V, HE)$ sont structurellement équivalents s'ils partagent le mème voisinage et on écrit $N(u) = N(v)$.

Cependant, cette définition se base sur l'équivalence. Dans notre cas où nous étudions le problème d'appariement inexact, cette définition est insuffisante pour mesurer la similarité entre deux nœuds. Elle peut être améliorée pour chercher l'approximation en prenant en compte les degrés des nœuds et en définissant la similarité structurelle locale (CASTRILLO, LEÓN et GÓMEZ, 2018) :

Définition 3.3.6. La similarité structurelle locale des nœuds u et v, désignée par $\sigma(u, v)$, est définie comme le cardinal de l'ensemble des voisins communs aux deux nœuds u, v, elle se note $\mid N(u) \cap N(v) \mid$ et est normalisée par la moyenne de leurs degrés, soit :

$$\sigma(u, v) = \frac{\mid N(u) \cap N(v) \mid}{\sqrt{deg(u) \times deg(v)}}$$

Nous adaptons cette définition et suggérons deux stratégies pour construire ce que nous appelons les *portes d'entrée*, (1) choisir une paire (ou plusieurs paires) de deux nœuds, chacun

dans un hypergraphe, dont nous savons qu'ils correspondent, ou (2) prendre les deux nœuds ayant la plus grande similarité. Sur cette base, nous proposons donc l'Algorithme 1.

Algorithm 1 Similarité structurelle

Entrée: Hypergraphes $HG_s = (V_s, HE_s, \lambda_s)$ et $HG_c = (V_c, HE_c, \lambda_c)$, *portes_d'entrée* g, deux nœuds $u \in HG_s$ et $v \in HG_c$
Sortie: $\sigma(u, v)$

1: $N(u) = voisins(HG_s, u)$
2: $N(v) = voisins(HG_c, v)$
3: $HG'_s = sous\_hypergraphe(N(u) \cup u)$ ▷ $HG'_s = (V'_s, HE'_s, \lambda'_s)$
4: $HG'_c = sous\_hypergraphe(N(v) \cup v)$ ▷ $HG'_c = (V'_c, HE'_c, \lambda'_c)$
5: $E = arêtes\_virtuelles(V'_s, V'_c)$
6: $voisins\_communs = E \cap g$
7: $\sigma(u, v) = \frac{|voisins\_communs|}{\sqrt{deg(u)_{HG'_s} \times deg(v)_{HG'_c}}}$

L'Algorithme 1 calcule la similarité structurelle entre deux nœuds sur la base des *portes d'entrée* g avec $g = \{(p^1_{HG_s}, p^1_{HG_c}), (p^2_{HG_s}, p^2_{HG_c}), \cdots, (p^r_{HG_s}, p^r_{HG_c})\}$, $p^i_{HG_s} \in V_s$ et $p^j_{HG_c} \in V_c$ pour $i, j = 1, 2, \cdots, r$. Ces portes d'entrée nous assurent un début certain et de comparer des voisinages qui ont un taux de ressemblance plutôt élevé avant ceux qui sont moins certains. Nous construisons l'ensemble des voisins pour chaque nœud u et v (lignes (1)-(2)) et les sous-hypergraphes induits par ces voisins incluant les nœuds u et v (lignes (3)-(4)) qui vont servir à calculer le degré des nœuds u et v. Puis, on construit les arêtes virtuelles qui correspondent à l'ensemble des correspondances possibles entre les deux sous-hypergraphes (ligne (5)). Ces arêtes vont être comparées à l'ensemble des *portes d'entrée* g. L'ensemble des *voisins_communs* est constitué des paires de nœuds communes entre ces deux ensembles (ligne (6)) et sa taille traduit le nombre de voisins communs des nœuds u et v. Au final, cette taille est normalisée en utilisant la moyenne des degrés des nœuds u et v dans les sous-hypergraphes respectifs (ligne (7)).

Ainsi, pour construire la matrice de similarité structurelle, cet algorithme est exécuté pour chaque paire de nœuds entre les deux hypergraphes. Cet Algorithme converge en au plus $\mathcal{O}(n^2)$ ($\simeq n \times m$ exécutions). Avec $| V_s | = n_s$, $| V_c | = n_c$ et $| g | \leq \min\{n_s, n_c\}$.

À titre d'exemple, nous reprenons les hypergraphes de la Figure 3.2. Soit :
- $g = \{invoicing\_table; invoice\_base\}$,
- $u = customer\_table$,
- $v = customer\_base$.

En appliquant l'Algorithme 1 nous obtenons les voisinages suivants :
- $N(u) = \{invoicing\_table\}$,
- $N(v) = \{transportation\_base; invoice\_base\}$.

Les sous-hypergraphes induits par les voisinages et les deux nœuds sont :
- $HG'_s = (V'_s, HE'_s, \lambda'_s)$:
  - $V'_s = (N(u) \cup u) = \{customer\_table; invoicing\_table\}$,
  - $HE'_s = \{customer\_num; invoice\_num\} = \{[invoicing\_table, customer\_table]; [invoicing\_table]\}$.
- $HG'_c = (V'_c, HE'_c, \lambda'_c)$:
  - $V'_c = (N(v) \cup v)\{transportation\_base; invoice\_base; customer\_base\}$,
  - $HE'_c = \{id\_customer; id\_invoice\}$
    $= \{[invoice\_base, customer\_base, transportation\_base]; [invoice\_base, transportation\_base]\}$.

En général, pour la construction des sous-hypergraphes, deux définitions sont proposées :

La définition dite *faible* (DEWAR et al., 2017) :

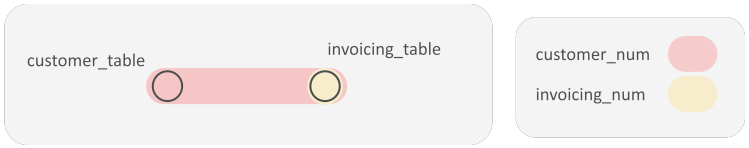
Définition 3.3.7. Soit $HG = (V, HE)$ un hypergraphe, $HG' = (V', HE')$ est un sous-hypergraphe de HG induit par V' avec $V' \subseteq V$ et $HE' \subseteq \{he' \cap V' \mid he' \in HE\}$.

Et la définition dite *forte* (BAHMANIAN et SAJNA, 2015) :

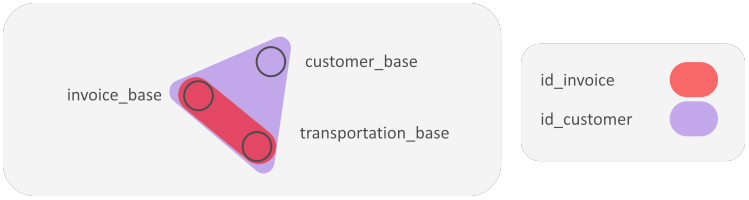
Définition 3.3.8. Soit $HG = (V, HE)$ un hypergraphe, $HG' = (V', HE')$ est un sous-hypergraphe de HG induit par V' avec $V' \subseteq V$ et $HE' = \{he' \in HE \mid he' \subseteq V'\}$.

La Définition 3.3.8 est une généralisation de la définition des sous-graphes. Elle permet de construire des sous-hypergraphes où les (hyper)arêtes HE' sont les (hyper)arêtes de HE qui contiennent uniquement des nœuds de V'. Tandis que la Définition 3.3.7 est une relaxation et est plus adaptée aux problèmes d'appariement inexact puisqu'elle permet d'inclure dans le sous-hypergraphe plus d'information sur les nœuds, autrement dit, si un nœud appartient à plusieurs (hyper)arêtes. Cette information n'est donc pas prise en compte par la Définition 3.3.8. Ainsi, la notion de voisinage est très restreinte et la description du degré d'un nœud est moins précise. La relaxation est alors plus adaptée aux problèmes d'appariement inexact où les structures d'hypergraphes sont très hétérogènes.

Alors, les sous-hypergraphes sont illustrés dans la Figure 3.9 :



(a) Sous-hypergraphe source HG'_s associé à V'_s



(b) Sous-hypergraphe cible HG'_c associé à V'_c

FIGURE 3.9 – Sous-hypergraphes HG'_s et HG'_c.

Ensuite, l'ensemble des arêtes virtuelles est donné par :

$E = \{(invoicing\_table, transportation\_base); (invoicing\_table, invoice\_base); (invoicing\_table, customer\_base); (customer\_table, transportation\_base); (customer\_table, invoice\_base); (customer\_table, customer\_base)\}$.

Et l'ensemble des voisins communs est : $E \cap g = \{(invoicing\_table, invoice\_base)\}$.

Au final, $\sigma(u,v) = \frac{|N(u) \cap N(v)|}{\sqrt{deg(u) \times deg(v)}} = \frac{1}{\sqrt{1 \times 1}} = 1$.

La matrice de similarité structurelle finale notée MF_{str} est de taille $n_s \times n_c = 4 \times 5$ et est présentée dans le Tableau 3.2 :

	invoice_base	customer_base	transportation_base	invoice_details_base	production_base
invoicing_table	0.500	0.707	0.500	0.500	0.000
customer_table	0.707	1.000	0.707	0.707	0.000
production_table	0.000	0.000	0.000	0.000	0.000
invoice_details_table	0.500	0.707	0.500	0.500	1.000

TABLEAU 3.2 – Matrice de similarité structurelle MF_{str} avec $\{invoicing\_table; invoice\_base\}$

3.3.1.2 Mesure de similarité structurelle pour les hypergraphes à large échelle

Pour les hypergraphes à large échelle [1], nous proposons l'Algorithme 2, où nous propageons la similarité par construction itérative des sous-ensembles. Nous commençons par prendre les sous-hypergraphes générés par les *portes d'entrée* en fixant une profondeur k pour le voisinage (*k-order neighbor* inspirée et réadaptée de (J. CHEN, R. ZHAO et Z. LI, 2004)) :

Définition 3.3.9. Deux nœuds (u, v) sont voisins d'ordre k, si la distance entre eux est $k - 1$.

Algorithm 2 Propagation à large échelle

Entrée: Hypergraphes $HG_s = (V_s, HE_s)$ et $HG_c = (V_c, HE_c)$, *portes_d'entrée* g, *profondeur* k, *seuil_min*, *seuil*, *iter_max*, *marge*
Sortie: *portes d'entrée*

1: $it = 0$
2: **Tant Que** *(seuil* $\geq$ *seuil_min) ou (it* $\leq$ *iter_max) ou (tous les nœuds ne sont pas appariés)* **Faire**
3: sHG_s, sHG_c = *sous_hypergraphe*$(HG_s, HG_c, pportes\_d'entrée, k)$
4: *optimisation*(sHG_s, sHG_c, *portes_d'entrée*, *seuil*)
5: **Si** *pas de nouvelles correspondances* **Alors**
6: $seuil = seuil - marge$
7: **Sinon**
8: *mise-à-jour(portes_d'entrée)*
9: **Fin Si**
10: $it = it + 1$
11: **Fin Tant Que**

Ainsi, les sous-hypergraphes générés ne se basent pas sur les voisins incidents, mais sur une profondeur de k voisins (ligne (3)). Ensuite, nous appliquons la procédure d'optimisation qu'on va définir par la suite pour trouver les correspondances (ligne (4)). Une fois les correspondances trouvées, nous mettons à jour l'ensemble des portes d'entrée (ligne (8)), et nous réappliquons la procédure d'optimisation sur les sous-hypergraphes générés par le nouvel ensemble (ligne (10)). Et ainsi de suite, jusqu'à ce que l'ensemble des correspondances soient trouvées (portes d'entrée mis à jour à la fin de l'algorithme).

1. Un hypergraphe est dit à *large échelle* si le nombre des nœuds est ≥ 1.000.

Pour ce processus, les conditions d'arrêt sont : (1) tous les nœuds sont appariés, (2) on fixe un seuil minimal de similarité, on exécute le processus. Si on n'obtient pas de nouvelles correspondances à ajouter et qu'on n'a pas visité tous les nœuds, on réduit le seuil et on répète, jusqu'à atteindre un seuil minimal (lignes (5)-(6)) (inspiré de (X. PENG et al., 2013)), (3) on fixe un nombre maximal d'itérations. Cet Algorithme converge en au plus $\mathcal{O}(n^3)$ mais cette convergence dépend de la complexité du problème d'optimisation à résoudre, en termes de contraintes et types de variables de décision (Section 2.3) et en termes de méthode de résolution utilisée (c-à-d, méthode heuristique [2] ou exacte [3]). Cependant, avec la méthode de propagation des voisins, l'espace des éléments à comparer est réduit et restreint à un sous-ensemble de nœuds et n'est donc pas exhaustif. Ainsi, pour ce type d'hypergraphes, la propagation basée sur de vraies correspondances réduit la marge d'erreur (DURAND et al., 2020).

Ainsi, pour des cas d'applications réelles, le choix des portes d'entrée est très important pour la qualité des résultats et doit être soutenu par une stratégie. La stratégie peut se baser sur la connaissance de la base de données par les experts (métier) et sur la structure des hypergraphes. Les nœuds choisis peuvent donc être des nœuds centraux dans les hypergraphes qui sont généralement des nœuds qui ont de nombreux liens avec d'autres nœuds. Ce choix peut être utilisé pour la propagation de la similarité dans des hypergraphes à large échelle qui ont une structure hiérarchique ou une forte densité de liens. Ainsi, la stratégie de propagation peut être utilisée en repartant des nœuds centraux vers les autres nœuds, ou en prenant les nœuds les plus éloignés du nœud central et de se propager vers ce dernier.

Il est important de noter que dans certains cas, l'équivalence ou l'approximation du voisinage ne reflète pas toujours l'équivalence ou l'approximation des concepts. Autrement dit, deux nœuds équivalents en termes de voisinage ou de similarité structurelle locale peuvent refléter des concepts différents. Une similarité structurelle peut être présente au-delà du voisinage local et au-delà du voisinage à k profondeur. Dans ce cas, la propagation avec un seuil de similarité élevée peut être une piste d'amélioration. Aussi, les mesures de similarité globale peuvent être prises en compte (W. WEN et al., 2020) et dans notre cas, la généricité de notre approche permet en outre l'association de plusieurs mesures de similarité pour la même dimension en calculant par exemple la moyenne entre la similarité structurelle locale et une autre mesure de similarité structurelle.

3.3.1.3 Mesure de similarité syntaxique

La similarité syntaxique fait référence à la similarité de la syntaxe, ou de la structure, de deux mots. Par exemple, les mots *roi* et *rois* peuvent être considérés comme très similaires, tandis que *voiture* et *automobile* peuvent être considérés comme différents.

Il existe plusieurs façons de quantifier la similarité syntaxique. Le choix de la méthode dépendra des caractéristiques structurelles (formes) des mots (J. WANG et Y. DONG, 2020) ; (X. ZHANG, MAO et CAMBRIA, 2022). Ainsi, les mesures se basent sur la longueur des mots ou encore la fréquence et l'ordre des caractères. D'autres mesures utilisent la décomposition des mots ou calculent le nombre minimum d'opérations nécessaires pour transformer un mot en l'autre (distance d'édition). Nous présenterons ci-dessous les mesures retenues dans le cadre du manuscrit :

— *Jaro-Winkler* (WINKLER, 1999) : est une mesure de similarité et une extension de la mesure de base *Jaro* (JARO, 1989). *Jaro* est basée sur le nombre et l'ordre des caractères communs entre deux chaînes de caractères et *Jaro_Winkler* utilise une échelle de préfixe

2. Méthode de calcul qui fournit rapidement une solution réalisable pour un problème d'optimisation dans un temps raisonable mais pas nécessairement optimale.
3. Méthode qui trouve une solution optimale pour une instance d'un problème d'optimisation.

qui donne plus de précision pour les chaînes de caractères qui ont un préfixe en commun dès le début. Cette mesure est donnée par l'expression suivante :

$$Jaro\_Winkler = s_j + (l \times p(1 - s_j))$$

où :

- s_j est la similarité de *Jaro* de deux chaînes de caractères s_1 et s_2, avec :
  - $s_j = \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m})$,
  - $\mid s_i \mid$ la longueur de la chaîne s_i avec $i = 1, 2$,
  - m le nombre de caractères qui correspondent entre les deux chaînes,
  - t le nombre de transpositions : deux caractères adjacents qui sont interchangés dans les deux chaînes.
- l est la longueur du préfixe commun à retenir au début de la chaîne avec $l \leq 4$ caractères,
- p est une constante suggérée comme étant fixée dans les travaux de *Winkler* à $p = 0.1$.

— *Levenshtein* (LEVENSHTEIN et al., 1966) : est une mesure de distance (*Levenshtein_Distance*) et est basée sur le nombre d'opérations d'édition nécessaires pour transformer une chaîne en l'autre, à savoir la suppression, l'insertion et le remplacement de caractères. Une forme normalisée de cette mesure de distance est proposée, garantissant que la valeur soit $\in [0, 1]$ et consiste à diviser le résultat sur la taille de la chaîne la plus longue. La transformation linéaire de la mesure de distance donne la mesure de similarité. Ainsi la similarité de *Levenshtein* entre deux chaînes de caractères s_1 et s_2 est :

$$Levenshtein\_Similarité = 1 - \frac{Levenshtein\_Distance}{\max_{i \in \{1,2\}}\{\mid s_i \mid\}}$$

— *Jaccard* (JACCARD, 1912) : est une mesure de similarité qui utilise la notion d'ensemble. On construit deux ensembles : le premier ensemble X composé de caractères communs et le second Y composé de caractères uniques entre les deux chaînes. La similarité est alors le rapport du nombre d'éléments contenus dans les deux ensembles :

$$Jaccard = \frac{\mid X \cap Y \mid}{\mid X \cup Y \mid}$$

— *Sorensen_Dice* (DICE, 1945) ; (SORENSEN, 1948) : qui est une autre variante de *Jaccard* utilisant des ensembles et peut être considéré comme le pourcentage de chevauchement entre deux ensembles :

$$Sorensen\_Dice = 2 \times \frac{\mid X \cap Y \mid}{\mid X \mid + \mid Y \mid}$$

— *Tversky* (TVERSKY, 1977) : est une généralisation des mesures de similarité de *Jaccard* et de *Sorensen-Dice* qui donne plus de flexibilité pour choisir d'accorder plus d'attention à un ensemble qu'à un autre grâce aux paramètres α et β :

$$Tversky = \frac{\mid X \cap Y \mid}{\mid X \cap Y \mid + \alpha \mid X \setminus Y \mid + \beta \mid Y \setminus X \mid}$$

Avec $X \setminus Y$ désigne le complément relatif de Y dans X, c-à-d., l'ensemble des éléments dans X mais pas dans Y.

— *Coefficient de chevauchement (Overlap coefficient)* (VIJAYMEENA et KAVITHA, 2016) : est une mesure de similarité pour le chevauchement entre deux ensembles, et est défini

comme le nombre de caractères de l'intersection des deux ensembles divisé par la taille du plus petit ensemble. Cette mesure considère deux chaînes de caractères comme une correspondance complète si l'une est un sous-ensemble de l'autre :

$$Overlap = \frac{| X \cap Y |}{min(| X |, | Y |)}$$

— *Ratcliff Obershelp* (RATCLIFF et METZENER, 1988) : est une mesure de similarité qui est basée sur l'utilisation de la notion de sous-chaîne commune :

$$Ratcliff\_Obershelp = 2 \times \frac{SetOfLCS}{| X | + | Y |}$$

Où : *SetOfLCS* est la somme des tailles de toutes les sous-chaines communes.

La mesure de similarité syntaxique est appliquée à deux niveaux, au niveau des attributs et au niveau des tables. En reprenant les hypergraphes de la Figure 3.2, nous appliquons la similarité syntaxique sur les noms d'attributs, puis, en utilisant un algorithme d'agrégation qu'on détaillera par la suite, nous obtenons une valeur de similarité basée uniquement sur les noms d'attributs. Cette valeur est ensuite combinée avec la valeur de la similarité syntaxique calculée en se basant sur les noms des tables grâce à une somme pondérée. Prenons par exemple deux nœuds de HG_s et de HG_c correspondant aux deux tables $u = customer\_table$ et $v = customer\_base$ avec les attributs associés. La similarité syntaxique est calculée en prenant, par exemple, pour le niveau des attributs, la mesure de similarité de *Levenshtein*.

Dans un premier temps, la mesure de similarité normalisée de *Levenshtein* sur les noms d'attributs des deux tables est donnée dans le Tableau 3.3.

	id_customer	customer	address
customer_num	0.416	0.666	0.083
postal_code	0.090	0.272	0.090
customer_name	0.384	0.615	0.076
customer_surname	0.312	0.500	0.125
address	0.181	0.000	1.000

TABLEAU 3.3 – Matrice de similarité syntaxique MA_{syn} entre les noms d'attributs des nœuds (tables) *customer_table* et *customer_base* avec la mesure de *Levenshtein*

Une fois un algorithme d'agrégation appliqué, la valeur finale de la similarité syntaxique basée sur les noms d'attributs est $VA_{syn}(customer\_table, customer\_base) = 0.410$. En appliquant ce procédé pour l'ensemble des couples de nœuds des deux hypergraphes HG_s et HG_c, nous obtenons la matrice de similarité syntaxique finale basée sur les noms d'attributs, notée MAF_{syn} et qui est de taille $n_s \times n_c$ présentée dans le Tableau 3.4.

	invoice_base	customer_base	transportation_base	invoice_details_base	production_base
invoicing_table	0.371	0.347	0.266	0.212	0.172
customer_table	0.153	0.410	0.343	0.131	0.112
production_table	0.221	0.226	0.206	0.215	0.472
invoice_details_table	0.195	0.138	0.230	0.608	0.240

TABLEAU 3.4 – Matrice de similarité syntaxique MAF_{syn} basée sur les noms d'attributs en utilisant la mesure de similarité normalisée de *Levenshtein*

Dans un second temps, la mesure de similarité de *Jaccard* appliquée sur les noms des tables donne la matrice notée MTF_{syn} et qui est de taille $n \times m$ est présentée dans le Tableau 3.5.

	invoice_base	customer_base	transportation_base	invoice_details_base	production_base
invoicing_table	0.588	0.333	0.360	0.590	0.428
customer_table	0.444	0.800	0.375	0.416	0.526
production_table	0.400	0.450	0.521	0.440	0.823
invoice_details_table	0.571	0.360	0.379	0.863	0.440

TABLEAU 3.5 – Matrice de similarité syntaxique MTF_{syn} basée sur les noms des tables en utilisant la mesure de similarité de *Jaccard*

Ainsi, on obtient au final deux matrices de similarité syntaxique MAF_{syn} et MTF_{syn} qui seront combinées en utilisant une somme pondérée.

3.3.1.4 Mesure de similarité sémantique

La similarité sémantique est un concept basé sur la similitude de sens ou de contenu sémantique. En effet, deux mots peuvent avoir une syntaxe différente, mais ont le même sens, comme les synonymes. Par exemple, les mots *voiture* et *automobile* peuvent être considérés sémantiquement similaires. Les mesures proposées se réfèrent à des sources externes à l'image des dictionnaires, des thésaurus, des ontologies ou des modèles entrainés. Ces mesures sont classées en deux grandes catégories :

— Les mesures basées sur la connaissance qui utilisent des structures sémantiques en forme de graphe à l'image de WordNet (G. A. MILLER, 1995) qui est une base de données lexicale de mots anglais (noms, verbes, adjectifs et adverbes) regroupés en ensembles de synsets (synonym set) représentant chaque mot avec des significations différentes sous forme d'arbre. ConceptNet (SPEER, CHIN et HAVASI, 2017) est une autre base de données multilingue ouverte qui couvre un éventail de concepts et de relations représentées sous forme d'arcs typés. Les données proviennent de diverses sources disponibles dans de nombreuses langues. Les mesures sont ensuite basées sur l'utilisation de la notion de chemin et de profondeur (LEACOCK et CHODOROW, 1998) ; (RADA et al., 1989) ; (Z. WU et PALMER, 1994) ou de contenu informationnel (LIN et al., 1998) ; (RESNIK, 1995).

— Les mesures basées sur un corpus utilisent des informations extraites de grands corpus qui représentent une grande collection de textes pour construire des relations entre les mots. On peut trouver, des méthodes basées sur des combinaisons des séquences de mots (DEERWESTER et al., 1990) ; (ROBERTSON et WALKER, 1994) ; (SALTON et BUCKLEY, 1988) ou des méthodes de transformations des mots en vecteurs en utilisant le plongement lexical (incorporation ou plongement de mots (Word Embeddings)) (BENGIO, DUCHARME et VINCENT, 2000).

Plongement lexical : le plongement lexical est un type de représentation des données textuelles sous la forme d'un vecteur numérique dans un espace vectoriel continu. Les vecteurs générés sont proches si les mots correspondants sont sémantiquement similaires dans des contextes similaires. On retrouve par exemple *Word2vec* (MIKOLOV et al., 2013), *GloVe (Global Vectors for Word Representation)* (PENNINGTON, SOCHER et MANNING, 2014), *BERT (Bidirectional Encoder Representations from Transformers)* (DEVLIN et al., 2018) ou encore *USE (Universal Sentence Encoder)* (Y. YANG et al., 2019). Ces modèles capturent le sens et le contexte des mots d'une manière plus robuste, mais se basent sur un entraînement sur une grande quantité de données pour être efficaces et peuvent ne pas comprendre des mots rares qui n'apparaissent pas dans les données entrainées.

Généralement, pour déterminer la similarité sémantique des deux mots s_1 et s_2, on utilise les vecteurs générés X et Y pour calculer le cosinus de l'angle θ entre les deux vecteurs non nuls X et Y :

$$Similarité\_Cosinus = cos(\theta) = \frac{X \times Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$$

Où : X_i et Y_i sont les composantes du vecteur X et Y respectivement.

La *Similarité_Cosinus* n'est pas toujours dans $[0,1]$ puisque la valeur $cos(\theta)$ est dans l'intervalle $[-1,1]$. Pour avoir une similarité dans $[0,1]$, les vecteurs sont normalisés en divisant chaque élément du vecteur par la norme euclidienne (également appelée norme L2) $\|X\|_2$ du vecteur. Cette norme représente la racine carrée de la somme des carrés des éléments du vecteur :

$$\|X\|_2 = (\sum_{i=1}^{n} X_i^2)^{1/2}$$

Diviser chaque élément du vecteur par cette norme permet de mettre à l'échelle le vecteur, tout en préservant la direction du vecteur. Ce processus est important pour le calcul de la *Similarité_Cosinus* car il garantit que les vecteurs sont sur la même échelle.

Comme pour la similarité syntaxique, la mesure de similarité sémantique est appliquée sur les deux niveaux des attributs et des tables. Pour les deux nœuds $u = customer\_table$ et $v = customer\_base$ de HG_s et de HG_c la similarité sémantique sur les noms d'attributs est donnée dans le Tableau 3.6.

	id_customer	customer	address
customer_num	0.751	0.852	0.333
postal_code	0.251	0.198	0.487
customer_name	0.680	0.736	0.406
customer_surname	0.538	0.589	0.404
address	0.337	0.401	1.000

TABLEAU 3.6 – Matrice de similarité sémantique MA_{sem} entre les noms d'attributs des nœuds (tables) *customer_table* et *customer_base* avec la mesure de similarité de *Similarité_Cosinus*

La valeur finale de la similarité sémantique basée sur les noms d'attributs est $VA_{sem}(customer\_table, customer\_base) = 0.506$. Pour les couples de nœuds des deux hypergraphes HG_s et HG_c, nous obtenons la matrice de similarité sémantique finale basée sur les noms d'attributs, notée MAF_{sem} et qui est de taille $n_s \times n_c$ Tableau 3.7.

	invoice_base	customer_base	transportation_base	invoice_details_base	production_base
invoicing_table	0.653	0.533	0.494	0.392	0.346
customer_table	0.286	0.506	0.477	0.254	0.201
production_table	0.333	0.347	0.263	0.392	0.809
invoice_details_table	0.468	0.239	0.398	0.846	0.393

TABLEAU 3.7 – Matrice de similarité sémantique MAF_{sem} basée sur les noms d'attributs en utilisant la mesure de similarité *Similarité_Cosinus*

Ensuite, la matrice de la similarité sémantique basée sur les noms des tables notée MTF_{sem} de taille $n_s \times n_c$ est comme suit dans Tableau 3.8.

	invoice_base	customer_base	transportation_base	invoice_details_base	production_base
invoicing_table	0.621	0.315	0.246	0.511	0.275
customer_table	0.437	0.724	0.169	0.303	0.267
production_table	0.461	0.331	0.222	0.353	0.675
invoice_details_table	0.680	0.240	0.119	0.819	0.237

TABLEAU 3.8 – Matrice de similarité sémantique MTF_{sem} basée sur les noms des tables en utilisant la mesure de similarité de *Similarité_Cosinus*

Ainsi, on obtient au final deux matrices de similarité sémantique MAF_{sem} et MTF_{sem} qui seront combinées en utilisant une somme pondérée.

3.3.1.5 Mesure de similarité des types de données

La similarité des types de données fait référence à la similarité entre des types de données sur la base de leurs caractéristiques et de leurs domaines de définition. Il existe plusieurs façons de mesurer la similarité entre les types de données selon les besoins et les exigences spécifiques :

- La similarité exacte où on considère que deux types de données sont similaires s'ils sont exactement les mêmes.
- Les mesures qui se basent sur ce que l'on appelle la table de compatibilité (MADHAVAN, BERNSTEIN et RAHM, 2001) ; (NAYAK et TRAN, 2007) qui est une solution générique. Cette table est construite soit manuellement, soit par des approches qui créent un arbre hiérarchique spécifique aux types de données utilisées (AL-BAKRI et FAIRBAIRN, 2012) ; (DONGO et al., 2017) ; (HONG-MINH et D. SMITH, 2007) ; (THUY, Y.-K. LEE et S. LEE, 2013), ensuite, les mesures de distance classiques sont appliquées (AL-HASSAN, H. LU et J. LU, 2015).
- En utilisant des techniques d'apprentissage, car les types de données sont généralement des informations universelles et interprétables (entier, réel, date, chaîne de caractère, etc.) (AIPE et GADIRAJU, 2018).
- Approche hybride qui combine plusieurs stratégies.

En reprenant l'exemple précédent des deux hypergraphes HG_s et HG_c, nous pouvons remarquer que les deux schémas correspondants utilisent les mêmes types de données. Ainsi, nous utiliserons par exemple une méthode de calcul de la similarité syntaxique à l'image de la mesure de *Jaccard* et qui sera appliquée uniquement au niveau des attributs. Le Tableau 3.9 représente les valeurs de la similarité des types de données pour les deux nœuds $u = customer\_table$ et $v = customer\_base$.

	id_customer	customer	address
customer_num	1.000	0.000	0.000
postal_code	0.444	0.100	0.100
customer_name	0.000	1.000	1.000
customer_surname	0.000	1.000	1.000
address	0.000	1.000	1.000

TABLEAU 3.9 – Matrice de similarité MA_{type} entre les types de données des attributs des nœuds $u = customer\_table$ et $v = customer\_base$ avec la mesure de similarité de *Jaccard*

En appliquant cette mesure de similarité sur les couples de nœuds des deux hypergraphes HG_s et HG_c, et après agrégation, nous obtenons la matrice de similarité des types de données finale basée uniquement sur les types d'attributs, notée MF_{type} et qui est de taille $n \times m$ présentée dans le Tableau 3.10.

	invoice_base	customer_base	transportation_base	invoice_details_base	production_base
invoicing_table	1.000	0.380	0.527	0.593	0.458
customer_table	0.317	0.600	0.288	0.333	0.444
production_table	0.458	0.740	0.270	0.500	1.000
invoice_details_table	0.593	0.305	0.770	1.000	0.500

TABLEAU 3.10 – Matrice de similarité des types de données MF_{type} de taille $n_s \times n_c$ en utilisant la mesure de similarité de *Jaccard*

Ainsi, on obtient au final une matrice de similarité des types de données MF_{type} qui sera combinée avec les autres matrices calculées précédemment.

3.3.2 Stratégies d'agrégation

Une fois que les différentes matrices de similarité sont calculées aux trois niveaux, pour construire la matrice finale notée $\mathcal{M}$, nous aurons besoin des matrices de similarités des quatre dimensions (structurelle, sémantique, syntaxique et types de données) entre les nœuds (tables). Les matrice de similarité des dimensions sémantique, syntaxique et types de données se calculent dans un premier temps au niveau des attributs puis doivent être agrégées pour obtenir une valeur de similarité pour les couples de tables (nœuds) correspondantes. Pour ce faire, nous utilisons une procédure inspirée et réadaptée des techniques de mesure de la similarité des phrases (similarité textuelle ou similarité des documents) (ACHANANUPARP, X. HU et SHEN, 2008) pour agréger les matrices. Le choix de cette orientation est motivé par le fait que dans ce type de techniques, on cherche à mesurer le degré de similarité entre deux phrases composées d'un ensemble de mots. Dans notre cas, les tables sont composées d'un ensemble d'attributs. Notons que pour la dimension structurelle, nous n'aurons pas besoin d'utiliser cette procédure d'agrégation, car le calcul de la similarité structurelle se fait déjà au niveau des tables uniquement et non au niveau des attributs. Une similarité structurelle peut être établie dans le cas où des relations de dépendances existent ou peuvent être trouvées entre les attributs, à l'image des dépendances fonctionnelles.

Pour ce faire, nous présenterons deux algorithmes gloutons [4] réajustés à partir des travaux de (KURTZBERG, 1962) et un modèle d'optimisation. Ces trois stratégies prennent en entrée une matrice et donnent en sortie une valeur agrégée correspondant à la similarité partielle entre deux nœuds u et v avec n le nombre d'attributs de la table source correspondante au nœud u et m nombre d'attributs de la table source correspondante au nœud v. La valeur finale agrégée est obtenue en appliquant une pénalité (R. FERREIRA et al., 2016).

3.3.2.1 Algorithme d'agrégation global

L'Algorithme 3 calcule la similarité partielle. Les lignes (1)-(4) construisent les paramètres de l'algorithme, à savoir les dimensions de la matrice et leur minimum, et l'initiation de la similarité totale. Dans les lignes (5)-(11), nous recherchons le maximum de la matrice de similarité et supprimons la ligne et la colonne associées. Nous incrémentons la similarité totale ainsi que le nombre d'itérations, et nous répétons le processus jusqu'à ce que nous atteignons le minimum. Ensuite, nous calculons la similarité partielle par le rapport entre la similarité totale et le nombre d'itérations effectuées (min). Cet algorithme converge en au plus $\mathcal{O}(n^3)$ et permet d'avoir une valeur globalement maximum.

4. Effectuer un choix à chaque étape de manière à optimiser localement (maximiser ou minimiser) une certaine quantité pour construire un résultat optimum global.

Algorithm 3 Algorithme d'agrégation global

Entrée: Matrice de similarité $\mathcal{A}$
Sortie: *similarité_partielle*
1: $n = size(\mathcal{A}.lignes)$
2: $m = size(\mathcal{A}.colonnes)$
3: $min = minimum(n, m)$
4: $similarité\_totale = 0$
5: **Pour** $it = 1, it \leq min, it++$ **Faire**
6: $max = max(\mathcal{A})$
7: $similarité\_totale = similarité\_totale + max$
8: $supprimer\_ligne(\mathcal{A}, ligne\_du(max))$
9: $supprimer\_colonne(\mathcal{A}, colonne\_du(max))$
10: **Fin Pour**
11: $similarité\_partielle = \frac{similarité\_totale}{min}$

3.3.2.2 Algorithme d'agrégation local

L'Algorithme 4 utilise le concept de balayage des lignes et des colonnes, qui se déroule en deux étapes. Après l'initialisation des paramètres (lignes (1)-(7)), la première étape consiste en un balayage ligne par ligne (lignes (8)-(13)), où la ligne de l'élément maximal en termes de valeur est sélectionnée. Cette ligne et la colonne correspondantes sont supprimées. En d'autres termes, pour chaque attribut de la table source, nous attribuons l'attribut de la table cible qui lui est le plus similaire. Ainsi, chaque attribut est affecté de manière unique, ce qui constitue une première solution d'affectation. La somme des valeurs maximales constitue donc la valeur de cette solution. La deuxième étape (lignes (14)-(19)) suit la même procédure. Les colonnes sont examinées et sont associées aux lignes qui leur sont similaires. Une deuxième solution est alors construite. La moyenne entre les deux solutions est prise en compte. Cette procédure est présentée dans l'Algorithme 4 et converge en au plus $\mathcal{O}(n^2)$ et permet d'avoir une valeur localement maximum.

3.3.2.3 Modèle d'optimisation pour l'agrégation globale

Le modèle d'optimisation permet d'avoir une valeur globalement optimale. Il prend en compte les contraintes de cardinalité des problèmes d'appariement, par exemple, pour une cardinalité *(1 : 1)* où un attribut de la table source correspond à un attribut de la table cible. Ainsi, le modèle (PLA) est adapté du modèle général d'optimisation qu'on va expliciter dans la prochaine section.

$$\text{maximiser :} \quad similarité\_totale = \sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}x_{ij}$$

$$\text{(PLA)} \qquad \text{sous contraintes} \quad \sum_{j=1}^{m} x_{ij} \leq 1, \qquad \forall i = 1, \cdots, n \tag{3.1}$$

$$\sum_{i=1}^{n} x_{ij} \leq 1, \qquad \forall j = 1, \cdots, m \tag{3.2}$$

$$x_{ij} \in \{0, 1\}, \quad \forall i = 1, \cdots, n; \forall j = 1, \cdots, m \tag{3.3}$$

Après résolution du modèle d'optimisation (ligne (3)), la fonction objectif à maximiser $\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}x_{ij}$ nous donne la similarité totale globalement optimale et les valeurs de la matrice d'affectation $\mathcal{X} = (x_{ij})$ désignent les correspondances entre les attributs de deux tables (ligne (4)). La similarité partielle est ensuite donnée par le rapport entre la valeur de la fonction objectif et le nombre d'affectations (ligne (5)). Cet Algorithme converge en au plus $\mathcal{O}(n^2)$ mais dépend de la complexité du problème d'optimisation à résoudre.

Algorithm 4 Algorithme d'agrégation local

Entrée: Matrice de similarité $\mathcal{A}$
Sortie: *similarité_partielle*

$n = size(\mathcal{A}.lignes)$
$m = size(\mathcal{A}.colonnes)$
$min = minimum(n, m)$
$max_R = 0$
$max_C = 0$
$similarité\_totale_R = 0$
$similarité\_totale_C = 0$
Pour $i = 1, i \leq n, i++$ **Faire**
    $max_R = max(\mathcal{A}(i, :))$
    $similarité\_totale_R = similarité\_totale_R + max_R$
    $supprimer\_ligne(\mathcal{A}, ligne\_du(max_R))$
    $supprimer\_colonne(\mathcal{A}, colonne\_du(max_R))$
Fin Pour
Pour $j = 1, j \leq m, j++$ **Faire**
    $max_C = max(\mathcal{A}(:, j))$
    $similarité\_totale_C = similarité\_totale_C + max_C$
    $supprimer\_ligne(\mathcal{A}, ligne\_du(max_C))$
    $supprimer\_colonne(\mathcal{A}, colonne\_du(max_C))$
Fin Pour
$similarité\_partielle = 1/2 \times (\frac{similarité\_totale_R}{min} + \frac{similarité\_totale_C}{min})$

Algorithm 5 Algorithme d'agrégation optimal

Entrée: Matrice de similarité $\mathcal{A}$
Sortie: *similarité_partielle*

$n = size(\mathcal{A}.lignes)$
$m = size(\mathcal{A}.colonnes)$
$similarité\_totale, \mathcal{X} = résolution\_modèle(PLA)$ ▷ Matrice $\mathcal{X} = (x_{ij})$
$nombre\_affectations = \mid \{x_{ij} = 1, \forall i = 1, \cdots, n \text{ et } j = 1, \cdots, m\} \mid$
$similarité\_partielle = \frac{similarité\_totale}{nombre\_affectations}$

Ensuite, l'étape suivante consiste à calculer la valeur de la pénalité à appliquer pour pallier la différence de taille des deux paramètres n et m. En d'autres termes, les nombres d'attributs entre les deux tables. Le calcul de la pénalité est donné par l'expression (R. FERREIRA et al., 2016) :

$$pénalité = \begin{cases} \frac{|n-m| \times similarité\_partielle}{n} & \text{si } n > m \\ \frac{|n-m| \times similarité\_partielle}{m} & \text{sinon} \end{cases} \tag{3.4}$$

Au final, la valeur agrégée qui reflète la similarité entre deux tables est calculée par l'expression générale (R. FERREIRA et al., 2016) :

$$similarité\_finale = similarité\_partielle - pénalité \tag{3.5}$$

En reprenant l'exemple précédent, les matrices de similarité basées sur les noms d'attributs et types de données sont agrégées par les trois algorithmes et les valeurs de similarité *MA* sont représentées dans le Tableau 3.11.

Matrice de similarité	Algorithme 3	Algorithme 4	Algorithme 5
MA_{syn}	0.410	0.285	0.410
MA_{sem}	0.506	0.450	0.506
MA_{type}	0.600	0.510	0.600

TABLEAU 3.11 – Valeurs de similarité agrégées basées sur les attributs pour les deux nœuds $u = customer\_table$ et $v = customer\_base$

Ainsi, l'Algorithme 3 et l'Algorithme 4 sont des heuristiques et permettent de trouver des solutions dans un temps raisonnable, alors que l'Algorithme 5 est une méthode exacte et permet de trouver la meilleure solution du point de vue de la fonction objectif, dans notre cas, la meilleure solution par rapport à la valeur de la similarité globale. Pour notre exemple, l'Algorithme 3 donne les mêmes résultats que l'Algorithme 5 et peut être utile quand il s'agit d'apparier des hypergraphes à large échelle.

3.3.2.4 Matrice finale $\mathcal{M}$

Le calcul des matrices de similarité s'est fait sur trois niveaux (des attributs, des tables et des schémas) :

1. **Niveau des attributs** : nous obtenons trois matrices de similarité basées sur les noms d'attributs et les type de données : la matrice de similarité syntaxique MAF_{syn}, la matrice de similarité sémantique MAF_{sem} et la similarité des types de données MF_{type}.
2. **Niveau des tables** : nous obtenons deux matrices de similarité basée sur les noms des tables : la matrice de similarité syntaxique MTF_{syn} et la matrice de similarité sémantique MTF_{sem}. Ces deux matrices sont combinées avec les deux matrices équivalentes du précédent niveau.
3. **Niveau des schémas** : nous obtenons une matrice de similarité structurelle MF_{str}.

La composition est basée sur la notion de somme pondérée qui est une expression mathématique qui consiste à multiplier chaque partie de la somme par un poids, puis à additionner les résultats. Les pondérations sont souvent utilisées pour donner à différents éléments différents niveaux d'importance ou d'influence dans la somme finale. Ainsi, les pondérations favorisent le choix d'une ou plusieurs dimensions (MARTINEZ-GIL et ALDANA-MONTES, 2011). D'autres techniques à l'image du maximum, minimum ou encore la moyenne sont utilisées. La technique du maximum prend la valeur de similarité la plus élevée tandis que la technique du minium prend la valeur la plus faible. La moyenne renvoie la moyenne des différentes similarités et qui est considérée comme un cas particulier de la somme pondérée. Ainsi, la somme pondérée est une stratégie linéaire qui cherche un compromis entre les différentes valeurs et est généralement définie comme suit :

Définition 3.3.10. Soit un vecteur de poids tel que $\sum_{i=1}^{n} w_i = 1$, la somme pondérée est alors : $w_1 \times x_1 + w_2 \times x_2 + \cdots + w_n \times x_n = \sum_{i=1}^{n} w_i \times x_i$.

Ainsi, la composition se fait dans un premier temps sur les matrices qui expriment la similarité syntaxique et sémantique basée sur les noms d'attributs et des tables :

$$MF_{sem} = w\_local\_table_{sem} \times MTF_{sem} + w\_local\_attribut_{sem} \times MAF_{sem} \quad (3.6)$$

$$MF_{syn} = w\_local\_table_{syn} \times MTF_{syn} + w\_local\_attribut_{syn} \times MAF_{syn} \quad (3.7)$$

Où *w_local* sont des coefficients de pondération locaux avec :

$$w\_local\_table_{sem} + w\_local\_attribut_{sem} = 1 \tag{3.8}$$
$$w\_local\_table_{syn} + w\_local\_attribut_{syn} = 1 \tag{3.9}$$

Dans un second temps, la composition finale produit la matrice finale $\mathcal{M}$ avec les coefficients de pondération *w_global* en utilisant l'expression suivante :

$$\begin{aligned}\mathcal{M} = w\_global_{sem} \times MF_{sem} + w\_global_{syn} \times MF_{syn} + \\ w\_global_{type} \times MF_{type} + w\_global_{str} \times MF_{str}\end{aligned} \tag{3.10}$$

Avec :

$$w\_global_{sem} + w\_global_{syn} + w\_global_{type} + w\_global_{str} = 1 \tag{3.11}$$

Le vecteur final est :

$$\begin{aligned}[w\_local\_table_{sem}, w\_local\_attribut_{sem}; w\_local\_table_{syn}, w\_local\_attribut_{syn}; \\ w\_global_{sem}, w\_global_{syn}, w\_global_{type}, w\_global_{str}]\end{aligned}$$

Plus une dimension est considérée comme importante, plus la valeur attribuée à son coefficient de pondération est élevée. Par exemple, si la signification des noms de tables est plus importante que celle des noms d'attributs, alors $w\_local\_table_{sem} > w\_local\_attribute_{sem}$. De plus, si une dimension n'est pas prise en compte, son coefficient est nul. Par exemple, si la dimension structurelle est absente ou qu'il est impossible de construire des relations, le poids $w\_global_{str} = 0$. Enfin, le vecteur nous permet d'ajouter d'autres dimensions autres que celles étudiées dans le cadre de nos travaux. Par exemple, si les données sont disponibles dans les deux schémas des bases de données, une mesure de similarité des instances de données de deux attributs peut être utilisée (S. GUPTA et al., 2007). Ainsi, nous pouvons calculer la valeur de la similarité entre deux tables (nœud) par la mesure de la similarité de données de leurs attributs. Un poids w_global_{data} est ajouté en respectant la condition que $w\_global_{sem} + w\_global_{syn} + w\_global_{type} + w\_global_{str} + w\_global_{data} = 1$. Aussi, dans des cas plus complexes, plusieurs mesures de similarité peuvent être utilisées pour la même dimension. Par exemple, pour la dimension syntaxique des noms des tables, on peut associer les mesures de similarité *Jaro-Winkler* et *Jaccard* par l'utilisation des poids comme suit : $MTF_{syn} = w\_local\_table_{syn}(Jaccard) \times MTF_{syn}(Jaccard) + w\_local\_table_{syn}(Jaro\_Winkler) \times MTF_{syn}(Jaro\_Winkler)$.

3.3.3 Formulation du modèle d'optimisation

Le modèle d'optimisation est formulé en se basant sur les deux hypergraphes attribués $HG_s = (V_s, HE_s, \lambda_s)$ et $HG_c = (V_c, HE_c, \lambda_c)$ représentant les deux schémas S_s et S_c où $\mid V_s \mid = n_s$, $\mid V_c \mid = n_c$, $\mid HE_s \mid = m_s$ et $\mid HE_c \mid = m_c$ tels que $n_s \neq n_c$ et $m_s \neq m_c$. Trouver les correspondances entre les deux schémas est équivalent à trouver les correspondances parmi l'ensemble des arêtes virtuelles pondérées entre les deux hypergraphes HG_s et HG_c et où les poids sur ces arêtes représentent les valeurs de similarité calculées dans la matrice $\mathcal{M}$. Chaque arête virtuelle est composée d'un nœud de HG_s et d'un nœud de HG_c. Un ensemble d'arêtes virtuelles qui représentent l'ensemble des correspondances potentielles est créé, et il s'agit de sélectionner les arêtes représentant les vraies correspondances.

Ainsi, la formulation est représentée par un programme linéaire (PLS) en considérant le graphe virtuel $G = (V, E)$ où V est l'ensemble de tous les nœuds des deux hypergraphes en

deux partitions $V_1 \cup V_2$ et E l'ensemble de toutes les arêtes virtuelles. Chaque nœud d'une partition est connecté à tous les nœuds de l'autre partition :

$$\begin{aligned}
& \text{maximiser} && \sum_{i=1}^{n_s}\sum_{j=1}^{n_c} s_{ij}x_{ij} && && \\
\text{(PLS)} \quad & \text{sous contraintes} && \sum_{j=1}^{n_c} x_{ij} \leq n, && \forall i = 1,\cdots,n_s && (3.12)\\
& && \sum_{i=1}^{n_s} x_{ij} \leq m, && \forall j = 1,\cdots,n_c && (3.13)\\
& && x_{ij} \in \{0,1\}, && \forall i = 1,\cdots,n_s; \forall j = 1,\cdots,n_c && (3.14)
\end{aligned}$$

Avec :

Variables de décision : qui sont les inconnues du programme qui décrivent (ou décident) des quantités à déterminer. Dans notre cas, les variables x_{ij} sont binaires, c'est-à-dire qu'elles prennent la valeur 1 si une arête virtuelle est choisie et 0 sinon.

Fonction objectif : qui est la fonction à optimiser composée de variables des décision et des paramètres. Cette fonction est utilisée comme critère pour déterminer la meilleure solution au problème modélisé en lui associant une valeur. Pour notre problème d'appariement, les paramètres s_{ij} sont les valeurs de la matrice de similarité $\mathcal{M}$ de taille $n_s \times n_c$ où $s_{ij} \in [0,1]$ quantifie le degré de similarité globale entre le nœud $i \in HG_s$ et $j \in HG_c$ [5].

Contraintes : il s'agit d'un ensemble d'équations ou d'inéquations composées des variables de décision et des paramètres que la solution doit satisfaire.

— Les contraintes (3.12) (resp. 3.13) concernent les cardinalités d'appariement *(n : m)* et expriment le fait que au plus n nœuds de l'hypergraphe HG_s (resp. m de HG_c) peuvent correspondre à au plus m nœuds dans l'hypergraphe HG_c (resp. n de HG_s).

Dans le cas où l'on cherche une correspondance entre des hypergraphes avec $n_s = n_c$ les inégalités peuvent devenir des égalités (mais pas nécessairement), et on écrit :

$$\sum_{j=1}^{n_c} x_{ij} = n, \quad \forall i = 1,\cdots,n_s \tag{3.12'}$$

$$\sum_{i=1}^{n_s} x_{ij} = m, \quad \forall j = 1,\cdots,n_c \tag{3.13'}$$

— Si nous recherchons les cardinalités *(1 : m)*, *(1 : 1)* ou *(m : 1)*, nous remplaçons le deuxième terme des contraintes. Par exemple, pour la cardinalité *(1 : 1)* à l'image du modèle d'optimisation d'agrégation (PLA), les contraintes deviennent :

$$\sum_{j=1}^{n_c} x_{ij} \leq 1, \quad \forall i = 1,\cdots,n_s \tag{3.12''}$$

$$\sum_{i=1}^{n_s} x_{ij} \leq 1, \quad \forall j = 1,\cdots,n_c \tag{3.13''}$$

5. Il est important de noter que la matrice finale $\mathcal{M}$ est modifiée pour prendre en compte les portes d'entrée g. En effet, les portes d'entrée constituent de vraies correspondances, ainsi, pour $(u,v) \in g$ la valeur de $x_{uv} = 1$, alors, la ligne u et la colonne v de la matrice s_{uv} sont soit supprimées, soit modifiées avec $s_{uv} = 1$.

— La contrainte (3.14) concerne les contraintes d'intégrité qui imposent le choix ou non d'une arête. La relaxation de cette contrainte permet au modèle de choisir plus d'une arête. Par exemple, cette relaxation peut être ajoutée dans le cas des cardinalités d'appariement *(n : m)* (BERRO, MEGDICHE et TESTE, 2015) :

$$x_{ij} \in \{0,1\}, \quad \forall i = 1, \cdots, n_s; \forall j = 1, \cdots, n_c \tag{3.14'}$$

— Si on veut prendre en compte des contraintes de seuil q, la contrainte suivante peut être ajoutée au modèle :

$$x_{ij} \times s_{ij} \geq q, \quad \forall i = 1, \cdots, n_s; \forall j = 1, \cdots, n_c \tag{3.15}$$

En appliquant le modèle d'optimisation (PLS) pour l'exemple Figure 3.4, nous obtenons les correspondances suivantes : *(customer_table, customer_base)* ; *(production_table, production_base)* et *(invoice_details_table, invoice_details_base)* qui sont désignées dans le modèle d'optimisation par $x_{ij} = 1$.

En somme, le modèle d'optimisation prend en compte différentes dimensions de similarité et est extensible pour la prise en compte des seuils de similarité et d'autres contraintes. Ce modèle est donc utilisé pour le processus d'agrégation (PLA) et de sélection (PLS).

3.4 Conclusion

Dans ce chapitre, nous avons présenté l'architecture de l'approche proposée pour l'appariement des schémas moyennant une modélisation des schémas par les hypergraphes et l'utilisation des modèles d'optimisation à différentes étapes du processus d'appariement.

L'approche est flexible, globale et générique, elle offre un environnement qui permet :

1. D'exploiter plusieurs types d'informations, à savoir, la sémantique, la syntaxe, les types de données et la structure ;
2. De prendre en compte différentes variations du problème d'appariement des schémas, telles que le seuil, les cardinalités d'appariement ou pallier l'indisponibilité où la difficulté d'interprétation d'informations ;
3. D'être appliqué aux problèmes d'appariement en réseau ou à large échelle ;
4. D'élever le niveau d'abstraction et de la représentation des informations et des relations complexes ;
5. D'inclure de plusieurs types de mesures de similarité pour les différentes dimensions et pour un même dimension ;
6. D'agréger des matrices et de sélectionner des correspondances guidé par les modèles d'optimisation qui offrent la meilleure solution globale.

Dans le chapitre suivant, nous présenterons l'implémentation de l'architecture de l'approche et les résultats des expérimentations effectuées.

4

Implémentation et expérimentations autour de l'approche

4.1 Introduction

Dans le Chapitre 3, nous avons présenté une approche flexible, globale et générique pour contribuer à la résolution de l'appariement de schémas des bases de données en offrant un environnement capable d'intégrer et de rassembler plusieurs étapes et stratégies à divers niveaux.

Dans ce chapitre, nous présenterons dans un premier temps l'environnement d'implémentation du prototype de l'approche dans la Section 4.2. Puis dans un second temps les illustrations des résultats d'évaluations sur deux cas d'études seront détaillées dans la Section 4.3. Les expérimentations sont donc menées sur deux cas d'études pour des processus de migration de bases de données relationnelles. Ces expérimentations ont été réalisées sur une série de simulations basées sur l'utilisation des vecteurs de poids présentés dans la Section 3.3.2.4. Les résultats obtenus sont ensuite analysés en s'appuyant sur des mesures d'évaluation. Enfin, la Section 4.4 concluera le chapitre.

4.2 Processus implémenté

Dans cette section, nous présentons l'architecture du prototype de l'approche développée. Le processus a cinq principaux composants où plusieurs algorithmes et procédures sont implémentés : (1) processus de prétraitement, (2) processus de calcul de similarité, (3) processus d'agrégation et de composition et enfin (4) processus de sélection et d'évaluation. L'ensemble de ces processus ont été mis en œuvre à l'aide du langage *Python* (VAN ROSSUM et DRAKE, 2009) car il offre un bon support pour l'intégration et le développement. Un ensemble de bibliothèques et une multitude de modules et de fonction utiles permette le développement d'algorithmes, la gestion des données, la connexion des bases de données et l'utilisation de techniques de plongement et la théorie des graphes.

4.2.1 Processus de prétraitement

L'objectif de ce processus est de préparer les données et les hypergraphes pour la suite de l'approche. Ce processus est présenté dans la Figure 4.1 et est décomposé en quatre grandes étapes :

(P1-1) : Connecter les bases de données : des connexions vers les bases de données sont réalisées où les schémas et les données sont importées via des requêtes *SQL* ;

(P1-2) : Parsing des fichiers *SQL* : dans le cas où la connexion est impossible, les fichiers *SQL* des schémas sont explorés pour extraire les informations données nécessaires ;

(P1-3) : Organisation des données : où les informations importées sont organisées en tables, attributs, types de données, clés primaires et étrangères. Dans des cas d'applications peu récentes où les bases de données n'ont pas été structurées rigoureusement, il arrive fréquemment de ne pas connaître les clés primaires des tables a priori. Dans ce cas, la méthode *primary_key_search()* basée sur le calcul des fréquences des instances de données et le principe d'unicité de la clé primaire est utilisée. Cette méthode est conditionnée par la fiabilité et la qualité des instances de données. Elle permet de détecter des clés primaires atomiques ou composites. Enfin, en se basant sur la notion de clé primaire, la méthode *foreign_key_search()* permet de construire les clés étrangères ;

(P1-4) : Modélisation des hypergraphes : dans cette étape, en se basant sur les données importées et les clés construites, les hypergraphes sont construits. Un exemple est proposé dans la Section 4.3.2.

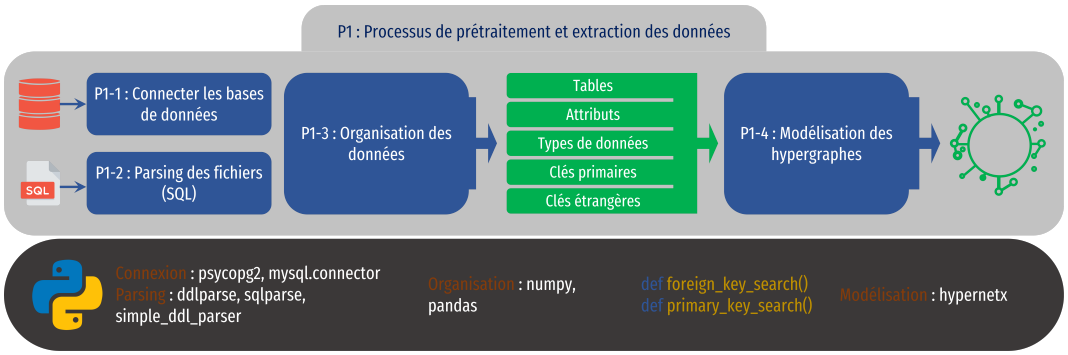


FIGURE 4.1 – Processus de prétraitement et extraction des informations.

4.2.2 Processus de calcul de similarité

Une fois les données importées et les hypergraphes construits, l'objectif de ce processus décrit dans la Figure 4.2 est de calculer les différentes matrices de similarité sur les quatre aspects comme décrit dans la Section 3.3.1 :

(P2-1) : Mesures de similarité structurelle ;
(P2-2) : Mesures de similarité syntaxique ;
(P2-3) : Mesures de similarité sémantique ;
(P2-4) : Mesures de similarité des types de données.

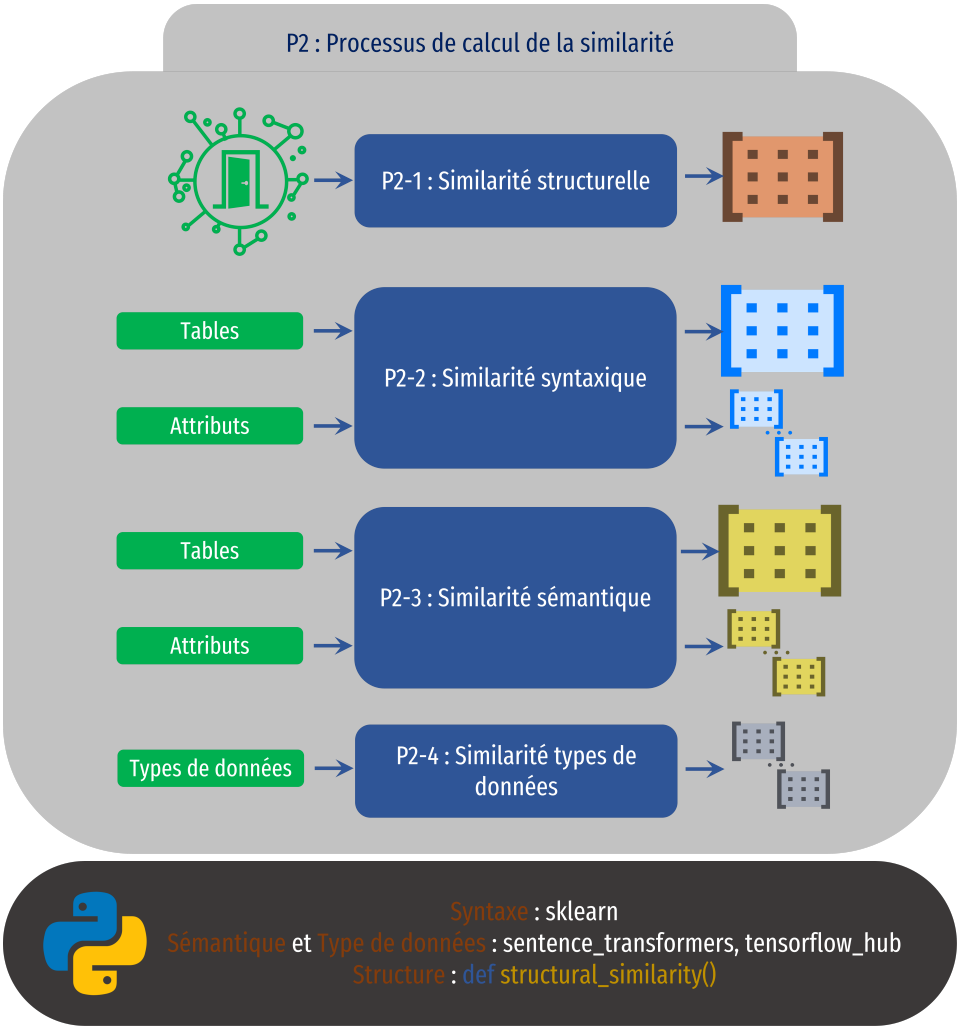


FIGURE 4.2 – Processus de calcul de la similarité.

4.2.3 Processus d'agrégation et de composition

Ce processus est illustré dans Figure 4.3 et prend en entrée les matrices de similarité précédemment calculées pour les agrégées **(P3-1)** puis les composées **(P3-2)** avec les vecteurs de poids. Le processus génère en sortie la matrice de similarité finale $\mathcal{M}$ comme présenté dans la Section 3.3.2.

4.2.4 Processus de sélection et d'évauation

Dans ce processus présenté dans la Figure 4.4, nous définissons les différentes parties du modèle d'optimisation **(P4-1)**, à savoir, les variables de décision, les contraintes et la fonction objectif comme présenté dans la Section 3.3.3. Le modèle est ensuite résolu par le solveur *Gurobi* (GUROBI OPTIMIZATION, LLC, 2022). Ensuite, les résultats sont évalués par l'utilisation de mesures d'évaluation **(P4-2)** présentées dans la prochaine section.

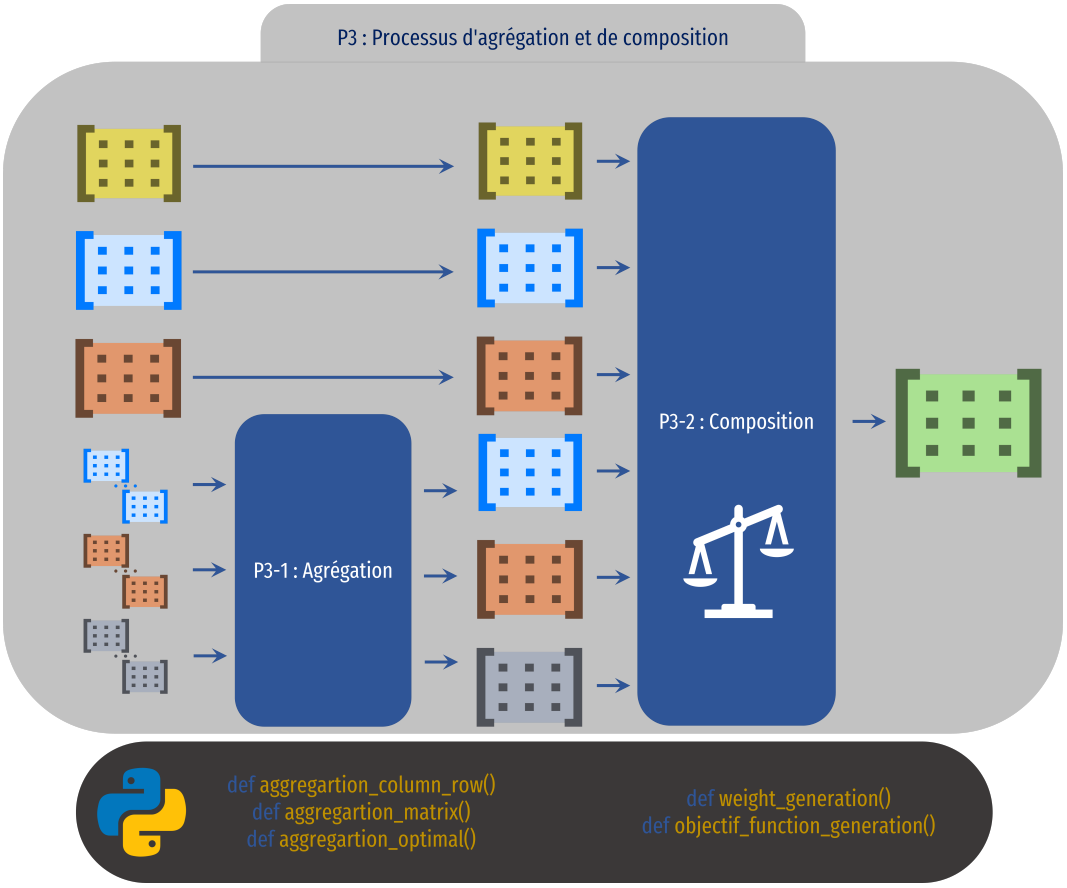


FIGURE 4.3 – Processus d'agrégation et de composition.

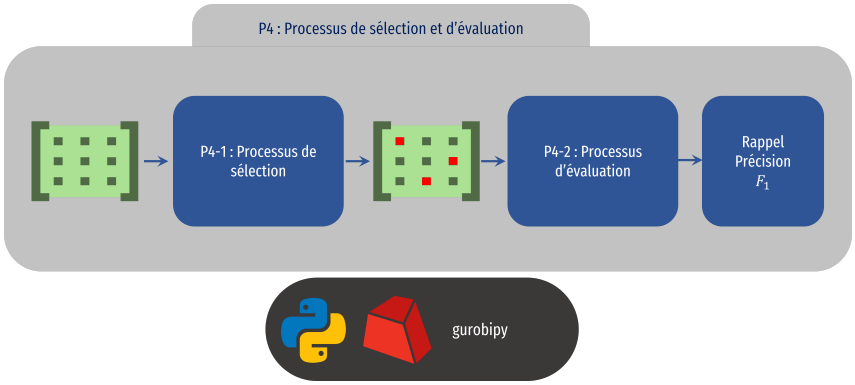


FIGURE 4.4 – Processus de sélection et d'évaluation.

4.3 Expérimentation

4.3.1 Cadre et mesures d'évaluation

Les expériences sont menées sur deux cas d'études qui sont des schémas de bases de données relationnelles présentant des hétérogénéités à plusieurs niveaux. Un cas d'étude où on a deux bases de données relationnelles représentant des informations géographiques et un

second cas d'étude où on a deux sous parties issues des bases de données relationnelles provenant de deux ERP industriels.

Afin d'évaluer la pertinence des résultats, il était important de pouvoir tester l'approche sur ce que l'on appelle des *benchmarks* qui sont un ensemble standardisé de problèmes accompagnés de leurs solutions et des résultats de tests servant de base de comparaison et donnant la solution attendue au problème, c'est-à-dire l'ensemble des correspondances à trouver (BELLAHSENE et al., 2011). Pour quantifier les résultats, des mesures sont utilisées, à savoir : la *Précision (Precision)*, le *Rappel (Recall)* et la *F-Mesure (F-Measure)* (DO, MELNIK et RAHM, 2002). Pour utiliser ces mesures, la base de comparaison *gold standard* est construite en général manuellement en désignant l'ensemble des (*correspondances de référence*) et est comparée à l'ensemble des correspondances trouvées par le biais de l'approche testée (*correspondances automatiques*). La Figure 4.5 montre les différents ensembles résultant de la comparaison entre les correspondances de référence et celles trouvées automatiquement et ces ensembles sont définis comme suit :

— *Vrais Positifs ou True Positives* (TP) (B) : les vraies correspondances correctement identifiées par l'approche ;
— *Faux Positifs ou False Positives* (FP) (C) : les correspondances trouvées par l'approche, mais qui sont incorrectes ;
— *Vrais Négatifs ou True Negatives* (TN) (D) : les fausses correspondances correctement écartées par l'approche ;
— *Faux Négatifs ou False Negatives* (FN) (A) : les correspondances identifiées comme fausses par l'approche, mais qui sont de vraies correspondances.

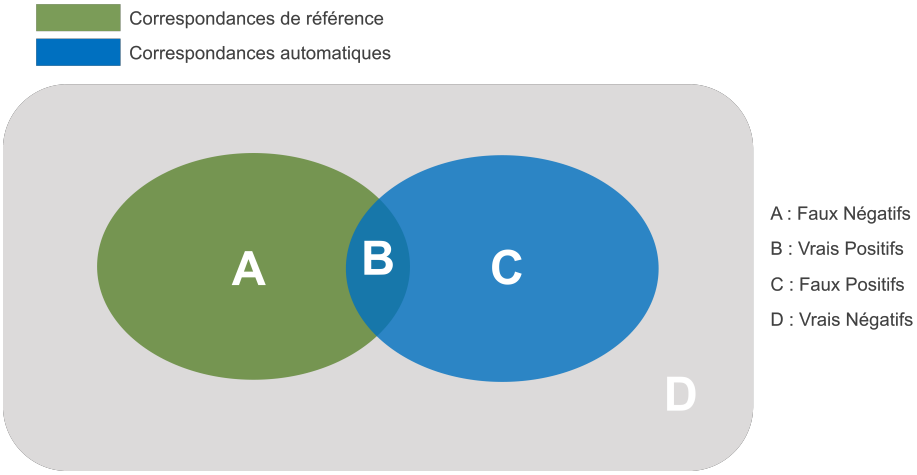


FIGURE 4.5 – Vue générale des ensembles résultant de la comparaison des correspondances de référence et des correspondances trouvées automatiquement.

Sur la base de ces ensembles, les mesures d'évaluation sont exprimées comme suit :

— *Rappel (R)* :

$$R = \frac{TP}{TP + FN}$$

Cette mesure calcule la proportion des correspondances correctement trouvées par rapport à toutes les correspondances. Plus R est élevé, plus l'approche maximise le nombre de correspondances positives, c'est-à-dire qu'avec un $R = 1$, toutes les correspondances ont été trouvées (100%). Néanmoins, cette mesure ne donne pas une indication sur le nombre de correspondances supplémentaires trouvées incorrectement identifiées comme vraies.

— *Précision (P)* :

$$P = \frac{TP}{TP + FP}$$

Mesure la part de correspondances correctement trouvées par rapport à l'ensemble des correspondances de référence. Autrement dit, mesure l'exactitude. Plus P est élevée, l'approche minimise le nombre de faux positifs. Une précision $P = 1$ exprime que toutes les correspondances trouvées sont correctes (100%).

— *F-Mesure* ($F_\alpha - score$) :

$$F_\alpha - score = (1 + \alpha^2) \times \frac{P \times R}{(\alpha^2 \times P) + R}$$

Cette mesure combine les deux précédentes mesures. Un paramètre de pondération α est utilisé pour ajuster l'influence d'une mesure sur une autre en fonction des besoins spécifiques. Si $\alpha = 1$, cela signifie que la précision et le rappel ont la même importance, tandis qu'une valeur de $\alpha < 1$ indique que la précision est plus importante que le rappel, et une valeur de $\alpha > 1$ indique que le rappel est plus important que la précision. Une valeur élevée de $F_\alpha - score$ indique qu'on a réussi à prédire correctement un grand nombre d'éléments pertinents (c'est-à-dire un rappel élevé) tout en ayant un faible taux d'erreur (c'est-à-dire une précision élevée). Dans les expérimentations, $\alpha = 1$, ainsi, $F_1 - score = 2 \times \frac{P \times R}{P + R}$.

Ainsi, ces mesures sont utilisées pour évaluer les résultats des différentes expériences et le fonctionnement des algorithmes et stratégies sur les deux cas d'études. En somme, ces expériences sont à la base d'un certain nombre de simulations où les mesures de similarité, les poids associés et les stratégies d'agrégation sont testés. Elles sont résumées dans le Tableau 4.1 ci-dessous.

Mesures sémantiques	*USE & Similarité_Cosinus*
Mesures syntaxiques	*Jaro_Winkler*, *Levenshtein_Similarité*, *Jaccard*
Mesures types de données	*USE & Similarité_Cosinus*, *BERT & Similarité_Cosinus*
Mesure structurelle	Algorithme 1
Stratégies d'agrégation	Algorithme 3
Stratégie de sélection	Modèle (PLS)
Stratégie de combinaison	Somme pondérée
Vecteurs de pondération	$\approx 18k$ vecteurs

TABLEAU 4.1 – Les mesures et stratégies retenues pour les évaluations

Pour la similarité sémantique, nous avons utilisé le modèle pré-entraîné *USE* pour encoder les noms des tables et d'attributs dans des vecteurs. *USE* est un modèle de traitement du langage développé par *Google* pré-entraîné sur des corpus multilingues, ce qui lui permet d'avoir une compréhension dans plusieurs langues et est adapté pour la similarité entre des phrases dans différentes langues. Ensuite, la mesure de similarité *Similarité_Cosinus* est utilisée pour calculer la similarité entre ces vecteurs. Dans le cas de la similarité syntaxique, les trois mesures retenues sont *Jaro_Winkler*, *Levenshtein_Similarité* et *Jaccard* et leur utilisation dépend des spécifications syntaxiques, des noms des tables et des attributs des cas d'études. Pour la similarité des types de données, nous avons utilisé *BERT & Similarité_Cosinus* pour encoder les types de données dans des vecteurs. *BERT* est un autre modèle dévelopé par *Google* qui a été pré-entraîné sur des corpus de données monolingues pour plusieurs langues, c'est-à-dire que les corpus utilisés pour son entraînement étaient uniquement écrits dans une seule langue. Cela lui permet d'avoir une compréhension précise et est donc efficace pour des tâches de similarité dans une seule langue. Cependant, il est important de noter que le choix entre ces modèles dépendra principalement des données et de l'évaluation de la performance des modèles sur les tâches à mener.

L'Algorithme 1 est utilisé pour mesurer la similarité structurelle étant donné que dans les cas d'études, le nombre de tables dans les bases de données est inférieur à 100. Cet algorithme utilise la notion de *portes d'entrée*. Le choix des portes d'entrée revient à choisir un (des) couple(s) de deux nœuds, chacun appartenant à un hypergraphe. Ce choix est très important pour la construction de la similarité structurelle, car l'Algorithme 1 se base sur ces portes pour construire le voisinage. Ceci nous permet donc de comparer deux voisinages qui partagent les mêmes données. Le choix des portes d'entrée peut donc être déterminant, surtout dans des cas où l'hétérogénéité sémantique, syntaxique ou des types de données est très présente. Ce choix permet donc d'éviter les fausses correspondances. Enfin, le nombre de portes d'entrée n'est pas fixé et comme initié dans la Section 3.3.1.1, le choix peut se faire suivant une stratégie, en faisant un calcul de similarité au préalable, puis choisir des correspondances avec une haute valeur de similarité. Dans le cas des bases de données industrielles, le choix est guidé et contrôlé par la connaissance des bases de données par l'utilisateur métier. Ainsi, on peut donc choisir différentes portes d'entrée traduisant des concepts métiers différents.

Pour la stratégie d'agrégation, nous avons retenu pour les expérimentations l'*Algorithme 3* qui s'exécute en un temps raisonnable et donne les mêmes résultats que le modèle d'optimisation (PLA). Le modèle d'optimisation (PLS) est adapté suivant le cas d'étude pour le processus de sélection et la notion de somme pondérée est utilisée pour le processus de combinaison (composition) des mesures de similarité.

Enfin, pour les simulations, nous avons $\approx 18k$ vecteurs de poids. Les simulations consistent alors à trouver les valeurs des poids qui maximisent les mesures d'évaluation. Pour chaque vecteur, on obtient une solution optimale exprimée par les valeurs des variables de décisions. Cette solution optimale maximise alors la valeur de la fonction objectif présentée dans la Section 3.3.3. Chaque variable traduit le choix ou pas d'une correspondance. L'ensemble des correspondances trouvées est ensuite comparé à l'ensemble des correspondances à trouver construit manuellement. Les mesures d'évaluation (R, P et $F_1 - score$) sont ensuite calculées.

Le modèle d'optimisation tente de trouver la meilleure solution qui maximise la fonction objectif tout en satisfaisant les contraintes. La solution optimale est alors un équilibre ou un compromis entre les poids à trouver.

4.3.2 Cas d'étude 1 : Bases de données géographiques

Pour ce premier cas d'étude, nous utilisons deux bases de données, la première est connue sous le nom de *Mondial* [1] (MAY, 1999) et représente une base de données contenant des données géographiques en anglais provenant de différentes sources web et issue du portail *Universität Freiburg, Institut für Informatik*. La seconde est une autre représentation d'une base de données d'informations géographiques, mais cette fois allemande provenant du portail *Sächsischer Bildungsserver, Serviceportal* et connue sous le nom de *Terra* [2] (DÜRR et RADERMACHER, 2013).

Ces deux bases de données sont utilisées dans divers articles (COFFMAN et WEAVER, 2010) ; (IZQUIERDO et al., 2018) ; (Y. SHI et al., 2021) et sont disponibles dans plusieurs formats. Nous nous intéressons au format relationnel de la base de données en considérant les schémas relationnels correspondant à chaque base de données. Le Tableau 4.2 résume les statistiques relatives aux informations présentes dans les deux schémas.

— **# tables** donne le nombre total des tables ;
— **# attributs** donne le nombre total d'attributs (non unique) ;
— **# clés primaires** donne le nombre de tables où on a des contraintes de clés primaires ;
— **# clés étrangères** fait référence au nombre de contraintes de clés étrangères.

Schéma	# tables	# attributs	# clés primaires	# clés étrangères
Terra	25	104	18	0
Mondial	33	135	31	0

TABLEAU 4.2 – Cas 1 : Statistiques des bases de données *Mondial* et *Terra*

Les deux schémas présentent des sources d'hétérogénéité à plusieurs niveaux en termes de nombre de tables, attributs et clés ainsi qu'en termes de syntaxe. Ces hétérogénéités apparaissent même lorsqu'il s'agit de deux entités traduisant les mêmes données comme le montre la Figure 4.6 et la Figure 4.7 qui représentent des données géographiques sur des montagnes.

```
CREATE TABLE `BERG` (
`B_NAME` varchar(20) DEFAULT NULL,
`GEBIRGE` varchar(25) DEFAULT NULL,
`HOEHE` double(16,4) DEFAULT NULL,
`JAHR` int(11) DEFAULT NULL,
`LAENGE` double(16,4) DEFAULT NULL,
`BREITE` double(16,4) DEFAULT NULL
) ENGINE=MyISAM DEFAULT CHARSET=latin1 PACK_KEYS=1;

ALTER TABLE `BERG`
ADD UNIQUE KEY `B_NAME` (`B_NAME`);
```

FIGURE 4.6 – Cas 1 : Déclaration de la table *'BERG'* au sein du schéma *Terra*.

```
CREATE TABLE mountain
(Name VARCHAR(35),
Mountains VARCHAR(35),
Height FLOAT,
Type VARCHAR(10),
Longitude FLOAT,
Latitude FLOAT,
CONSTRAINT MountainKey PRIMARY KEY(Name),
CONSTRAINT CHECK ((Longitude >= -180) AND (Longitude <= 180)
      AND  (Latitude >= -90) AND (Latitude <= 90)));
```

FIGURE 4.7 – Cas 1 : Déclaration de la table *mountain* au sein du schéma *Mondial*.

Avant de commencer à réaliser les simulations, les correspondances de référence sont construites manuellement pour permettre une base d'évaluation. Ensuite, les schémas sont importés en considérant le schéma *Terra* comme schéma source et le schéma *Mondial* comme schéma cible. L'ensemble des données est extrait et organisé en tables, attributs, types de données et clés. Des opérations de prétraitement sont aussi effectuées afin d'avoir plus de précision dans les calculs, car nous traitons du texte. En effet, certaines mesures de similarité sont sensibles à la présence des majuscules ou des caractères spéciaux. Par conséquent, le prétraitement pour ces bases de données vise à améliorer la précision des calculs en uniformisant la syntaxe par la suppression des caractères spéciaux comme le trait du bas sans altérer le sens des mots et sans pour autant changer la syntaxe générale.

Une fois ces données importées et classées, elles sont utilisées pour construire les hypergraphes. Le manque de référence sur les clés étrangères dans ces bases de données est une contrainte importante dans la construction des hypergraphes, car elles contribuent non

1. https ://www.dbis.informatik.uni-goettingen.de/Mondial/
2. https ://www.sachsen.schule/ terra2014/

seulement à renforcer les relations entre les tables et à garantir l'intégrité des données, mais aussi facilitent l'interrogation et la manipulation des données dans la base de données.

Ainsi, afin de pallier cette absence, nous utilisons la notion de clé primaire pour construire les clés étrangères dans les tables correspondantes. Par exemple, pour la base de données *Mondial*, nous prenons une partie de son schéma illustrée dans la Figure 4.8 et composée de quatre tables.

Les clés primaires et les attributs de chaque table sont résumés dans le Tableau 4.3.

Tables	Clé primaire	Attributs
country	(Code)	[Name, Code, Capital, Province, Area, Population]
city	(Name, Country, Province)	[Name, Country, Province, Population, Longitude, Latitude]
province	(Name, Country)	[Name, Country, Population, Area, Capital, CapProv]
language	(Name, Country)	[Name, Country, Percentage]

TABLEAU 4.3 – Cas 1 : Tables, clés primaires et attributs des tables de la Figure 4.8

L'attribut *Country* dans les clés primaires des trois tables *city*, *province* et *language* fait référence à la clé primaire *(Code)* clé de la table *country*. Cet attribut constitue donc une clé étrangère. Aussi, l'attribut *Province* faisant partie de la clé primaire de la table *city* appartient aux attributs de la table *country* et constitue donc une clé étrangère. Ainsi, les clés étrangères sont présentées dans le Tableau 4.4.

Tables	Clé primaire	Attributs	Clés étrangères
country	(Code)	[Name, Code, Capital, Province, Area, Population]	Province
city	(Name, Country, Province)	[Name, Country, Province, Population, Longitude, Latitude]	Country, Province
province	(Name, Country)	[Name, Country, Population, Area, Capital, CapProv]	Country
language	(Name, Country)	[Name, Country, Percentage]	Country

TABLEAU 4.4 – Cas 1 : Tables, clés primaires, attributs et clés étrangères des tables de la Figure 4.8

Ainsi, chaque clé étrangère sera une (hyper)arête dans l'hypergraphe correspondant, et on aura les (hyper)arêtes dans le Tableau 4.5.

(Hyper)arêtes	[Province]	[Country]
Nœuds	{country, city, province }	{country, city, province, language}

TABLEAU 4.5 – Cas 1 : (Hyper)arêtes correspondantes aux tables de la Figure 4.8

Dans le cas où des tables n'ont pas de clé étrangère, nous rajoutons une (hyper)arête qui regroupe les nœuds isolés correspondants aux tables sans clé étrangère. Au final, nous obtenons les statistiques présentées dans le Tableau 4.6.

Schéma	# tables	# attributs	# clés primaires	# clés étrangères
Terra	25	104	18	8
Mondial	33	135	31	11

TABLEAU 4.6 – Cas 1 : Statistiques des bases de données *Mondial* et *Terra* après prétraitement

```
CREATE TABLE country
(Name VARCHAR(35) NOT NULL UNIQUE,
Code VARCHAR(4),
Capital VARCHAR(35),
Province VARCHAR(35),
Area FLOAT,
Population INT,
CONSTRAINT CountryKey PRIMARY KEY(Code),
CONSTRAINT CountryArea CHECK (Area >= 0),
CONSTRAINT CountryPop CHECK (Population >= 0));

CREATE TABLE city
(Name VARCHAR(35),
Country VARCHAR(4),
Province VARCHAR(35),
Population INT,
Longitude FLOAT,
Latitude FLOAT,
CONSTRAINT CityKey PRIMARY KEY (Name, Country, Province),
CONSTRAINT CityPop CHECK (Population >= 0),
CONSTRAINT CityLon CHECK ((Longitude >= -180)
    AND (Longitude <= 180)),
CONSTRAINT CityLat CHECK ((Latitude >= -90)
    AND (Latitude <= 90)));

CREATE TABLE province
(Name VARCHAR(35) NOT NULL,
Country VARCHAR(4) NOT NULL ,
Population INT,
Area FLOAT,
Capital VARCHAR(35),
CapProv VARCHAR(35),
CONSTRAINT PrKey PRIMARY KEY (Name, Country),
CONSTRAINT PrPop CHECK (Population >= 0),
CONSTRAINT PrAr CHECK (Area >= 0));

CREATE TABLE language
(Country VARCHAR(4),
Name VARCHAR(50),
Percentage FLOAT,
CONSTRAINT LanguageKey PRIMARY KEY (Name, Country),
CONSTRAINT LanguagePercent
CHECK ((Percentage > 0) AND (Percentage <= 100)));
```

FIGURE 4.8 – Cas 1 : Partie du schéma *Mondial* avec quatre tables.

À l'aide des hypergraphes, la formulation du modèle d'optimisation est présentée dans le modèle linéaire (ModLin1).

$$\text{maximiser} \quad \sum_{i=1}^{n_s}\sum_{j=1}^{n_c} s_{ij}x_{ij}$$

$$\text{(ModLin1)} \quad \text{sous contraintes} \quad \sum_{j=1}^{n_c} x_{ij} \leq 1, \qquad \forall i = 1, \cdots, n_s \tag{4.1}$$

$$\sum_{i=1}^{n_s} x_{ij} \leq 1, \qquad \forall j = 1, \cdots, n_c \tag{4.2}$$

$$x_{ij} \in \{0,1\}, \quad \forall i = 1, \cdots, n_s; \forall j = 1, \cdots, n_c \tag{4.3}$$

Avec :

- L'hypergraphe source $HG_s = (V_s, HE_s)$ modélisant le schéma relationnel de la base de données source *Terra* avec $n_s = \mid V_s \mid = 25$, $\mid HE_s \mid = 9$. À noter que le nombre d'(hyper)arête reflète le nombre de clés étrangères + l'(hyper)arête des nœuds isolés.
- L'hypergraphe cible $HG_c = (V_c, HE_c)$ modélisant le schéma relationnel de la base de données cible *Mondial* avec $n_c = \mid V_c \mid = 33$, $\mid HE_s \mid = 12$.
- $G = (V, E)$ le graphe virtuel où V est l'ensemble de tous les nœuds des deux hypergraphes HG_s et HG_c en deux partitions $V_s \cup V_c$, E l'ensemble de toutes les arêtes virtuelles.

Variable de décision : x_{ij} qui est binaire, c'est-à-dire, égale à 1 si l'arête e_{ij} est sélectionnée où les nœuds i et j correspondent à deux tables de *Terra* et *Mondial* respectivement, et 0 sinon.

Fonction objectif : où le paramètre s_{ij} représente la valeur de la similarité totale entre les deux nœuds i et j.

Contraintes : pour ce cas d'étude, les cardinalités retenues sont celles des cardinalités d'appariement des schémas $(1:1)$ qui expriment le fait qu'une table de *Terra* peut correspondre au maximum à une table de *Mondial* et vice versa. Comme le nombre de tables des deux schémas est différent, toutes les tables ne peuvent trouver leur correspondance, ainsi, les contraintes sont sous forme d'inégalités. La dernière contrainte concerne les contraintes d'intégrité des variables qui imposent le choix ou non d'une arête.

Une fois les hypergraphes construits et le modèle d'optimisation spécifié, nous entamons une simulation générale avec les options par défaut résumées dans le Tableau 4.7.

Mesure sémantique pour les noms des tables	*USE & Similarité_Cosinus*
Mesure sémantique pour les noms des attributs	*USE & Similarité_Cosinus*
Mesure syntaxique pour les noms des tables	*Jaccard*
Mesure syntaxique pour les noms des attributs	*Jaccard*
Mesure pour les types de données	*BERT & Similarité_Cosinus*
Mesure pour les informations structurelle	Algorithme 1
Stratégie d'agrégation	Algorithme 3
Stratégie de sélection	Modèle (ModLin1)
Stratégie de combinaison	Somme pondérée
Porte d'entrée	$g = (\text{'}MEER\text{'}, sea)$
Pondération	$\approx 18k$ *vecteurs*

TABLEAU 4.7 – Cas 1 : Les mesures, stratégies et paramètres par défaut pour les simulations

Étant donné que les deux schémas sont fortement hétérogènes en termes de syntaxe, l'un conçu en allemand et l'autre en anglais, il était nécessaire de pouvoir interpréter ces informations. Pour ce faire, nous utilisons *USE* [3] pour représenter les mots par des vecteurs de nombres réels. Cette représentation est possible, car les syntaxes des deux schémas (noms des tables et des attributs) sont reconnaissables, significatives et issues du même contexte. Cette technique permet de représenter des chaînes de caractères dans un environnement sémantique unique basé sur des modèles de traduction pré-entraînés. Cela permet de couvrir l'aspect sémantique et de couvrir les lacunes que peuvent avoir les mesures syntaxiques. Chaque mot aura donc sa propre représentation en un vecteur normalisé.

Pour la similarité des types de données, nous utilisons *BERT* [4] pour représenter les types de données écrits en anglais par des vecteurs de nombres réels. Ensuite, nous utilisons la *Similarité_Cosinus* pour le calcul de la similarité et les résultats sont présentés dans le Tableau 4.8.

	date	**varchar**	**int**	**float**
char	0.55	0.80	0.78	0.66
varchar	0.61	1.00	0.84	0.68
int	0.65	0.84	1.00	0.68
double	0.45	0.72	0.79	0.64

TABLEAU 4.8 – Cas 1 : Similarité des types de données avec *BERT & Similarité_Cosinus*

La porte d'entrée choisie est $g = (\text{'}MEER\text{'}, sea)$. Ce choix a été fait manuellement. D'autres portes d'entrée ont également été testées et les résultats seront présentés par la suite.

4.3.2.1 Résultats généraux

Les simulations sont basées sur le test de $\approx 18k$ compositions différentes des poids appliqués aux mesures de similarité. Ces vecteurs de poids sont de la forme suivante :

$$[w\_local\_table_{sem}, w\_local\_attribut_{sem}; w\_local\_table_{syn}, w\_local\_attribut_{syn}; \\ w\_global_{sem}, w\_global_{syn}, w\_global_{type}, w\_global_{str}]$$

Où :

- $w\_local\_table_{sem}$ et $w\_local\_attribut_{sem}$ correspondent aux poids attribués à la similarité locale sémantique des tables et des attributs avec $w\_local\_table_{sem} + w\_local\_attribut_{sem} = 1$;
- $w\_local\_table_{syn}$ et $w\_local\_attribut_{syn}$ correspondent aux poids attribués à la similarité locale syntaxique des tables et des attributs avec $w\_local\_table_{syn} + w\_local\_attribut_{syn} = 1$;
- w_global_{sem}, w_global_{syn}, w_global_{type} et w_global_{str} correspondent aux poids attribués à la similarité globale sémantique, syntaxique, structurelle et type de données avec $w\_global_{sem} + w\_global_{syn} + w\_global_{type} + w\_global_{str} = 1$.

Ces simulations exhaustives prennent en compte des poids qui varient de 0 à 1 avec des pas égaux à 0.1 et en respectant la propriété que la somme des poids doit être égale à 1 comme présenté dans la Section 3.3.2.4. Nous avons donc $\approx 18k$ résultats pour les mesures d'évaluation (rappel, précision et $F_1 - score$). Ainsi, plusieurs compositions de poids peuvent donner la même valeur de rappel, de précision ou de $F_1 - score$.

Pour ce cas d'étude, le nombre de correspondances à trouver est de 23. Étant donné que la contrainte d'appariement est $(1:1)$, deux tables de la base de données *Terra* et dix tables de

3. https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3
4. https://tfhub.dev/google/collections/bert/1

la base de données *Mondial* n'auront pas de correspondances. Le nombre de correspondances trouvé par le modèle d'optimisation est le nombre de fois que la variable de décision $x_{ij} = 1$

Pour mieux illustrer les résultats, nous avons pris des résultats distincts des mesures d'évaluation. L'objectif de la Figure 4.9 est de montrer l'influence du changement des poids sur les résultats où on peut voir qu'en proposant différentes compositions, nous pouvons avoir des résultats différents.

FIGURE 4.9 – Cas 1 : Mesures d'évaluation des résultats d'appariement pour 22 compositions de poids différentes parmi 18k donnant des résultats différents.

Pour le $F_1 - score$ le meilleur résultat atteint 0.96 pour un rappel de 1.00 ce qui signifie que 100% des correspondances ont été trouvées et pour une précision de 0.91 équivalente à dire que 91% des correspondances trouvées sont correctes. Ces résultats sont obtenus en associant les dimensions sémantiques, syntaxique, type de données et structurelles. Ce résultat peut être comparé aux travaux dans (HÄTTASCH et al., 2022) où, sur le même jeu de données, les expériences consistent à utiliser les données sur les tables, attributs et les instances de données qu'elles portent. Deux paramètres sont utilisés, les variations du seuil et la contrainte du ratio qui détermine la fraction minimale d'attributs qui doivent correspondre pour considérer une table comme candidate à l'appariement. L'approche propose ainsi, pour une table, un ensemble de tables potentielles vérifiant les deux paramètres précédents.

Dans nos expériences, les contraintes de seuil n'ont pas été appliquées. Dans les cas où l'hétérogénéité est fortement présente, l'application d'un seuil peut réduire l'espace des solutions, mais peut également éliminer certaines correspondances positives. Dans ces cas, les valeurs de similarité sont inférieures au seuil appliqué qui est plus élevé et ne sont pas prises en compte. Même si elles reflètent de faibles degrés de similarité, elles peuvent être correctes. Ensuite, la suggestion de plusieurs candidats implique la sélection manuelle de la

bonne correspondance parmi les suggestions, ce qui se fait au détriment du rappel en raison des fausses correspondances suggérées (faux positifs).

La fonction objectif nous permet une optimisation de la similarité globale en proposant les meilleures correspondances qui n'altèrent pas la qualité globale de la ou les solutions proposées. Dans le cas de problèmes à grande échelle, la suggestion peut être lourde pour un utilisateur. La proposition d'une solution de meilleure qualité globale est une alternative à cela puisque nous n'essayons pas de trouver les meilleures correspondances individuellement. Des résultats de bonne qualité peuvent être obtenus automatiquement en utilisant les modèles d'optimisation.

En somme, sur les $\approx 18k$ vecteurs de poids, seulement deux vecteurs donnent la meilleure valeur du $F_1 - score = 0.96$ qui sont :

$$[w\_local\_table_{sem}, w\_local\_attribut_{sem}; w\_local\_table_{syn}, w\_local\_attribut_{syn}; \\ w\_global_{sem}, w\_global_{syn}, w\_global_{type}, w\_global_{str}] = [0.7, 0.3; 0.3, 0.7; 0.6, 0.1, 0.0, 0.3]$$

Et :

$$[w\_local\_table_{sem}, w\_local\_attribut_{sem}; w\_local\_table_{syn}, w\_local\_attribut_{syn}; \\ w\_global_{sem}, w\_global_{syn}, w\_global_{type}, w\_global_{str}] = [0.7, 0.3; 0.3, 0.7; 0.7, 0.1, 0.0, 0.2]$$

La Figure 4.10 montre les variations des types de poids des deux vecteurs.

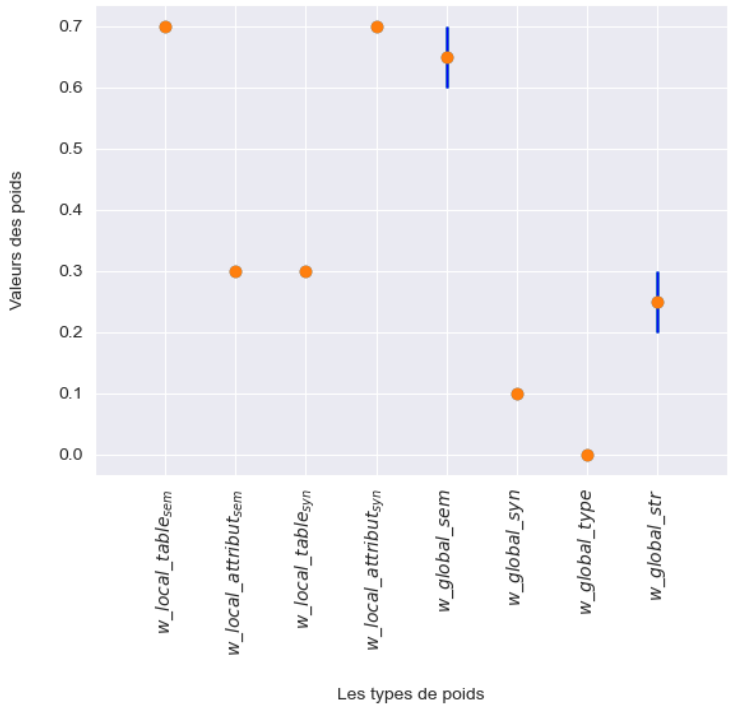


FIGURE 4.10 – Cas 1 : Variations des poids pour $F_1 - score = 0.96$ des deux vecteurs.

Nous pouvons constater que pour les poids locaux sémantiques, le poids sur les noms des tables est à 0.7 et en parallèle, le poids sur les noms des attributs est à 0.3. Pour les poids locaux syntaxiques, le poids sur les noms des tables est de 0.3 et le poids sur les noms des attributs est de 0.7. Pour les poids globaux, le poids global sémantique varie de 0.6 à 0.7, le

poids global structurel varie de 0.2 à 0.3. Enfin, le poids global syntaxique est de 0.1 tandis que le poids des types de données est de 0.0.

Ces observations sont ainsi analysées en détail dans la prochaine section en analysant le $F_1 - score$ qui tient compte à la fois de la précision et du rappel.

4.3.2.2 Analyse détaillée des résultats

Dans cette analyse détaillée, nous comparons l'importance et l'influence des noms de tables et des noms d'attributs, puis des données sémantiques, syntaxiques, structurelles et des types de données avec les paramètres par défaut du Tableau 4.7.

L'analyse va se faire sur les résultats du $F_1 - score$. Une valeur du $F_1 - score$ élevée signifie qu'on a une bonne précision et un bon rappel. Inversement, une valeur du $F_1 - score$ faible signifie qu'on a une mauvaise précision ou un mauvais rappel, ou les deux. Les résultats concernent donc des simulations des $\approx 18k$ vecteurs de poids. Pour plus de lisibilité, nous regroupons les vecteurs par catégorie de poids (poids locaux sémantique, poids locaux syntaxique et poids globaux) et nous prenons la valeur maximale des mesures d'évaluation. Ainsi, nous allons comparer, dans un premier temps, les poids sémantiques sur les noms des tables et des attributs, puis dans un deuxième temps, les poids syntaxiques sur les noms des tables et des attributs, et enfin, les poids sémantiques, syntaxiques, types de données et structurelles.

Poids sémantiques des tables et des attributs : la Figure 4.11 montre les valeurs maximales du $F_1 - score$ pour des vecteurs de poids :

$$[w\_local\_table_{sem}, w\_local\_attribut_{sem}]$$

On peut observer que la valeur maximale du $F_1 - score$ est atteinte pour des valeurs de $w\_local\_table_{sem} = 0.7$ avec en parallèle des valeurs de $w\_local\_attribut_{sem} = 0.3$.

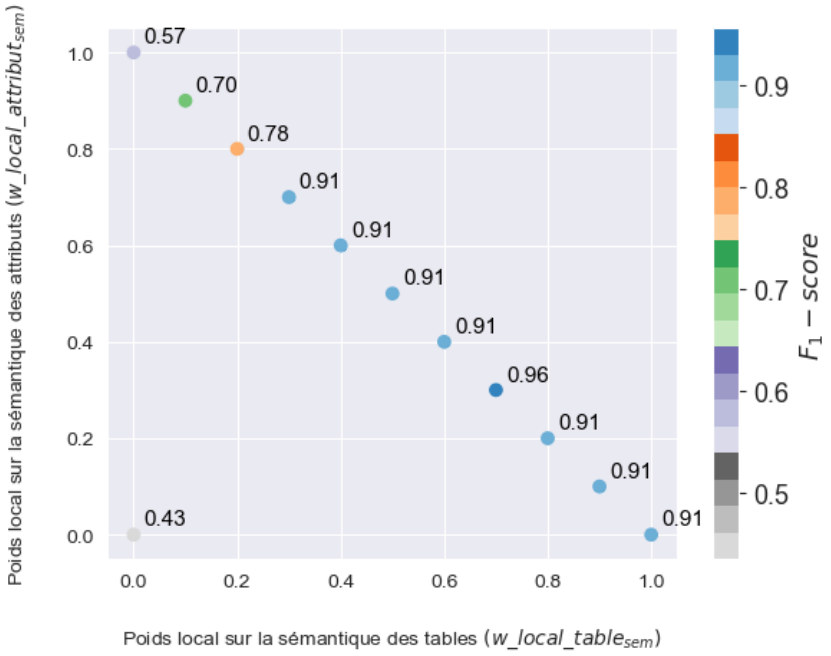


FIGURE 4.11 – Cas 1 : Valeurs du $F_1 - score$ en fonction des valeurs des poids sémantiques sur les noms des tables et les noms des attributs.

Ainsi, on peut constater que, plus on donne un poids élevé pour la sémantique des noms des tables, plus on améliore les mesures d'évaluation. En variant $w\_local\_table_{sem}$ de 0.0 à 0.7, le $F_1 - score$ est amélioré, à l'inverse des poids sur la sémantique des attributs où le $F_1 - score$ se dégradent en augmentant $w\_local\_attribut_{sem}$ au-delà de 0.3 ou en diminuant en dessous de 0.3. La sémantique des noms des tables présente donc des pondérations plus élevées et indiquent un effet important dans la valeur du $F_1 - score$. En donnant un poids $[w\_local\_table_{sem}, w\_local\_attribut_{sem}] = [0.0, 0.0]$, c'est-à-dire, en enlevant la similarité sémantique, la valeur du $F_1 - score = 0.43$.

Poids syntaxiques des tables et des attributs : Figure 4.12 montre les valeurs maximales du $F_1 - score$ pour des vecteurs de poids :

$$[w\_local\_table_{syn}, w\_local\_attribut_{syn}]$$

En analysant de la même manière ces résultats, on peut constater qu'à l'inverse des poids locaux sémantiques, le meilleur résultat est donné avec un poids moins élevé sur la syntaxe des noms des tables avec $w\_local\_table_{syn} = 0.3$ et plus de poids sur la syntaxe des noms des attributs avec $w\_local\_attribut_{syn} = 0.7$.

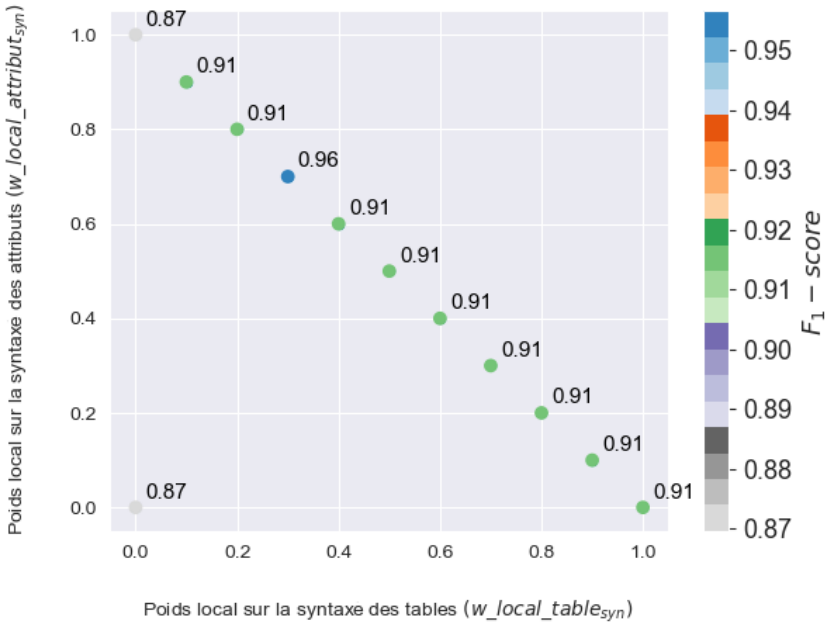


FIGURE 4.12 – Cas 1 : Valeurs du $F_1 - score$ en fonction des valeurs des poids syntaxiques sur les noms des tables et les noms des attributs.

En analysant les variations du $F_1 - score$ on peut voir qu'en augmentant les poids sur les noms des tables jusqu'à $w\_local\_table_{syn} = 0.3$, la valeur du $F_1 - score$ s'améliore. En parallèle, les poids syntaxiques sur les noms des attributs doivent être $w\_local\_attribut_{syn} \geq 0.7$. Ainsi, la syntaxe des noms des tables a moins d'influence que la syntaxe des noms des attributs.

En somme, l'importance de la sémantique des noms des tables par rapport à la sémantique des noms des attributs vient du fait que les noms des tables peuvent être interprétés par les techniques de plongement. Parallèlement, l'importance de la syntaxe des noms des attributs par rapport à la syntaxe des noms des tables vient du fait que dans certaines tables, les noms des attributs partagent des caractères en communs.

Pour illustrer, nous reprenons les deux tables dans la Figure 4.6 et la Figure 4.7 qui représentent une correspondance vraie et nous résumons les valeurs de la similarité sémantique

et syntaxique sur les noms des tables et les valeurs de la similarité sémantique et syntaxique agrégés sur les noms des attributs dans le Tableau 4.9.

		Sémantique	Syntaxe
'BERG'	mountain	0.81	0.00
'B_NAME' 'GEBIRGE' 'HOEHE' 'JAHR' 'LAENGE' 'BREITE'	Name Mountains Height Type Longitude Latitude	0.55	0.31

TABLEAU 4.9 – Cas 1 : Similarité sémantique et syntaxique sur les noms des tables et les noms des attributs de 'BERG' de la base de données *Terra* et mountain de la base de données *Mondial*

La similarité sémantique des noms des tables (0.81) est plus élevée que la similarité syntaxique des noms des tables (0.00). Les deux tables ne partagent pas la même syntaxe, mais ont une similarité sémantique qui les rend proches. Parallèlement, on ne note pas autant de différence pour la similarité agrégée sémantique (0.55) et syntaxique (0.31) des noms des attributs. Les deux tables partagent des attributs en commun proche syntaxiquement à l'image des attributs *'B_NAME'* et *Name*. Cependant, ce cas n'est pas le cas dans toutes les correspondances trouvées. En effet, en prenant les deux tables, 'GEO_BERG' de la base de données *Terra* et geo_mountain de la base de données *Mondial*. Nous résumons les valeurs de la similarité dans le Tableau 4.10.

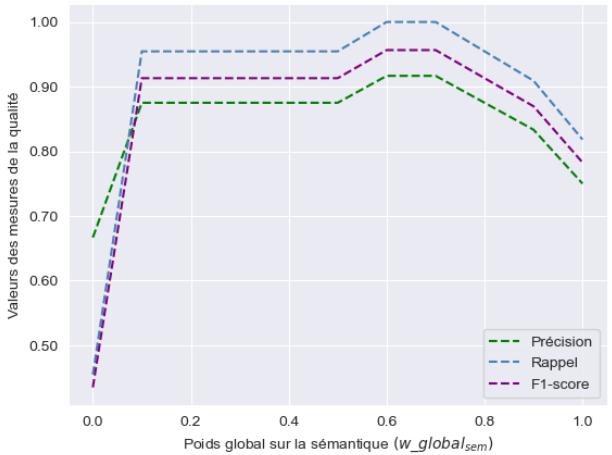
		Sémantique	Syntaxe
'GEO_BERG'	geo_mountain	0.91	0.23
'ID' 'L_ID' 'LT_ID' 'B_NAME'	Mountain Country Province	0.20	0.12

TABLEAU 4.10 – Cas 1 : Similarité sémantique et syntaxique sur les noms des tables et les noms des attributs de 'GEO_BERG' de la base de données *Terra* et de geo_mountain de la base de données *Mondial*

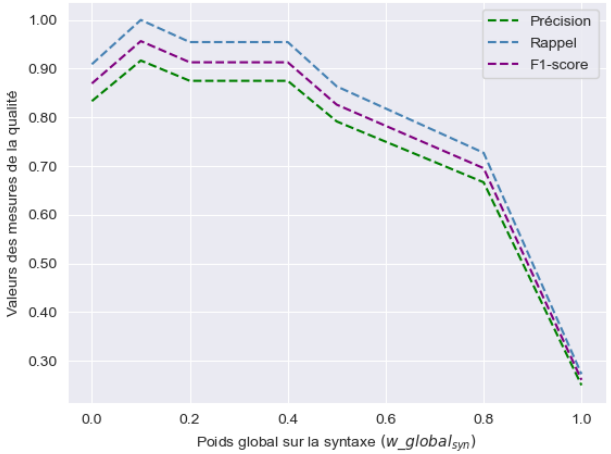
On constate que dans la table 'GEO_BERG' on a des attributs avec des noms abrégés ou avec des préfixes, et des attributs portant le nom (*ID*). Ces deux spécificités ne sont pas présentes dans les tables de la base de données *Mondial* ce qui rend la similarité syntaxique sur les noms des attributs moins performante pour trouver certaines correspondances.

Poids de la similarité globale : l'objectif des deux premières analyses était de voir l'influence des poids locaux sur les noms des tables et les noms des attributs pour trouver les correspondances en comparant les poids sémantiques et syntaxiques qui maximisent le $F_1 - score$. Pour cette analyse, nous allons prendre les vecteurs de poids globaux qui décrivent mieux l'influence des poids sur les résultats. La Figure 4.13 présente les valeurs maximales des mesures d'évaluation pour les vecteurs de poids :

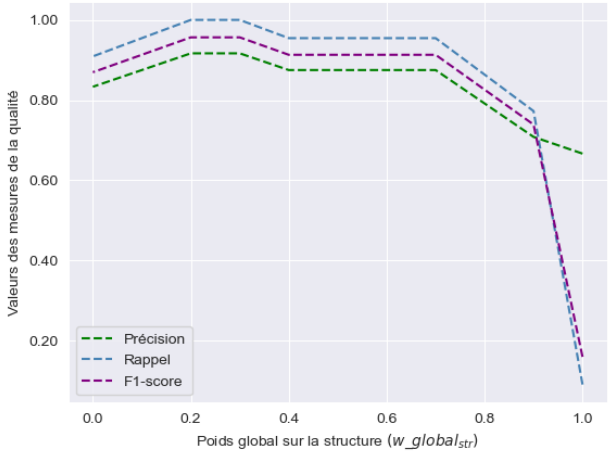
$$\left[w\_global_{sem}, w\_global_{syn}, w\_global_{type}, w\_global_{str}\right]$$



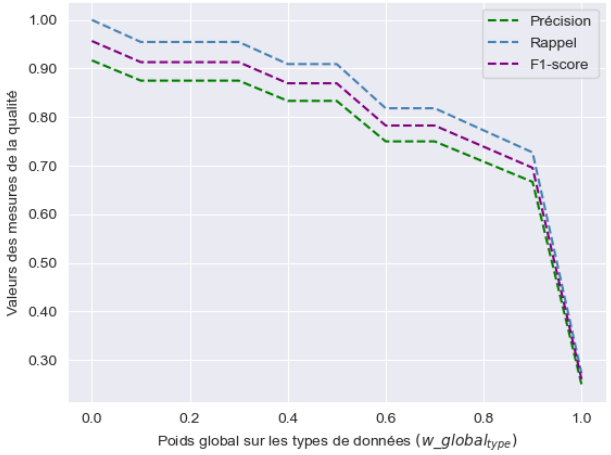
(a) Cas 1 : Mesures d'évaluation pour les variations du poids global sémantique w_global_{sem}



(b) Cas 1 : Mesures d'évaluation pour les variations du poids global syntaxique w_global_{syn}



(c) Cas 1 : Mesures d'évaluation pour les variations du poids global structurel w_global_{str}



(d) Cas 1 : Mesures d'évaluation pour les variations du poids global des types de données w_global_{type}

FIGURE 4.13 – Cas 1 : Mesures d'évaluation pour les variations des poids globaux.

En analysant ces résultats, on peut constater que seule la variation du poids relatif à la similarité sémantique w_global_{sem} améliore significativement les mesures d'évaluation. Ainsi, la valeur de ce poids reste très élevée et doit être entre 0.6 et 0.7 alors que pour le reste des poids (w_global_{syn}, w_global_{type} et w_global_{str}), la somme de leurs valeurs doit rester ≤ 0.4. Ces poids contribuent en général avec moins d'impact sur les résultats. La raison est que la syntaxe des schémas est très hétérogènes et ne présente pas assez de similitudes sur l'ensemble des noms des tables et noms des attributs sauf dans très peu de cas comme nous l'avons montré dans la précédente analyse.

Ensuite, la structure est construite à l'aide des contraintes sur les clés étrangères qui peuvent être différentes d'un schéma à un autre. Ceci rend la topologie des hypergraphes résultants très hétérogènes. Dans l'Algorithme 1, nous avons défini les *portes_d'entrée*. En faisant des simulations sur les différentes portes d'entrées possibles, nous pouvons observer que le choix de ce paramètre est significativement important dans le résultat final. Le Tableau 4.11 présente les meilleurs résultats des simulations en changeant les *portes_d'entrée*.

portes_d'entrée	***P***	***R***	$F_1 - \textbf{\textit{score}}$
g = ('STADT', city)	0.79	0.86	0.82
g = ('GEO_SEE', geo_lake)	0.87	0.95	0.91
g = ('SEE', lake)	0.83	0.90	0.86

TABLEAU 4.11 – Cas 1 : Mesures d'évaluation pour différentes *portes_d'entrée*

Le but des simulations exhaustives sur les portes d'entrée étaient de voir l'impact que peut avoir ce choix sur les résultats. Cependant, comme expliqué dans la Section 3.3.1.2, le choix peut être soutenu par une stratégie.

Enfin, pour la similarité des types de données, la valeur du poids w_global_{type} est égale à zéro. Cela signifie que les types de données ne contribuent pas dans le meilleur résultat, et cela peut être causé d'un côté par l'hétérogénéité liée aux types de données, car différents types sont utilisés pour exprimer la même information, et certains types de données sont utilisés dans un schéma et pas dans l'autre, à l'image du type de données *date*. Le poids w_global_{type} n'est donc pas pertinent pour une évaluation globale pour la fonction objectif du modèle d'optimisation et est dominé par les autres poids. Le modèle d'optimisation tente donc de trouver la meilleure solution qui maximise la fonction objectif en fonction du vecteur de poids. Le vecteur de poids maximisant les mesures d'évaluation reflète donc un compromis entre les poids.

4.3.3 Cas d'étude 2 : Bases de données industrielles

Pour ce deuxième cas, nous utilisons une partie de deux bases de données réelles issues de deux ERP industriels fournit par *Forterro Sylob*. La première base de données (*BDS* pour base de données source) se caractérise par l'utilisation d'abréviations difficilement interprétables, ainsi que par l'absence de la structure liée à la déclaration des clés primaires et étrangères dans la majorité des tables. La seconde base de données (*BDC* pour base de données cible) a été récemment développée et est destinée à remplacer la première. Elle se caractérise par une structuration bien définie des relations ainsi que par la clarté de la syntaxe des noms de tables et des attributs. Le Tableau 4.12 résume les statistiques relatives aux informations présentent dans les deux schémas des deux bases de données.

Les deux schémas présentent des sources d'hétérogénéité à plusieurs niveaux, notamment, en termes de nombre d'attributs, de clés ainsi qu'en termes de syntaxe. La Figure 4.14 et la Figure 4.15 sont des tables extraites des deux bases de données qui représentent les principales données d'un article.

Schéma	# tables	# attributs	# clés primaires	# clés étrangères
BDS	11	186	0	0
BDC	11	101	11	11

TABLEAU 4.12 – Cas 2 : Statistiques des bases de données *BDS* et *BDC*

```
create table "root".bas_art(
    no_art char(17) not null,
    design1 char(40),
    design2 char(80),
    typ_art char(1),
    cd_famart char(7),
    cd_famcom char(7),
    cd_quest char(10),
    no_form char(10),
    dte_cre date,
    dte_dernmvt date,
    val_sort float,
    cns1 char(30),
    cns2 char(30),
    gest_stk smallint,
    nb_cartons integer,
    ...
    nb_pieces_car float);
```

FIGURE 4.14 – Cas 2 : Déclaration de la table *bas_art* au sein du schéma *BDS*.

```
CREATE  TABLE "public".dm1_article (
    id                    varchar(60)  NOT NULL  ,
    datefinvalidite       timestamp    ,
    conversiondeb         numeric(31,8)    ,
    nomenclatureproduit   varchar(255)    ,
    code                  varchar(255)    ,
    commentaire           varchar(2000)    ,
    designation           varchar(255)    ,
    indicevalide          varchar(255)    ,
    origine               varchar(255)    ,
    poidsunitaire_valeur  numeric(31,8)    ,
    prixrevient_valeur    numeric(31,8)    ,
    typearticle           varchar(255)    ,
    id_famillearticle     varchar(60)    ,
    id_compteurnumeroserie varchar(60)    ,
    id_etablissementapprovisionnement varchar(60)    ,
        ...
        CONSTRAINT dm1_article_pkey PRIMARY KEY ( id )
 );
CREATE INDEX ix_20c8043b5a066fe039e5bc4f90e553ed ON
"public".dm1_article  ( strategieproduction );
...
ALTER TABLE "public".dm1_article ADD CONSTRAINT
fkee9cb07fa235f9da FOREIGN KEY ( id_compteurnumeroserie )
REFERENCES "public".dm1_compteurserielot( id );
...
```

FIGURE 4.15 – Cas 2 : Déclaration de la table *dm1_article* au sein du schéma *BDC*.

Comme pour le premier cas d'étude, les correspondances de référence sont construites manuellement et les données sont extraites et organisées en tables, attributs, types de données et clés. Les opérations de prétraitement concernent principalement la suppression de certains attributs spécifiques créés à la base pour les systèmes qui gèrent ces bases de données et qui n'ont pas de correspondance, par exemple, l'attribut *datefinvalidite*. Aussi, étant donné que les clés primaires et étrangères sont absentes de la base de données source (BDS), nous avons procédé au calcul de la distribution des fréquences des données des attributs en se basant sur le principe d'unicité des clés primaires. Ensuite, nous utilisons la clé primaire pour construire les clés étrangère. Les données sur les schémas sont résumées comme suit dans le Tableau 4.13.

Schéma	# tables	# attributs	# clés primaires	# clés étrangères
BDS	11	186	11	9
BDC	11	101	11	11

TABLEAU 4.13 – Cas 2 : Statistiques des bases de données *Mondial* et *Terra* après prétraitement

Ainsi, à l'aide de ces données, les hypergraphes sont construits. Le modèle d'optimisation est le même que celui utilisé pour le premier cas d'étude (ModLin1) avec aussi des contraintes d'appariement $(1:1)$, où :

— Hypergraphe source $HG_s = (V_s, HE_s)$ modélisant le schéma relationnel de la base de données source *BDS* avec $n_s = \mid V_s \mid = 11, \mid HE_s \mid = 10$.

— Hypergraphe cible $HG_c = (V_c, HE_c)$ modélisant le schéma relationnel de la base de données cible *BDC* avec $n_c = \mid V_c \mid = 11, \mid HE_s \mid = 12$.

— $G = (V, E)$ le graphe virtuel où V est l'ensemble de tous les nœuds des deux hypergraphes HG_s et HG_c en deux partitions $V_s \cup V_c$, E l'ensemble de toutes les arêtes virtuelles.

Pour la similarité des types de données, le modèle *BERT* est utilisé avec la *Similarité_Cosinus* et les résultats sont présentés dans le tableau Tableau 4.14.

	integer	boolean	timestamp without time zone	character varying	numeric
integer	1.00	0.62	0.46	0.57	0.76
char	0.63	0.67	0.34	0.67	0.60
smallint	0.59	0.62	0.50	0.66	0.57
float	0.51	0.61	0.32	0.55	0.57
date	0.53	0.47	0.45	0.50	0.61
decimal	0.81	0.65	0.51	0.60	0.81

TABLEAU 4.14 – Cas 2 : Similarité des types de données avec *BERT & Similarité_Cosinus*

Ainsi, pour mieux voir l'impact du choix des mesures de similarité, nous n'allons pas prendre en compte la dimension sémantique et nous considérons les options résumées dans le Tableau 4.15.

4.3.3.1 Résultats généraux

Les simulations sont basées sur le test de 682 compositions différentes des poids qui varient de 0 à 1 avec des pas égaux à 0.1. Nous avons donc 682 résultats pour les mesures d'évaluation (rappel, précision et $F_1 - score$) et plusieurs compositions donnent la même valeur de rappel, de précision ou de $F_1 - score$. La Figure 4.16 montre l'influence du changement des poids sur les résultats où on peut observer qu'avec différentes compositions de poids, nous pouvons avoir des résultats différents.

Le meilleur résultat atteint pour $F_1 - score$ est de 1.00 et le rappel et la précision correspondants sont aussi de 1.00, ce qui se traduit par le fait que toutes les correspondances

Mesure sémantique pour les noms des tables	non appliquée
Mesure sémantique pour les noms des attributs	non appliquée
Mesure syntaxique pour les noms des tables	*Jaro_Winkler*
Mesure syntaxique pour les noms des attributs	*Jaro_Winkler*
Mesure pour les types de données	*BERT & Similarité_Cosinus*
Mesure pour les informations structurelle	Algorithme 1
Stratégie d'agrégation	Algorithme 3
Stratégie de sélection	Modèle (ModLin1)
Stratégie de combinaison	Somme pondérée
Porte d'entrée	*g = (bas_contcli, dm1_contact)*
Pondération	*682 vecteurs*

TABLEAU 4.15 – Cas 2 : Les mesures, stratégies et paramètres par défaut pour les simulations

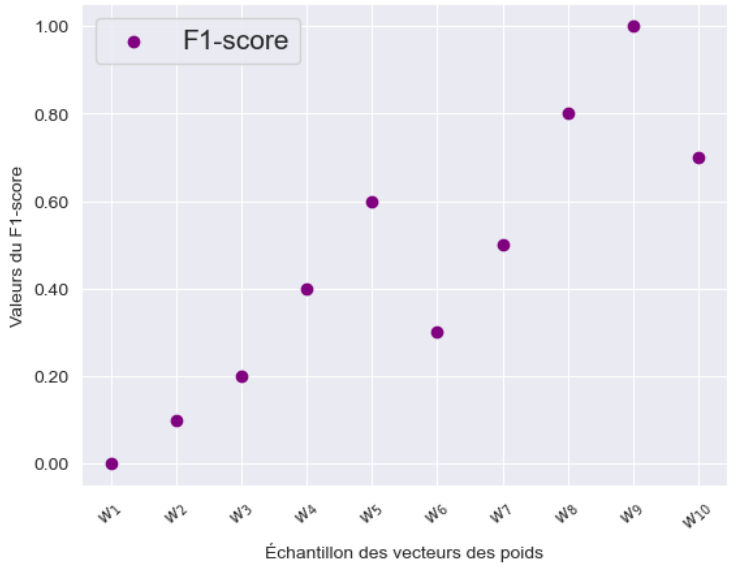


FIGURE 4.16 – Cas 2 : $F_1 - score$ des résultats d'appariement pour 10 compositions de poids différentes parmi 682 donnant des résultats différents.

sont trouvées et que toutes les correspondances trouvées sont correctes. En effet, le modèle d'optimisation nous a permis de trouver une correspondance pour chacune des tables sans avoir recours à la mesure de la similarité sémantique, et ceci étant possible, car les deux schémas présentent de fortes similitudes syntaxiques.

La Figure 4.17 montre les variations des types de poids pour des vecteurs donnant $F_1 - score = 1.00$.

Nous pouvons constater que pour les poids locaux syntaxiques, le poids sur les noms des tables est entre 0.2 et 1.0, en parallèle, le poids sur les noms des attributs est entre 0.0 et 0.8. Pour les poids globaux, le poids global syntaxique varie de 0.1 à 0.9, le poids global structurel varie de 0.1 à 0.7. Enfin, le poids des types de données varie de 0.0 à 0.4. Pour plus de détails, nous comparons l'importance de la syntaxe des noms de tables et des noms d'attributs, puis l'importance des informations syntaxiques, structurelles et des types de données.

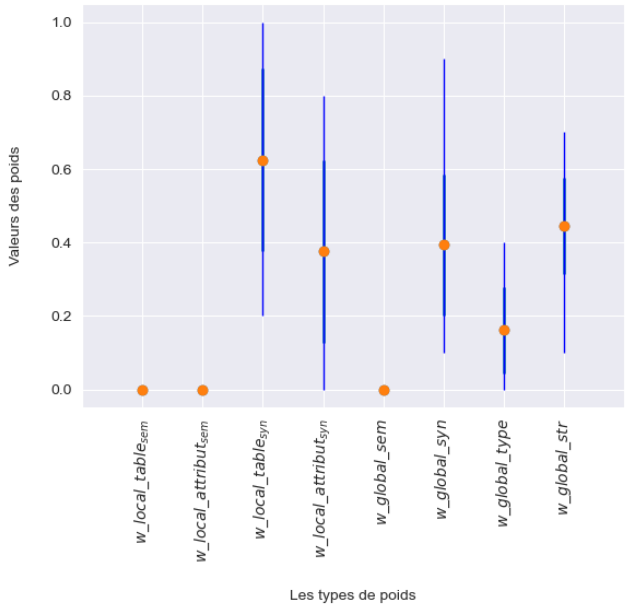


FIGURE 4.17 – Cas 2 : Variations des poids pour $F_1 - score$.

4.3.3.2 Analyse détaillée des résultats

Pour ce cas d'étude, les vecteurs des poids sont :

$$[w\_local\_table_{syn}, w\_local\_attribut_{syn};$$
$$w\_global_{syn}, w\_global_{type}, w\_global_{str}]$$

Poids syntaxiques des tables et des attributs : la Figure 4.18 montre les valeurs maximales du $F_1 - score$ pour des vecteurs de poids $[w\_local\_table_{syn}, w\_local\_attribut_{syn}]$.

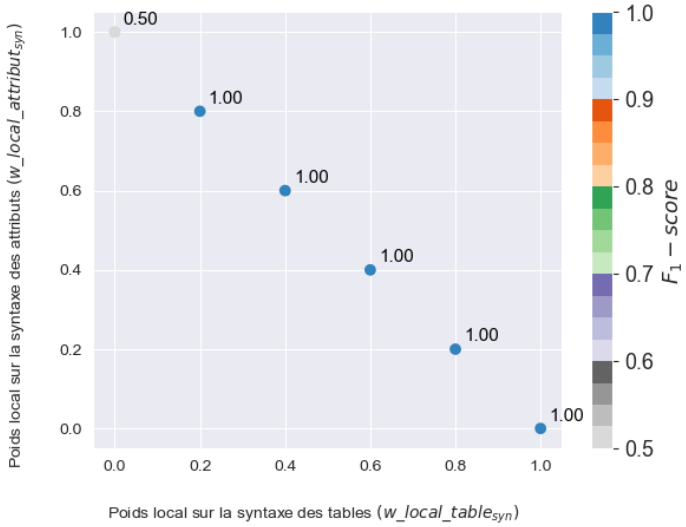


FIGURE 4.18 – Cas 2 : $F_1 - score$ pour les variations des poids syntaxiques sur les tables et les attributs.

On constate qu'on obtient la valeur $F_1 - score = 1.0$ pour des poids sur les noms des tables ≥ 0.2, comparé aux poids sur les noms des attributs qui doivent être ≤ 0.8 voir $= 0.0$. Ainsi, les noms des tables présentent les pondérations les plus importantes dans la détermination des correspondances. L'importance des noms de tables vient du fait que d'une part, il existe une différence en termes de nombre d'attributs dans les bases de données où plusieurs attributs ne peuvent pas être appariés et d'autre part, les schémas des bases de données présentent de fortes similitudes syntaxiques puisque dans le schéma de la base de données source les acronymes utilisés pour les noms des tables sont pour certains dérivés ou constitués de plusieurs abréviations de mots complets présents dans le schéma de la base de données cible. Par exemple, pour les deux tables *dm1_article* et *bas_art*, *art* est une sous-chaîne de *article*. À cet effet, l'aspect syntaxique prime également sur les deux autres mesures de similarité, à savoir les types de données et la structure, comme on va le voir dans la prochaine analyse.

Poids de la similarité globale : la Figure 4.19, la Figure 4.20 et la Figure 4.21 montrent les valeurs du $F_1 - score$ en fonction des variations des poids des vecteurs de poids

$$[w\_global_{syn}, w\_global_{type}, w\_global_{str}]$$

On peut remarquer que pour les meilleurs résultats du $F_1 - score$, le poids global relatif au calcul syntaxique w_global_{syn} doit être ≥ 0.1. Entre 0.1 et 1.0 la valeur du $F_1 - score$ est constante, car toutes les correspondances ont été trouvées et toutes les correspondances trouvées sont vraies.

Pour les poids structurels et les poids des types de données, la variation des poids n'améliorent pas les valeurs du $F_1 - score$. En effet, les résultats sont constants et égales à 1.0 quand les poids w_global_{type} et w_global_{str} sont ≤ 0.6 et ≤ 0.7 respectivement. Ceci peut s'expliquer par la différence structurelle entre les deux schémas des bases de données en termes de définition de la notion des clés et la construction de relations différentes. Ainsi, pour deux tables ayant une syntaxe similaire, leurs voisinages peuvent être très différents étant donné que le schéma de la base de données cible a pour ambition d'améliorer les lacunes du schéma de la base de données source et de mieux représenter la réalité. Dans la même lignée, la similarité des types de données ne contribue pas car différents types de données sont utilisés dans les deux schémas.

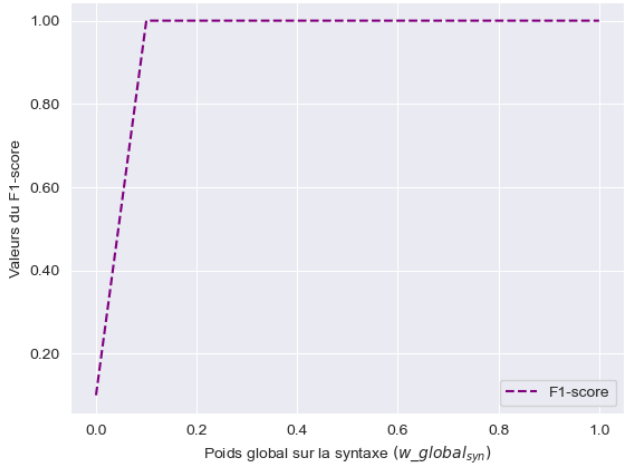


FIGURE 4.19 – Cas 2 : $F_1 - score$ en fonction des variations du poids global syntaxique w_global_{syn}.

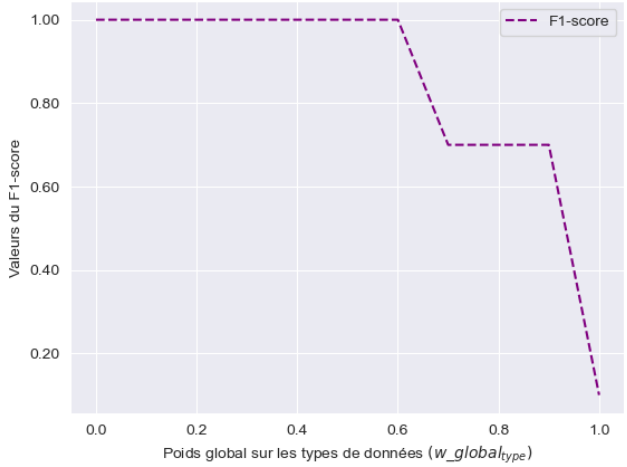


FIGURE 4.20 – Cas 2 : $F_1 - score$ en fonction des variations du poids global des types de données w_global_{type}.

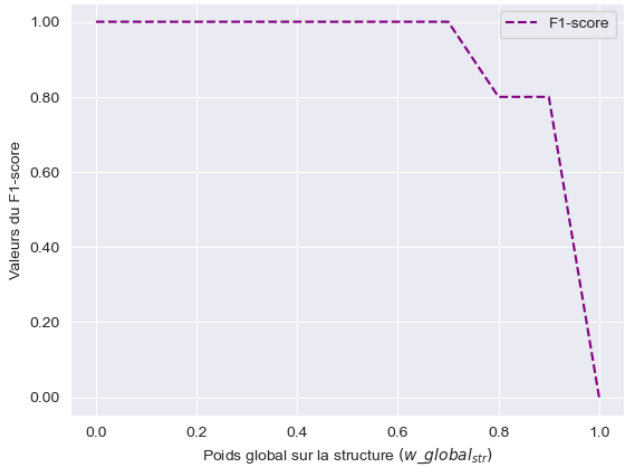


FIGURE 4.21 – Cas 2 : $F_1 - score$ en fonction des variations du poids global sur la structure w_global_{str}.

4.3.4 Évaluation du temps d'exécution

Le temps d'exécution de l'approche a été évalué sur trois niveaux, (1) prétraitement et extraction des informations, (2) calcul de la similarité et processus d'agrégation, et (3) processus de sélection. Le Tableau 4.16 résume les résultats.

	Hypergraphe		Graphe		Vecteurs	(1)	(2)	(3)
	Source	Cible	Nœuds	Arêtes				
Cas 1	(25, 9)	(33, 12)	58	825	$\approx 18k$ $= 682$	$\approx 10(s)$	$\approx 40(s)$ $\approx 5(s)$	$\approx 20(m)$ $\approx 40(s)$
Cas 2	(11, 10)	(11, 12)	22	121	$\approx 18k$ $= 682$	$\approx 10(s)$	$\approx 15(s)$ $\approx 5(s)$	$\approx 5(m)$ $\approx 1(m)$

TABLEAU 4.16 – Temps d'exécution de l'approche sur les deux cas d'études

Les simulations ont été réalisées sur un ordinateur portable exécutant Python 3.9.13 64 bits sur 4 cœurs Intel i5-8350U CPU @ 1.70GHz. La mémoire principale est de 16 Go. Le système d'exploitation était Windows 10 64 bits.

Le tableau reflète donc les temps d'exécution pour les deux cas d'études avec deux configurations de vecteurs. Les temps de prétraitement et d'extraction des informations sont de l'ordre de *10 secondes*. Pour le calcul de la similarité et le processus d'agrégation, le temps d'exécution est de l'ordre de *40 secondes* pour le premier cas d'étude et de *15 secondes* pour le second cas d'étude avec la configuration de $\approx 18k$ vecteurs ($\approx 18k$ exécutions). Pour la deuxième configuration des poids, les temps d'exécution sont de l'ordre de *5 secondes* pour les deux cas d'étude (682 exécutions). Cette différence est due à la différence dans la dimension des matrices de similarité calculées en fonction du nombre de nœuds de chaque hypergraphe. Ainsi, pour le premier cas, la matrice est d'ordre (25×33) correspondant à 825 valeurs et pour le deuxième cas, la matrice est d'ordre (11×11) correspondant à 121 valeurs. Par ailleurs, la prise en compte du calcul sémantique par la technique de plongement qui correspond à 70% du temps total de calcul de la similarité. Les temps d'agrégation des matrices sont négligeables par rapport au reste. Enfin, pour le processus de sélection, la configuration avec $\approx 18k$ vecteurs consomme plus de temps, car le modèle d'optimisation est exécuté $\approx 18k$ fois, ce qui est un temps de résolution raisonnable face aux nombres de fois où le modèle est résolu pour obtenir une solution optimale. Ainsi, l'approche nous fournit un ensemble de solutions optimales globales avec différents vecteurs de poids.

4.4 Conclusion

Dans ce chapitre, avons présenté dans un premier temps l'environnement d'implémentation de l'approche pour l'appariement des schémas pour contribuer à l'interopérabilité des données. Ensuite, dans un second temps, nous avons expérimenté cette dernière sur deux cas d'étude avec comme objectif principal l'évaluation de sa flexibilité, sa globalité et sa généricité.

Pour ce faire, nous avons mené un ensemble de simulations avec des mesures de similarité sur quatre types dimensions, à savoir, la sémantique, la syntaxe, les types de données et la structure, et guidée avec différents vecteurs de poids en utilisant des stratégies d'agrégation, de sélection et des mesures de similarité. Au travers des expérimentations et l'analyse des résultats, l'approche nous a permis de prendre en compte les variations du problème d'appariement, telles que l'indisponibilité des relations ou l'interprétation des données.

Enfin, l'approche propose un environnement capable d'intégrer et de rassembler plusieurs mesures de similarité et de trouver un équilibre entre les différents poids sur les dimensions qui maximise la fonction objectif qui traduit la similarité globale.

5

Conclusion

Les travaux de cette thèse abordent la question de l'interopérabilité en général et l'interopérabilité des données en particulier. Nous avons traité ces questions avec deux principaux objectifs.

Le premier objectif a été abordé sous l'angle d'un travail de recherche, de synthèse et de proposition bibliographique avec l'ambition de contribuer à la théorie de l'interopérabilité en proposant une conceptualisation des connaissances autour de cette notion ainsi que de proposer une orientation des recherches et solutions futures vers l'implémentation d'une approche fédérée.

Le deuxième objectif a été abordé sous un angle applicatif et industriel avec l'ambition de développer une approche pour la mise en œuvre de l'interopérabilité des données par l'implémentation d'une approche fédérée dans le cadre d'une migration de base de données.

Dans la suite, nous rappelons et synthétisons les principales contributions, les perspectives d'améliorations, et les futures directions que pourraient prendre ces travaux.

5.1 Synthèse

Dans le Chapitre 1 nous avons mené une recherche bibliographique sur l'interopérabilité, ces concepts et les domaines qui lui sont adjacents en repartant du postulat que l'interopérabilité est une notion multidimensionnelle et transdisciplinaire. Le sujet de l'interopérabilité a été largement abordé dans la littérature et dans plusieurs domaines d'application à travers les nombreuses avancées, développements, normes et concepts proposés traduisant son importance tant académique que socioéconomique. Ceci a été motivé par la croissance des besoins en termes d'interopérabilité, soutenu par l'essor des nouvelles technologies de l'information et de la communication d'une part et, d'autre part, l'explosion du volume et de la complexité des données. L'étude de l'état de l'art a montré une fragmentation dans la définition et la compréhension des notions de l'interopérabilité ainsi que les solutions proposées, qui sont parfois spécifiques à un problème et sont difficiles à reproduire. Encouragés par l'étude et l'analyse de cet état de l'art, nous avons proposé un cadre général qui rassemble les différents concepts d'interopérabilité proposés dans la littérature. Ce cadre propose une suite d'étapes pour assister le processus de mise en œuvre de l'interopérabilité en s'appuyant sur des bases

théoriques et pratiques. Ensuite, en analysant les besoins et défis actuels et futurs, nous nous sommes orientés vers l'implémentation de l'interopérabilité en utilisant une approche fédérée qui par définition n'impose pas un format de données commun et les entités (base de données, logiciels, systèmes, composants de systèmes, simulateurs, entreprise, organisation, etc.) impliquées dans le processus d'interopérabilité doivent s'adapter à la volée. L'analyse de l'état de l'art sur les approches fédérées montre que l'application de ces approches est souvent conseillée, mais moins couramment utilisées que les approches unifiées ou intégrées. L'interopérabilité fédérée est nécessaire lorsque différentes entités hétérogènes qui ont des capacités techniques différentes, fonctionnent avec des technologies différentes, ou évoluent dans des environnements dynamiques doivent partager des données ou des ressources afin d'atteindre un objectif commun où il n'est pas pratique ou possible de centraliser les données par un modèle commun. Les approches fédérées qui sont proposées sont souvent basées sur l'utilisation des bases de connaissances, des ontologies et des métamodèles, cependant, ces concepts peuvent être complexes à créer et à maintenir et nécessitent des mises à jour au fur et à mesure que des changements environnementaux et technologiques apparaissent. Ainsi, ces concepts ne suffisent pas toujours surtout dans des cas où les entités doivent interopérer en temps réel avec de nouvelles entités qui peuvent être non connues à l'avance. Au final, nous nous sommes recentrés sur une des dimensions de l'interopérabilité qui est l'interopérabilité des données et nous avons proposé une approche implémentant une interopérabilité fédérée entre des bases de données.

Le Chapitre 2 s'est concentré sur l'importance de l'interopérabilité des données qui est une condition nécessaire à la réalisation de l'interopérabilité en général et une étape importante pour répondre aux besoins actuels et futurs. En s'appuyant sur les analyses et conclusions du Chapitre 1, nous avons adressé et présenté l'approche que nous proposons pour implémenter l'interopérabilité des données. En effet, cette approche a pour objectif de fédérer des bases de données à travers la résolution des problèmes d'appariement, qui sont un moyen pour faciliter l'échange et l'utilisation des données par différentes entités. L'approche est implémentée dans le cadre d'une migration de données qui représente une des ambitions du projet soutenu par l'entreprise Forterro Sylob et porté par le Centre de Génie Industriel de l'IMT Mines Albi. Ce projet de collaboration a pour but d'assister les processus de migration, la montée de version logicielle ou encore l'implémentation de connecteurs. Pour ce faire, nous avons présenté les notions de base de la théorie des graphes et ses avantages en termes de modélisation à travers les graphes, les hypergraphes, leurs propriétés et leurs caractéristiques. Ensuite, nous avons introduit la notion d'optimisation pour la traduction des problèmes à travers la définition des variables de décision, des contraintes et de la fonction objectif. Puis, après avoir défini la notion de bases de données relationnelles, l'interopérabilité des données et positionné le mécanisme de migration des données, nous avons défini le problème d'appariement comme une des tâches les plus importantes dans le mécanisme de migration. Ce mécanisme consiste à identifier les concepts et les structures équivalentes dans différents schémas ou modèles de données. Dans ce contexte, les techniques issues du traitement du langage naturel sont utilisées pour faciliter cette identification en extrayant et en mettant en correspondance des schémas ou des structures en général et des schémas correspondants aux bases de données en particulier. En analysant l'état de l'art actuel sur les problèmes d'appariement, nous avons constaté qu'il est nécessaire d'avoir une approche, flexible, globale et générique, capable de prendre en compte les diverses variations et caractéristiques des problèmes d'appariement. Finalement, l'approche présentée devra donc contribuer à répondre à ce besoin.

Dans le Chapitre 3, nous avons présenté une approche pour résoudre les problèmes d'appariement qui peut traiter différentes phases (modélisation, agrégation ou sélection), techniques (mesures de similarité), variantes (par paire, holistique ou à large échelle) et contraintes (seuil, cardinalité ou structure) liées à ces problèmes. L'analyse de l'état de l'art relatif à ces problèmes montre aussi que la théorie des graphes et l'optimisation sont des techniques utilisées à plusieurs niveaux et peuvent ainsi être proposées pour combler certaines lacunes

des approches actuelles. Ainsi, nous utilisons la modélisation des schémas fournis en entrée par les hypergraphes qui permettent une modélisation plus expressive et plus représentative des relations complexes des bases de données, comme les contraintes de clés étrangères dans les bases de données relationnelles. Ensuite, nous traduisons le problème d'appariement des schémas en un problème d'appariement des hypergraphes grâce à un modèle d'optimisation où sont spécifiées les variables de décisions, les contraintes et la fonction objectif. Dans ce modèle, les variables de décision traduisent le choix des correspondances à trouver parmi l'ensemble de toutes les correspondances possibles, et qui satisferont aux contraintes des problèmes d'appariement *((1 : 1), (1 : m), (n : 1) et (n : m))* traduit sous forme d'équations ou d'inéquations. La fonction objectif qui est une expression mathématique qui sert à quantifier la similarité globale entre les schémas. Cette similarité est calculée en se basant sur les informations des schémas disponibles. Ces informations sont classées en quatre dimensions (sémantiques, syntaxiques, type de données et structurelle). Les calculs se font en utilisant des mesures de similarité issues du traitement du langage naturel et en proposant un algorithme pour le calcul de la similarité structurelle se basant sur la notion de voisinage. Des algorithmes d'agrégation, dont l'un est basé sur un modèle d'optimisation, sont proposés pour agréger les dimensions des matrices de similarité et une somme pondérée est utilisée pour renvoyer une seule valeur de similarité. Ainsi, l'approche permet de modéliser des schémas de base de données, de retrouver les relations entre eux et d'optimiser le processus de mise en relation global.

Le Chapitre 4 a été dédié à la validation l'approche proposée. Nous avons présenté tout d'abord l'environnement d'implémentation puis expérimenté l'approche sur deux cas d'étude de migration entre deux bases de données relationnelles avec l'hypothèse que les données ne sont pas disponibles dans la base de données cible et aussi en considérant des contraintes *(1 : 1)*. Le premier est un cas d'étude académique où les bases de données contiennent des informations géographiques et le second est un cas d'étude industriel issu des bases de données de deux systèmes ERP de l'entreprise Forterro Sylob. L'objectif de l'évaluation est de valider l'approche utilisée en effectuant une série de simulations intégrant une variation sur plusieurs paramètres et guidées par les vecteurs de poids traduisant l'influence de chaque dimension (sémantique, syntaxique, types de données et structurelle). Des métriques d'évaluation ont été utilisées pour quantifier et comparer les résultats obtenus. Ainsi, nous avons montré que l'approche est capable de trouver l'ensemble des correspondances avec de bonnes valeurs de précision et de rappel sans imposer d'autres contraintes de seuil ou de ratio. La robustesse de l'approche a permis de prendre en compte plusieurs dimensions des données, et grâce à la fonction objectif, l'approche est capable de faire face à l'absence ou l'imprécision des informations des bases de données.

5.2 Principales contributions

Comme précisé au début de ce chapitre, deux objectifs ont été fixés qui ont abouti à deux principales contributions. La première contribution répond à la *QR1* et fait référence au **cadre d'interopérabilité** proposé dont l'ambition n'est pas seulement de rassembler les concepts pratiques et théoriques de l'interopérabilité autour d'un unique cadre général, mais aussi, de proposer une feuille de route à suivre pour les futurs travaux relatifs à l'interopérabilité. Ceci a fait l'objet d'un article soumis pour publication (LABRECHE et al., 2023b) et en cours de relecture au 15/01/2023. La seconde contribution est la proposition d'une **approche flexible, globale et générique pour l'appariement des schémas**. Cette approche est capable d'intégrer et de rassembler plusieurs techniques et stratégies. Elle permet de faire face à l'absence d'information et de contribuer à l'automatisation de ce processus ainsi qu'à la transformation à la volée grâce aux modèles d'optimisation et à la modélisation par la théorie des graphes. Cette contribution répond à la *QR2* et fait l'objet d'une première publication (LABRECHE et al., 2020) et a été étendue dans un deuxième article soumis pour publication (LABRECHE et al., 2023a) en cours de relecture au 15/01/2023.

5.3 Améliorations et Perspectives

Au cours de la réalisation de ces travaux de recherche, de nombreux questionnements se sont posés et plusieurs perspectives ont été envisagées.

En suivant le cadre général d'interopérabilité proposé, la question d'évaluation et d'amélioration de l'interopérabilité n'a pas été traitée. La théorie des graphes peut être utilisée et des algorithmes sont proposés dans la littérature pour évaluer par exemple l'effort requis pour atteindre ou maintenir le niveau d'interopérabilité souhaité (BLANC-SERRIER, DUCQ et VALLESPIR, 2018) ou de mesurer la quantité des échanges des données (G. JIANG, CYBENKO et HENDLER, 2006).

L'approche d'appariement proposée n'est pas exclusive à l'appariement des schémas de bases de données, mais aussi à l'appariement des ontologies ou des métamodèles. Il serait donc intéressant d'étendre l'évaluation à ces types d'appariement.

Le point de départ de cette approche est la modélisation des schémas par la théorie des graphes. Le choix du bon type de graphe et des éléments à modéliser (tables, données, attributs, vues, etc.) est alors important. Ainsi, quelques difficultés peuvent survenir si l'on souhaite pousser encore plus loin la modélisation à la volée :

— La complexité : les graphes peuvent devenir très complexes, en particulier lorsque le nombre de nœuds et d'arêtes augmente. Cela peut rendre difficile la compréhension et l'analyse du graphe dans son ensemble, et peut également rendre difficile l'exécution de certaines opérations. Cependant, les avancées sur les graphes à larges échelles et complexes peuvent être utiles, à l'image des techniques de partitionnement ;
— La qualité des données : l'exactitude des données utilisées pour construire les graphes peut avoir un impact important. Des données absentes ou de mauvaise qualité peuvent conduire à des résultats incorrects ou trompeurs. C'est pour cela qu'il serait intéressant d'avoir des mécanismes de nettoyage et de préparation des données ;
— Évolutivité : certains graphes peuvent donner des imprécisions lorsqu'ils changent ou évoluent, ce qui rend difficile l'adaptation à la volée. Les techniques d'intégration de l'incertitude et les techniques de prédiction des liens peuvent être une voie prometteuse (SAÏS, 2019) ; (THOMAS et al., 2022).

Le calcul de la similarité peut être plus précis en procédant à des améliorations, par exemple :

— L'utilisation des sommes pondérées non-linéaires pour capter plus de relations de dépendance entre les différentes dimensions (A. A. A. ALGERGAWY, 2010) ;
— L'amélioration de la sémantique par des procédés d'enrichissement (SORRENTINO et al., 2009) ;
— La création de nouvelles connaissances (KOK et DOMINGOS, 2009) ;
— L'amélioration de la structure par des techniques de plongement structurel (DI-JORIO, LAURENT et TEISSEIRE, 2009) ; (VOIGT, IVANOV et RUMMLER, 2010) ;
— La découverte et l'extraction des schémas (CASTELLTORT et LAURENT, 2017) ; (CHILLÓN et al., 2021) ; (LAMMARI, COMYN-WATTIAU et AKOKA, 2007) ; (MANSURI et SARAWAGI, 2006).

Comme nous l'avons vu dans le Chapitre 1, l'interopérabilité est cruciale pour toute architecture de fédération (Federated Learning ou Federated Computing). Les avancées technologiques permettent de connecter de plus en plus d'entités hétérogènes. Les tendances actuelles en matière de recherche mettent en avant des approches qui évitent la centralisation et permettent des communications entre les entités. Ainsi, les travaux de cette thèse peuvent contribuer à la mise en place d'une fédération favorisant les échanges directs à la volée entre les différentes entités. Cela peut améliorer les performances de la fédération, augmenter la flexibilité et permettre une meilleure utilisation des ressources.

Pour conclure, l'importance de l'interopérabilité doit être soulignée, mais des points de débats méritent d'être abordés et départagés afin de soutenir l'industrie vers une interopérabilité durable et autonome. Une vision unifiée de l'interopérabilité peut être dans la fédération des efforts.

A

Revue Systématique de la littérature

Afin d'explorer les concepts d'interopérabilité dans la littérature, la méthodologie de recherche présentée dans la Figure A.1 s'inspire des travaux de (KEELE et al., 2007) ; (KITCHENHAM et al., 2009) ; (XIAO et WATSON, 2019), et est décrite comme suit : le processus de recherche décrit dans la Section A.1, les critères d'inclusion et d'exclusion détaillés dans la Section A.2 et la classification dans la Section A.3.

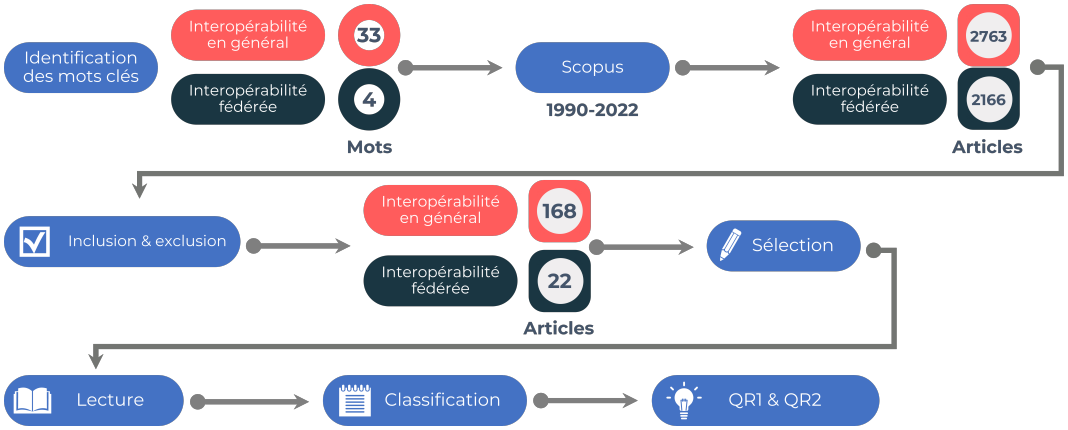


FIGURE A.1 – Méthodologie de la revue systématique de la littérature.

A.1 Processus de recherche

Le processus de recherche a été effectué sur Scopus, qui est destiné à être « *la plus grande base de données de résumés et de citations de la littérature scientifique* » (SCOPUS, 2004) et fournit des références sous forme de résumés et de citations provenant de grandes bibliothèques

numériques indexées à partir de nombreux éditeurs (Elsevier, Springer, Taylor & Francis, etc.). Toutes les étapes de la méthodologie étant guidées par les questions de recherche, deux axes principaux émergent, à savoir l'interopérabilité avec ses notions relatives (Section 1.4) et l'interopérabilité fédérée (Section 1.6). Nous nous concentrerons d'abord sur les articles qui ont examiné et passé en revue les principaux aspects de l'interopérabilité, les connaissances pratiques et les orientations de recherche. Sur cette base, nous nous concentrons ensuite sur les études qui traitent de solutions pour des aspects spécifiques de l'interopérabilité et plus précisément des approches de fédération.

Il convient de noter que certains des articles cités ne sont pas seulement le résultat des requêtes principales autour des deux axes mentionnés ci-dessus, mais sont le résultat de recherches individuelles et informelles pour avoir plus de détails. En outre, d'autres références ont été trouvées en utilisant les techniques de recherche *backward* et *forward* (Y. LEVY et ELLIS, 2006) qui sont utilisées pour trouver des références supplémentaires à partir des articles déjà identifiés. La première permet de remonter aux travaux précédents afin de suivre l'évolution des connaissances autour d'un sujet et d'identifier ses spécialistes (laboratoires ou chercheurs), la seconde également appelée recherche de référence, permet de recueillir les nouveaux progrès et développements. Avec ces techniques, nous avons essayé de couvrir la meilleure zone de recherche possible afin d'identifier d'éventuels articles qui auraient pu être manqués, sachant qu'une couverture complète est impossible et que certains articles pourraient ne pas être inclus.

De plus, Scopus nous permet de mener des recherches avancées en utilisant des requêtes construites avec les opérateurs *AND* et *OR* sur plusieurs champs (résumés, titres, mots-clés, etc.), d'appliquer les techniques de recherche susmentionnées et d'exporter les résultats de recherche pour l'analyse. D'autres filtres peuvent être appliqués pour sélectionner une plage d'années, une zone de sujet ou un type de document, mais il était important de voir l'interprétation de l'interopérabilité dans différentes zones et l'évolution des recherches. Enfin, la date de recherche a été lancée pour la première fois le *09/08/2021*, actualisée le *24/11/2021* et pour la dernière fois le *10/08/2022*, et la recherche couvre la période de 1990 à 2022.

A.1.1 Revue des notions de l'interopérabilité

Afin de recueillir autant d'articles que possible, nous nous sommes particulièrement tournés vers des études dédiées exclusivement à l'examen ou à l'analyse des concepts d'interopérabilité existants. Cependant, des mots-clés tels que *survey* ou *review* ne sont pas suffisants et ne nous permettent pas d'avoir une vision globale de la recherche, car il est difficile d'avoir une vue d'ensemble et cohérente des propositions d'interopérabilité (K. S. S. SANTOS, PINHEIRO et MACIEL, 2021). Il était donc important d'élargir le nombre de mots-clés en ajoutant des synonymes alternatifs et des termes connexes pour inclure plus de références et parvenir à un compromis entre précision et exhaustivité. Cependant, il était important pour nous d'être plus exhaustifs que précis, comme le suggère (WANDEN-BERGHE et SANZ-VALERO, 2012).

Les termes identifiés sont : *state-of-the-art*, *survey*, *review*, *taxonomies*, *challenges*, *future*, *comparative*, *models*, *methodologies*, *frameworks*, *techniques*, *initiatives*, *guidelines*, *approaches*, *platforms*, *technologies*, *tools*, *middlewares*, *methods*, *procedures*, *practices*, *protocols*, *bibliography*, *literature*, *vocabularies*, *definitions*, *issues*, *research*, *comparison*, *applications*, *systems*, *mediators*, *architectures*.

Chaque mot-clé est associé au terme *interoperability* via l'opérateur *AND*. Une recherche exhaustive sur les résumés, les titres et les mots-clés (identifiés sur Scopus avec *TITLE-ABS-KEY(mot-clé AND interoperability)*) donne plus de 40.000 résultats, ce qui est normal pour ce sujet mais énorme en termes de revue. Ainsi, afin de se concentrer sur l'étude, il a été décidé de ne se concentrer que sur les titres qui incluent les associations précédentes de mots-clés. Ainsi, la requête comprend des articles où la revue était le principal objectif et d'autres articles où la revue était une partie, nous avons alors collecté 2.763 articles.

Une autre clarification doit être apportée, la recherche sur certains mots-clés a été effectuée en utilisant *{ }* car Scopus utilise automatiquement des techniques de racinisation et de lemmatisation pour trouver les occurrences d'un mot.

La requête complète est la suivante :

TITLE(interoperability AND state AND of AND the AND art) OR TITLE(interoperability AND survey) OR TITLE(interoperability AND review) OR TITLE(interoperability AND taxonomies) OR TITLE(interoperability AND challenges) OR TITLE(interoperability AND future) OR TITLE(interoperability AND comparative) OR TITLE(interoperability AND {models}) OR TITLE(interoperability AND {methodologies}) OR TITLE(interoperability AND {frameworks}) OR TITLE(interoperability AND {techniques}) OR TITLE(interoperability AND {initiatives}) OR TITLE(interoperability AND {guidelines}) OR TITLE(interoperability AND {approaches}) OR TITLE(interoperability AND {platforms}) OR TITLE(interoperability AND {technologies}) OR TITLE(interoperability AND {tools}) OR TITLE(interoperability AND {middlewares}) OR TITLE(interoperability AND {methods}) OR TITLE(interoperability AND {procedures})OR TITLE(interoperability AND {practices}) OR TITLE(interoperability AND {protocols}) OR TITLE(interoperability AND bibliography) OR TITLE(interoperability AND literature) OR TITLE(interoperability AND vocabularies) OR TITLE(interoperability AND definitions) OR TITLE(interoperability AND {issues}) OR TITLE(interoperability AND research) OR TITLE(interoperability AND comparison) OR TITLE(interoperability AND {applications}) OR TITLE(interoperability AND {systems}) OR TITLE(interoperability AND {mediators}) OR TITLE(interoperability AND {architectures}).

A.1.2 Revue de l'interopérabilité fédérée

En poursuivant notre première recherche, nous avons décidé d'étudier les approches proposées dans la résolution des problèmes liés à l'interopérabilité. Plusieurs travaux ont été proposés ces dernières années pour résumer les documents existants et une direction émerge, à savoir l'interopérabilité fédérée. Pour collecter des références sur ce sujet, la requête consistait à combiner les mots clés *federated* et *interoperability* en utilisant l'opérateur *AND* et les termes relatifs *federation* et *interoperation*. La requête étant :

TITLE-ABS-KEY(federated AND interoperability) OR TITLE-ABS-KEY(federation AND interoperability) OR TITLE-ABS-KEY(federated AND interoperation) OR TITLE-ABS-KEY(federation AND interoperation) a donné 2.166 résultats.

A.2 Critéres d'inclusion et d'exclusion

Après avoir collecté les articles (les deux requêtes ont été exécutées de manière indépendante), l'étape suivante consiste à sélectionner les articles à analyser. Les articles non pertinents ont été exclus. Les articles sélectionnés ont été lus afin de s'assurer que leur contenu est lié à nos sujets de recherche. Pour ce faire, des critères d'inclusion et d'exclusion ont été appliqués :

- Critères d'exclusion des articles qui :
  - ne sont pas écrit en anglais
  - n'ont pas d'accès au texte intégral (même dans les pages personnelles de ResearchGate)
  - sont courts, avec moins de trois pages
  - sont des doublons ou contenant les mêmes références
  - ne se concentrent pas sur les concepts et les solutions d'interopérabilité et ne les détaillent pas

— Critères d'inclusion des articles qui :
  — sont entièrement dédiés aux concepts d'interopérabilité ou seulement une partie de l'article y est consacrée
  — sont explicitement dédiés au concept de fédération et d'interopérabilité
  — représentent la version la plus complète et la plus récente des initiatives précédentes
  — contribuent ou fournissent de nouvelles perspectives pour au moins l'une des questions de recherche

Nous avons finalement sélectionné 190 articles traitant de concepts d'interopérabilité et les fédérations.

A.3 Classification

L'interopérabilité a été largement abordée dans la littérature et dans plusieurs domaines d'application. Curieusement, une partie de la fragmentation de la littérature semble être liée à cette variété de domaines d'application. En effet, les solutions proposées dans le contexte de l'interopérabilité restent très spécifiques au domaine d'application et restent difficiles à reproduire. De plus, les solutions ne sont pas bien définies pour être adaptées à un autre problème d'un autre domaine. Les catégories choisies pour la classification qui rassemble les articles des deux requêtes sont définies dans le Tableau A.1.

Catégorie	Description
Définitions de l'interopérabilité	Identifier les articles qui répertorient ou proposent plusieurs définitions de l'interopérabilité ou qui sont basés sur une définition unique
Niveaux d'interopérabilité	Description des différents niveaux d'interopérabilité où les incompatibilités et les dépendances sont identifiées et doivent être prises en compte pour construire des solutions d'interopérabilité
Exigences en matière d'interopérabilité	Les exigences et attributs qui soutiennent l'action d'interopérabilité et sa mise en œuvre
Approches en matière d'interopérabilité	Aperçu des propositions dans le cadre de la résolution des problèmes d'interopérabilité
Évaluation et amélioration de l'interopérabilité	Collection de modèles et de méthodes pour l'évaluation, l'évaluation, la mesure et l'amélioration de la performance et du degré d'interopérabilité
Interopérabilité et fédération	Propositions qui abordent l'interopérabilité fédérée et l'interopérabilité dans les fédérations, leurs propriétés et caractéristiques
Contributions à la théorie ou à la science de l'interopérabilité	Études qui abordent la formalisation ou la théorisation de l'interopérabilité et qui fournissent des initiatives qui peuvent être considérées dans la construction d'une théorie ou d'une science de l'interopérabilité

TABLEAU A.1 – Classification proposée de la littérature

Cette classification a naturellement conduit à la construction du cadre général pour les concepts d'interopérabilité présenté dans le Chapitre 1, car elle a permis de converger enfin vers une compréhension globale de l'interopérabilité en tant qu'initiative générique consistant en étapes claires, chacune nécessitant un cadre approprié pour en faire un succès.

Table des figures

Liste des tableaux

Bibliographie

(Abedjan, Golab et Naumann, 2015) Z. Abedjan, L. Golab et F. Naumann. "Profiling relational data: a survey". In: *The VLDB Journal* 24.4 (2015), p. 557-581 (cf. p. 57).

(Abukwaik et Rombach, 2017) H. Abukwaik et D. Rombach. "Software interoperability analysis in practice: a survey". In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. 2017, p. 12-20 (cf. p. 15, 23).

(Achananuparp, X. Hu et Shen, 2008) P. Achananuparp, X. Hu et X. Shen. "The evaluation of sentence similarity measures". In: *International Conference on data warehousing and knowledge discovery*. Springer. 2008, p. 305-316 (cf. p. 77).

(Adolphs, 2015) P. Adolphs. "RAMI 4.0: An Architectural Model for Industrie 4.0". In: *Platform Industrie* 4 (2015) (cf. p. 19).

(Agarwal et al., 2016) R. Agarwal, D. G. Fernandez, T. Elsaleh, A. Gyrard, J. Lanza, L. Sanchez, N. Georgantas et V. Issarny. "Unified IoT ontology to enable interoperability and federation of testbeds". In: *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*. IEEE. 2016, p. 70-75 (cf. p. 30).

(Aggarwal, H. Wang et al., 2010) C. C. Aggarwal, H. Wang et al. *Managing and mining graph data*. T. 40. Springer, 2010 (cf. p. 63).

(Agostinho et al., 2016) C. Agostinho, Y. Ducq, G. Zacharewicz, J. Sarraipa, F. Lampathaki, R. Poler et R. Jardim-Goncalves. "Towards a sustainable interoperability in networked enterprise information systems: Trends of knowledge and model-driven technology". In: *Computers in industry* 79 (2016), p. 64-76 (cf. p. 27).

(Aipe et Gadiraju, 2018) A. Aipe et U. Gadiraju. "Similarhits: Revealing the role of task similarity in microtask crowdsourcing". In: *Proceedings of the 29th on Hypertext and Social Media*. 2018, p. 115-122 (cf. p. 76).

(M. Alam et al., 2021) M. Alam, C. F. Ahmed, M. Samiullah, C. K. Leung et al. "Mining frequent patterns from hypergraph databases". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2021, p. 3-15 (cf. p. 57).

(A. Algergawy, Nayak et Saake, 2010) A. Algergawy, R. Nayak et G. Saake. "Element similarity measures in XML schema matching". In: *Information Sciences* 180.24 (2010), p. 4975-4998 (cf. p. 52).

(A. A. A. Algergawy, 2010) A. A. A. Algergawy. "Management of XML data by means of schema matching". Thèse de doct. Magdeburg, Univ., Diss., 2010, 2010 (cf. p. 42, 114).

(Allahyari et al., 2017) M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez et K. Kochut. "Text summarization techniques: a brief survey". In: *arXiv preprint arXiv:1707.02268* (2017) (cf. p. 44).

(F. Almeida et Xexéo, 2019) F. Almeida et G. Xexéo. "Word embeddings: A survey". In: *arXiv preprint arXiv:1901.09069* (2019) (cf. p. 44).

(Almeida Prado Cestari et al., 2020) J. M. Almeida Prado Cestari, E. d. F. R. Loures, E. A. P. Santos et H. Panetto. "A capability model for public administration interoperability". In: *Enterprise Information Systems* 14.8 (2020), p. 1071-1101 (cf. p. 21, 23).

(Alter, 2008) S. Alter. "Defining information systems as work systems: implications for the IS field". In: *European Journal of Information Systems* 17.5 (2008), p. 448-469 (cf. p. 7).

(Alwan et al., 2017) A. A. Alwan, A. Nordin, M. Alzeber et A. Z. Abualkishik. "A survey of schema matching research using database schemas and instances". In: *International Journal of Advanced Computer Science and Applications* 8.10 (2017), p. 2017 (cf. p. 44).

(Angelopoulos et al., 2019) A. Angelopoulos, E. T. Michailidis, N. Nomikos, P. Trakadas, A. Hatziefremidis, S. Voliotis et T. Zahariadis. "Tackling faults in the industry 4.0 era—a survey of machine-learning solutions and key aspects". In: *Sensors* 20.1 (2019), p. 109 (cf. p. 11).

(Anthes, 2010) G. Anthes. "Happy birthday, RDBMS!" In: *Communications of the ACM* 53.5 (2010), p. 16-17 (cf. p. 42).

(Appandairajan, N. Z. A. Khan et Madiajagan, 2012) P. Appandairajan, N. Z. A. Khan et M. Madiajagan. "ERP on Cloud: Implementation strategies and challenges". In: *2012 International Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM)*. IEEE. 2012, p. 56-59 (cf. p. 9).

(Arvind et Köbler, 2006) V. Arvind et J. Köbler. "On hypergraph and graph isomorphism with bounded color classes". In: *Annual Symposium on Theoretical Aspects of Computer Science*. Springer. 2006, p. 384-395 (cf. p. 62).

(Assis et Bittencourt, 2016) M. R. Assis et L. F. Bittencourt. "A survey on cloud federation architectures: Identifying functional and non-functional properties". In: *Journal of Network and Computer Applications* 72 (2016), p. 51-71 (cf. p. 30).

(Atencia, David et Scharffe, 2012) M. Atencia, J. David et F. Scharffe. "Keys and pseudo-keys detection for web datasets cleansing and interlinking". In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer. 2012, p. 144-153 (cf. p. 56).

(ATHENA, 2004-2007) ATHENA. *List of ATHENA deliverables, INTEROP-VLab, the European Virtual Laboratory for Enterprise Interoperability (I-VLab)*. 2004-2007. URL: %7Bhttp://interop-vlab.eu/athena/%7D. (accessed: 16.09.2021) (cf. p. 18).

(Aumueller et al., 2005) D. Aumueller, H.-H. Do, S. Massmann et E. Rahm. "Schema and ontology matching with COMA++". In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. 2005, p. 906-908 (cf. p. 48).

(Azad et al., 2020) P. Azad, N. J. Navimipour, A. M. Rahmani et A. Sharifi. "The role of structured and unstructured data managing mechanisms in the Internet of things". In: *Cluster computing* 23.2 (2020), p. 1185-1198 (cf. p. 40).

(Bahmanian et Sajna, 2015) M. A. Bahmanian et M. Sajna. "Connection and separation in hypergraphs". In: *Theory and Applications of Graphs* 2.2 (2015), p. 5 (cf. p. 69).

(Al-Bakri et Fairbairn, 2012) M. Al-Bakri et D. Fairbairn. "Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources". In: *International Journal of Geographical Information Science* 26.8 (2012), p. 1437-1456 (cf. p. 76).

(Banerjee et Parui, 2022) A. Banerjee et S. Parui. "On Some General Operators of Hypergraphs". In: *arXiv preprint arXiv:2203.00396* (2022) (cf. p. 67).

(Barthe-Delanoë et al., 2014) A.-M. Barthe-Delanoë, S. Truptil, F. Bénaben et H. Pingaud. "Event-driven agility of interoperability during the Run-time of collaborative processes". In: *Decision Support Systems* 59 (2014), p. 171-179 (cf. p. 3).

(Barut, Faisst et Kanet, 2002) M. Barut, W. Faisst et J. J. Kanet. "Measuring supply chain coupling: an information system perspective". In: *European Journal of Purchasing & Supply Management* 8.3 (2002), p. 161-171 (cf. p. 22).

(Bates et Samal, 2018) D. W. Bates et L. Samal. "Interoperability: What is it, how can we make it work for clinicians, and how should we measure it in the future?" In: *Health services research* 53.5 (2018), p. 3270 (cf. p. 17).

(Batet et Sánchez, 2015) M. Batet et D. Sánchez. "A review on semantic similarity". In: *Encyclopedia of Information Science and Technology, Third Edition* (2015), p. 7575-7583 (cf. p. 46).

(Bazzanella et Tzitzikas, 2013) B. Bazzanella et Y. Tzitzikas. "Interoperability objectives and approaches: Results from the APARSEN NoE". In: *10th International Conference on Preservation of Digital Objects*. 2013, p. 53 (cf. p. 15, 18).

(Belchior et al., 2021) R. Belchior, A. Vasconcelos, S. Guerreiro et M. Correia. *A Survey on Blockchain Interoperability: Past, Present, and Future Trends*. 2021. arXiv: 2005.14282 [cs.DC] (cf. p. 18, 23, 26).

(Bellahsene et al., 2011) Z. Bellahsene, A. Bonifati, F. Duchateau et Y. Velegrakis. "On evaluating schema matching and mapping". In: *Schema matching and mapping*. Springer, 2011, p. 253-291 (cf. p. 43, 52, 89).

(Bellahsene et Duchateau, 2011) Z. Bellahsene et F. Duchateau. "Tuning for schema matching". In: *Schema Matching and Mapping*. Springer, 2011, p. 293-316 (cf. p. 52).

(Beltran, Jaudoin et Pivert, 2015) W. C. Beltran, H. Jaudoin et O. Pivert. "Découverte de proportions analogiques dans les bases de données: Une première approche". In: *15e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'15)*. 2015, p. 12 (cf. p. 41).

(Benaben, 2012) F. Benaben. "Conception de Système d'Information de Médiation pour la prise en charge de l'Interopérabilité dans les Collaborations d'Organisations". Thèse de doct. Institut National Polytechnique de Toulouse, 2012 (cf. p. 1, 6, 7).

(Benaben et al., 2013) F. Benaben, N. Boissel-Dallier, H. Pingaud et J.-P. Lorre. "Semantic issues in model-driven management of information system interoperability". In: *International Journal of Computer Integrated Manufacturing* 26.11 (2013), p. 1042-1053 (cf. p. 3).

(Benaben et al., 2012) F. Benaben, M. Lauras, S. Truptil et J. Lamothe. "Mise 3.0: an agile support for collaborative situation". In: *Working Conference on Virtual Enterprises*. Springer. 2012, p. 645-654 (cf. p. 3).

(Bénaben et al., 2008a) F. Bénaben, C. Hanachi, M. Lauras, P. Couget et V. Chapurlat. "A metamodel and its ontology to guide crisis characterization and its collaborative management". In: *Proceedings of the 5th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Washington, DC, USA, May*. 2008, p. 4-7 (cf. p. 29).

(Bénaben et al., 2008b) F. Bénaben, J. Touzi, V. Rajsiri, S. Truptil, J.-P. Lorré et H. Pingaud. "Mediation information system design in a collaborative SOA context through a MDD approach". In: *Proceedings of MDISIS* 8 (2008), p. 1-17 (cf. p. 16, 23, 25).

(Bengio, Ducharme et Vincent, 2000) Y. Bengio, R. Ducharme et P. Vincent. "A neural probabilistic language model". In: *Advances in neural information processing systems* 13 (2000) (cf. p. 74).

(Benson et Grieve, 2021) T. Benson et G. Grieve. "Why interoperability is hard". In: *Principles of health interoperability*. Springer, 2021, p. 21-40 (cf. p. 22).

(Bento et Costa, 2013) F. Bento et C. J. Costa. "ERP measure success model; a new perspective". In: *Proceedings of the 2013 International Conference on Information Systems and Design of Communication*. 2013, p. 16-26 (cf. p. 9).

(Bento, Costa et Aparicio, 2017) F. Bento, C. J. Costa et M. Aparicio. "SI success models, 25 years of evolution". In: *2017 12th Iberian conference on information systems and technologies (CISTI)*. IEEE. 2017, p. 1-6 (cf. p. 9).

(Berge, 1984) C. Berge. *Hypergraphs: combinatorics of finite sets*. T. 45. Elsevier, 1984 (cf. p. 38).

(Berlin et Motro, 2002) J. Berlin et A. Motro. "Database schema matching using machine learning with feature selection". In: *International Conference on Advanced Information Systems Engineering*. Springer. 2002, p. 452-466 (cf. p. 44).

(Bernstein, Madhavan et Rahm, 2011) P. A. Bernstein, J. Madhavan et E. Rahm. "Generic schema matching, ten years later". In: *Proceedings of the VLDB Endowment* 4.11 (2011), p. 695-701 (cf. p. 48).

(A.-J. Berre et al., 2007) A.-J. Berre, B. Elvesæter, N. Figay, C. Guglielmina, S. G. Johnsen, D. Karlsen, T. Knothe et S. Lippe. "The ATHENA interoperability framework". In: *Enterprise interoperability II*. Springer, 2007, p. 569-580 (cf. p. 18).

(Berro, Megdiche et Teste, 2015) A. Berro, I. Megdiche et O. Teste. "A linear program for holistic matching: Assessment on schema matching benchmark". In: *Database and Expert Systems Applications*. Springer. 2015, p. 383-398 (cf. p. 50, 83).

(Bézivin, 2005) J. Bézivin. "On the unification power of models". In: *Software & Systems Modeling* 4.2 (2005), p. 171-188 (cf. p. 29).

(Bianchini et al., 2006) D. Bianchini, V. De Antonellis, M. Melchiori et D. Salvi. "Semantic-enriched service discovery". In: *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE. 2006, p. 38-38 (cf. p. 22).

(Bick et al., 2021) C. Bick, E. Gross, H. A. Harrington et M. T. Schaub. "What are higher-order networks?" In: *arXiv preprint arXiv:2104.11329* (2021) (cf. p. 38).

(Bidoux, 2016) L. Bidoux. "Planification avec préférences basée sur la Théorie de l'Utilité Multi-Attribut couplée à une intégrale de Choquet: application à l'interopérabilité des organisations en gestion de crise". Thèse de doct. Ecole nationale des Mines d'Albi-Carmaux, 2016 (cf. p. 3).

(Blanc-Serrier, Ducq et Vallespir, 2018) S. Blanc-Serrier, Y. Ducq et B. Vallespir. "Organisational interoperability characterisation and evaluation using enterprise modelling and graph theory". In: *Computers in Industry* 101 (2018), p. 67-80 (cf. p. 114).

(Bodkhe et al., 2020) U. Bodkhe, S. Tanwar, K. Parekh, P. Khanpara, S. Tyagi, N. Kumar et M. Alazab. "Blockchain for industry 4.0: A comprehensive review". In: *IEEE Access* 8 (2020), p. 79764-79800 (cf. p. 11).

(Boell et Cecez-Kecmanovic, 2015) S. K. Boell et D. Cecez-Kecmanovic. "What is an information system?" In: *2015 48th Hawaii International Conference on System Sciences*. IEEE. 2015, p. 4959-4968 (cf. p. 7).

(Boersma et Kingma, 2005) K. Boersma et S. Kingma. "From means to ends: The transformation of ERP in a manufacturing company". In: *The Journal of Strategic Information Systems* 14.2 (2005), p. 197-219 (cf. p. 9).

(Boissel-Dallier et al., 2015) N. Boissel-Dallier, F. Benaben, J.-P. Lorré et H. Pingaud. "Mediation information system engineering based on hybrid service composition mechanism". In: *Journal of Systems and Software* 108 (2015), p. 39-59 (cf. p. 3).

(Bonifati et al., 2010) A. Bonifati, E. Chang, T. Ho, L. V. Lakshmanan, R. Pottinger et Y. Chung. "Schema mapping and query translation in heterogeneous P2P XML databases". In: *The VLDB Journal* 19.2 (2010), p. 231-256 (cf. p. 43).

(Boonstra, 2013) A. Boonstra. "How do top managers support strategic information system projects and why do they sometimes withhold this support?" In: *International Journal of Project Management* 31.4 (2013), p. 498-512 (cf. p. 7).

(Borst, 1999) W. N. Borst. "Construction of engineering ontologies for knowledge sharing and reuse." In: (1999) (cf. p. 29).

(Bouacida et Mohapatra, 2021) N. Bouacida et P. Mohapatra. "Vulnerabilities in Federated Learning". In: *IEEE Access* 9 (2021), p. 63229-63249 (cf. p. 30).

(Boukhari, Bellatreche et Jean, 2012) I. Boukhari, L. Bellatreche et S. Jean. "An ontological pivot model to interoperate heterogeneous user requirements". In: *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*. Springer. 2012, p. 344-358 (cf. p. 21).

(Bourey et al., 2007) J.-P. Bourey, R. Grangel, G. Doumeingts et A. Berre. "Deliverable DTG 2.3-Report On Model Driven Interoperability". In: *INTEROP-NOE Project* (2007) (cf. p. 14, 19).

(Bourey et al., 2006) J.-P. Bourey, R. G. Uji, G. Doumeingts, A. J. Berre, S. Pantelopoulos et K. Kalampoukas. "Deliverable DTG2. 3 Report on model driven". In: *Update* 2.3 (2006) (cf. p. 14).

(Bourrières, 2006) J.-P. Bourrières. "The interop network of excellence". In: *Interoperability of Enterprise Software and Applications*. Springer, 2006, p. 455-457 (cf. p. 19).

(Boussaïd, Siarry et Ahmed-Nacer, 2017) I. Boussaïd, P. Siarry et M. Ahmed-Nacer. "A survey on search-based model-driven engineering". In: *Automated Software Engineering* 24.2 (2017), p. 233-294 (cf. p. 34).

(S. Boyd, S. P. Boyd et Vandenberghe, 2004) S. Boyd, S. P. Boyd et L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004 (cf. p. 39).

(Boza et al., 2015) A. Boza, L. Cuenca, R. Poler et Z. Michaelides. "The interoperability force in the ERP field". In: *Enterprise Information Systems* 9.3 (2015), p. 257-278 (cf. p. 12).

(Breque, De Nul et Petridis, 2021) M. Breque, L. De Nul et A. Petridis. "Industry 5.0: towards a sustainable, human-centric and resilient European industry". In: *Luxembourg, LU: European Commission, Directorate-General for Research and Innovation* (2021) (cf. p. 11).

(Bretto, 2013) A. Bretto. "Hypergraph theory". In: *An introduction. Mathematical Engineering. Cham: Springer* (2013) (cf. p. 38, 58).

(Bruzzone et al., 2019) A. G. Bruzzone, G. Fancello, M. Daga, B. Leban et M. Massei. "Mixed reality for industrial applications: interactions in human-machine system and modelling in immersive virtual environment". In: *International Journal of Simulation and Process Modelling* 14.2 (2019), p. 165-177 (cf. p. 11).

(Bunke, X. Jiang et Kandel, 2000) H. Bunke, X. Jiang et A. Kandel. "On the minimum common supergraph of two graphs". In: *Computing* 65.1 (2000), p. 13-25 (cf. p. 62).

(Burns, Cosgrove et Doyle, 2019) T. Burns, J. Cosgrove et F. Doyle. "A Review of Interoperability Standards for Industry 4.0." In: *Procedia Manufacturing* 38 (2019), p. 646-653 (cf. p. 18).

(C4ISR et al., 1998) C. A. W. G. C4ISR et al. "Levels of information systems interoperability (LISI)". In: *United States of America Department of Defense* (1998) (cf. p. 18).

(M. Camara, Ducq et Dupas, 2010) M. Camara, Y. Ducq et R. Dupas. "Methodology for prior evaluation of interoperability". In: *Working Conference on Virtual Enterprises*. Springer. 2010, p. 697-704 (cf. p. 22).

(M. S. Camara, Ducq et Dupas, 2014) M. S. Camara, Y. Ducq et R. Dupas. "A methodology for the evaluation of interoperability improvements in inter-enterprises collaboration based on causal performance measurement models". In: *International Journal of Computer Integrated Manufacturing* 27.2 (2014), p. 103-119 (cf. p. 23).

(Cappuzzo, Papotti et Thirumuruganathan, 2020) R. Cappuzzo, P. Papotti et S. Thirumuruganathan. "Creating embeddings of heterogeneous relational datasets for data integration tasks". In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2020, p. 1335-1349 (cf. p. 51).

(Cárdenas-Robledo et al., 2022) L. A. Cárdenas-Robledo, Ó. Hernández-Uribe, C. Reta et J. A. Cantoral-Ceballos. "Extended reality applications in industry 4.0.-A systematic literature review". In: *Telematics and Informatics* (2022), p. 101863 (cf. p. 12).

(Carney, J. Smith et Place, 2005) D. Carney, J. Smith et P. Place. *Topics in interoperability: Infrastructure replacement in a system of systems*. Rapp. tech. CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 2005 (cf. p. 13).

(Casanova et al., 2007) M. A. Casanova, K. K. Breitman, D. F. Brauner et A. L. Marins. "Database conceptual schema matching". In: *Computer* 40.10 (2007), p. 102-104 (cf. p. 43).

(Castelltort et Laurent, 2017) A. Castelltort et A. Laurent. "Exploiting nosql graph databases and in memory architectures for extracting graph structural data summaries". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 25.01 (2017), p. 81-109 (cf. p. 57, 114).

(Castelnovo et Simonetta, 2006) W. Castelnovo et M. Simonetta. "Networks of SLGOs: from systems interoperability to organizational cooperability". In: *Proceedings of the 6th European Conference on E-government*. 2006, p. 47-56 (cf. p. 18).

(Castrillo, León et Gómez, 2018) E. Castrillo, E. León et J. Gómez. "Dynamic structural similarity on graphs". In: *arXiv preprint arXiv:1805.01419* (2018) (cf. p. 67).

(Cazabet, Amblard et Hanachi, 2010) R. Cazabet, F. Amblard et C. Hanachi. "Detection of overlapping communities in dynamical social networks". In: *2010 IEEE second international conference on social computing*. IEEE. 2010, p. 309-314 (cf. p. 35).

(Chalmeta et Grangel, 2008) R. Chalmeta et R. Grangel. "Methodology for the implementation of knowledge management systems". In: *Journal of the American Society for Information Science and technology* 59.5 (2008), p. 742-755 (cf. p. 9).

(Chalmeta et Pazos, 2015) R. Chalmeta et V. Pazos. "A step-by-step methodology for enterprise interoperability projects". In: *Enterprise Information Systems* 9.4 (2015), p. 436-464 (cf. p. 18).

(Chapurlat et Roque, 2009) V. Chapurlat et M. Roque. "Interoperability constraints and requirements formal modelling and checking framework". In: *IFIP International Conference on Advances in Production Management Systems*. Springer. 2009, p. 219-226 (cf. p. 21).

(Charalabidis et al., 2008) Y. Charalabidis, G. Gionis, K. Moritz Hermann et C. Martinez. "Enterprise interoperability research roadmap, draft version 5.0". In: *Brussels: European Commission* (2008) (cf. p. 2, 13, 27).

(Charalabidis, 2014) Y. Charalabidis. *Revolutionizing enterprise interoperability through scientific foundations*. IGI Global, 2014 (cf. p. 16).

(Charalabidis, R. J. Gonçalves et Popplewell, 2011) Y. Charalabidis, R. J. Gonçalves et K. Popplewell. "Towards a scientific foundation for interoperability". In: *Interoperability in digital public services and administration: Bridging E-government and E-business*. IGI Global, 2011, p. 355-373 (cf. p. 14).

(Charatsis et al., 2005) K. Charatsis, A. Kalogeras, M. Georgoudakis, J. Gialelis et G. Papadopoulos. "Home/building automation environment architecture enabling interoperability, flexibility and reusability". In: *Proceedings of the IEEE International Symposium on Industrial Electronics, 2005. ISIE 2005.* T. 4. IEEE. 2005, p. 1441-1446 (cf. p. 17).

(Checkland, 2000) P. Checkland. "Systems thinking, systems practice: includes a 30-year retrospective". In: *Journal-Operational Research Society* 51.5 (2000), p. 647-647 (cf. p. 6).

(D. Chen, 2006) D. Chen. "Enterprise Interoperability Framework." In: *EMOI-INTEROP*. 2006 (cf. p. 19).

(D. Chen, 2017) D. Chen. "Framework for enterprise interoperability". In: *Enterprise interoperability: INTEROP-PGSO vision* 1 (2017), p. 1-18 (cf. p. 19, 27).

(D. Chen et Daclin, 2006) D. Chen et N. Daclin. "Framework for enterprise interoperability". In: *Interoperability for Enterprise Software and Applications: Proceedings of the Workshops and the Doctorial Symposium of the Second IFAC/IFIP I-ESA International Conference: EI2N, WSI, IS-TSPQ 2006*. Wiley Online Library. 2006, p. 77-88 (cf. p. 13).

(D. Chen et Doumeingts, 2003a) D. Chen et G. Doumeingts. "Basic concepts and approaches to develop interoperability of enterprise applications". In: *Working Conference on Virtual Enterprises*. Springer. 2003, p. 323-330 (cf. p. 12).

(D. Chen et Doumeingts, 2003b) D. Chen et G. Doumeingts. "European initiatives to develop interoperability of enterprise applications—basic concepts, framework and roadmap". In: *Annual reviews in control* 27.2 (2003), p. 153-162 (cf. p. 13).

(D. Chen, Doumeingts et F. Vernadat, 2008) D. Chen, G. Doumeingts et F. Vernadat. "Architectures for enterprise integration and interoperability: Past, present and future". In: *Computers in industry* 59.7 (2008), p. 647-659 (cf. p. 2, 13, 16-20, 33).

(D. Chen, Knothe et Zelm, 2004) D. Chen, T. Knothe et M. Zelm. "ATHENA integrated project and the mapping to international standard ISO 15704". In: *Knowledge Sharing in the Integrated Enterprise*. Springer, 2004, p. 67-77 (cf. p. 18).

(D. Chen, Vallespir et Daclin, 2008) D. Chen, B. Vallespir et N. Daclin. "An Approach for Enterprise Interoperability Measurement." In: *MoDISE-EUS* 341 (2008), p. 1-12 (cf. p. 22).

(J. Chen, R. Zhao et Z. Li, 2004) J. Chen, R. Zhao et Z. Li. "Voronoi-based k-order neighbour relations for spatial analysis". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 59.1-2 (2004), p. 60-72 (cf. p. 70).

(Chillón et al., 2021) A. H. Chillón, J. R. Hoyos, J. Garcia-Molina et D. S. Ruiz. "Discovering entity inheritance relationships in document stores". In: *Knowledge-Based Systems* 230 (2021), p. 107394 (cf. p. 114).

(Chiticariu et al., 2007) L. Chiticariu, M. A. Hernández, P. G. Kolaitis et L. Popa. "Semi-Automatic Schema Integration in Clio." In: *VLDB*. T. 7. 2007, p. 1326-1329 (cf. p. 44).

(Chituc, 2019) C.-M. Chituc. "Interoperability frameworks for networked information systems: A comparative analysis and discussion". In: *International Journal of Cooperative Information Systems* 28.01 (2019), p. 1950002 (cf. p. 19).

(Christen et Vatsalan, 2013) P. Christen et D. Vatsalan. "Flexible and extensible generation and corruption of personal data". In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013, p. 1165-1168 (cf. p. 45).

(Christophides et al., 2019) V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis et K. Stefanidis. "End-to-end entity resolution for big data: A survey". In: *arXiv preprint arXiv:1905.06397* (2019) (cf. p. 42).

(Cleveland, Small et Brunetto, 2008) F. Cleveland, F. Small et T. Brunetto. "Smart grid: Interoperability and standards an introductory review". In: *Utility Standard Board* (2008) (cf. p. 16).

(Coche et al., 2020) J. Coche, A. Montarnal, A. Tapia et F. Benaben. "Automatic Information Retrieval from Tweets: A Semantic Clustering Approach". In: *ISCRAM 2020-17th International conference on Information Systems for Crisis Response and Management*. 2020, p-134 (cf. p. 43).

(Codd, 2007) E. F. Codd. "Relational database: A practical foundation for productivity". In: *ACM Turing award lectures*. 2007, p. 1981 (cf. p. 56).

(Coffman et Weaver, 2010) J. Coffman et A. C. Weaver. "A framework for evaluating database keyword search strategies". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, p. 729-738 (cf. p. 91).

(Conte et al., 2004) D. Conte, P. Foggia, C. Sansone et M. Vento. "Thirty years of graph matching in pattern recognition". In: *International journal of pattern recognition and artificial intelligence* 18.03 (2004), p. 265-298 (cf. p. 63).

(Cordella et al., 2004) L. P. Cordella, P. Foggia, C. Sansone et M. Vento. "A (sub) graph isomorphism algorithm for matching large graphs". In: *IEEE transactions on pattern analysis and machine intelligence* 26.10 (2004), p. 1367-1372 (cf. p. 62).

(Cram, Brohman et R. B. Gallupe, 2016) W. A. Cram, K. Brohman et R. B. Gallupe. "Information systems control: A review and framework for emerging information systems processes". In: *Journal of the Association for Information Systems* 17.4 (2016), p. 2 (cf. p. 7).

(Crescenzi et al., 2021) V. Crescenzi, A. De Angelis, D. Firmani, M. Mazzei, P. Merialdo, F. Piai et D. Srivastava. "Alaska: A flexible benchmark for data integration tasks". In: *arXiv preprint arXiv:2101.11259* (2021) (cf. p. 52).

(Cristani et Cuel, 2005) M. Cristani et R. Cuel. "A survey on ontology creation methodologies". In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 1.2 (2005), p. 49-69 (cf. p. 52).

(Croteau et P. Li, 2003) A.-M. Croteau et P. Li. "Critical success factors of CRM technological initiatives". In: *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration* 20.1 (2003), p. 21-34 (cf. p. 9).

(Cruz, Antonelli et Stroe, 2009) I. F. Cruz, F. P. Antonelli et C. Stroe. "AgreementMaker: efficient matching for large real-world schemas and ontologies". In: *Proceedings of the VLDB Endowment* 2.2 (2009), p. 1586-1589 (cf. p. 50).

(Culot et al., 2020) G. Culot, G. Nassimbeni, G. Orzes et M. Sartor. "Behind the definition of Industry 4.0: Analysis and open questions". In: *International Journal of Production Economics* 226 (2020), p. 107617 (cf. p. 11, 17).

(D'Ambrogio, Gianni et Iazeolla, 2007) A. D'Ambrogio, D. Gianni et G. Iazeolla. "Software technologies for the interoperability, reusability and adaptability of distributed simulators". In: *Proceedings of the 2007 European Simulation Interoperability Workshop (EuroSIW-07)*. 2007 (cf. p. 17).

(Daclin, D. Chen et Vallespir, 2006) N. Daclin, D. Chen et B. Vallespir. "Enterprise interoperability measurement-Basic concepts." In: *EMOI-INTEROP* 6 (2006) (cf. p. 22).

(Daclin, D. Chen et Vallespir, 2016) N. Daclin, D. Chen et B. Vallespir. "Developing enterprise collaboration: a methodology to implement and improve interoperability". In: *Enterprise Information Systems* 10.5 (2016), p. 467-504 (cf. p. 23).

(Daclin et al., 2016) N. Daclin, S. M. Daclin, V. Chapurlat et B. Vallespir. "Writing and verifying interoperability requirements: Application to collaborative processes". In: *Computers in Industry* 82 (2016), p. 1-18 (cf. p. 21, 22).

(Daclin et Mallek, 2014) N. Daclin et S. Mallek. "Capturing and structuring interoperability requirements: a framework for interoperability requirements". In: *Enterprise Interoperability VI*. Springer, 2014, p. 239-249 (cf. p. 21).

(Damiani et al., 2018) L. Damiani, M. Demartini, G. Guizzi, R. Revetria et F. Tonelli. "Augmented and virtual reality applications in industrial systems: A qualitative review towards the industry 4.0 era". In: *IFAC-PapersOnLine* 51.11 (2018), p. 624-630 (cf. p. 11).

(Dassisti et al., 2013) M. Dassisti, R. Jardim-Goncalves, A. Molina, O. Noran, H. Panetto et M. M. Zdravković. "Sustainability and interoperability: Two facets of the same gold medal". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2013, p. 250-261 (cf. p. 25).

(Davis, 2000) G. B. Davis. "Information systems conceptual foundations: looking backward and forward". In: *Organizational and social perspectives on information technology*. Springer, 2000, p. 61-82 (cf. p. 7).

(Deerwester et al., 1990) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer et R. Harshman. "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6 (1990), p. 391-407 (cf. p. 74).

(Deshmukh et al., 2021) R. A. Deshmukh, D. Jayakody, A. Schneider et V. Damjanovic-Behrendt. "Data spine: A federated interoperability enabler for heterogeneous iot platform ecosystems". In: *Sensors* 21.12 (2021), p. 4010 (cf. p. 18, 26, 27, 30, 31).

(Devlin et al., 2018) J. Devlin, M.-W. Chang, K. Lee et K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018) (cf. p. 74).

(Dewar et al., 2017) M. Dewar, J. Healy, X. Pérez-Giménez, P. Prałat, J. Proos, B. Reiniger et K. Ternovsky. "Subhypergraphs in non-uniform random hypergraphs". In: *arXiv preprint arXiv:1703.07686* (2017) (cf. p. 69).

(Di Martino et al., 2018) B. Di Martino, M. Rak, M. Ficco, A. Esposito, S. A. Maisto et S. Nacchia. "Internet of things reference architectures, security and interoperability: A survey". In: *Internet of Things* 1 (2018), p. 99-112 (cf. p. 26).

(Diallo, 2010) S. Y. Diallo. *Towards a formal theory of interoperability*. Old Dominion University, 2010 (cf. p. 14).

(Diallo et al., 2011) S. Y. Diallo, H. Herencia-Zapana, J. J. Padilla et A. Tolk. "Understanding interoperability". In: *Proceedings of the 2011 Emerging M&S Applications in Industry and Academia Symposium*. 2011, p. 84-91 (cf. p. 14).

(Dice, 1945) L. R. Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), p. 297-302 (cf. p. 72).

(Dictionaries, 2022) O. L. Dictionaries. *System*. In: *oxfordlearnersdictionaries.com dictionary* (cf. p. 6).

(Ding, H. Dong et G. Wang, 2012) G. Ding, H. Dong et G. Wang. "Appearance-order-based schema matching". In: *International Conference on Database Systems for Advanced Applications*. Springer. 2012, p. 79-94 (cf. p. 52).

(Do, 2006) H.-H. Do. "Schema matching and mapping-based data integration". In: (2006) (cf. p. 48).

(Do, Melnik et Rahm, 2002) H.-H. Do, S. Melnik et E. Rahm. "Comparison of schema matching evaluations". In: *Net. ObjectDays: International Conference on Object-Oriented and Internet-Based Technologies, Concepts, and Applications for a Networked World*. Springer. 2002, p. 221-237 (cf. p. 52, 89).

(Do et Rahm, 2002) H.-H. Do et E. Rahm. "COMA—a system for flexible combination of schema matching approaches". In: *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier. 2002, p. 610-621 (cf. p. 48).

(Doan, Domingos et A. Y. Halevy, 2001) A. Doan, P. Domingos et A. Y. Halevy. "Reconciling schemas of disparate data sources: A machine-learning approach". In: *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*. 2001, p. 509-520 (cf. p. 47).

(Doan, A. Halevy et Ives, 2012) A. Doan, A. Halevy et Z. Ives. *Principles of data integration*. Elsevier, 2012 (cf. p. 42).

(Dongo et al., 2017) I. Dongo, F. Al Khalil, R. Chbeir et Y. Cardinale. "Semantic web datatype similarity: towards better RDF document matching". In: *International Conference on Database and Expert Systems Applications*. Springer. 2017, p. 189-205 (cf. p. 76).

(Dorneles, R. Gonçalves et Santos Mello, 2011) C. F. Dorneles, R. Gonçalves et R. dos Santos Mello. "Approximate data instance matching: a survey". In: *Knowledge and Information Systems* 27.1 (2011), p. 1-21 (cf. p. 44).

(Doumeingts, Ducq et D. Chen, 2009) G. Doumeingts, Y. Ducq et D. Chen. "System theory to support enterprise interoperability science base". In: *2009 IEEE International Technology Management Conference (ICE)*. IEEE. 2009, p. 1-12 (cf. p. 14).

(Doyle-Kent, 2021) M. Doyle-Kent. "Collaborative robotics in industry 5.0". Thèse de doct. Wien, 2021 (cf. p. 11).

(Drouot, Golra et Champeau, 2019) B. Drouot, F. R. Golra et J. Champeau. "A role modeling based approach for cyber threat analysis". In: *International Conference on Model-Driven Engineering and Software Development*. Springer. 2019, p. 76-100 (cf. p. 29).

(Drumm, 2008) C. Drumm. "Improving schema mapping by exploiting domain knowledge". Thèse de doct. Karlsruhe, Univ., Diss., 2008, 2008 (cf. p. 43).

(Drumm et al., 2007) C. Drumm, M. Schmitt, H.-H. Do et E. Rahm. "Quickmig: automatic schema matching for data migration projects". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 2007, p. 107-116 (cf. p. 42, 52).

(Duan et Da Xu, 2021) L. Duan et L. Da Xu. "Data analytics in industry 4.0: A survey". In: *Information Systems Frontiers* (2021), p. 1-17 (cf. p. 11).

(Duchateau, Bellahsene et Hunt, 2007) F. Duchateau, Z. Bellahsene et E. Hunt. "XBenchMatch: a benchmark for XML schema matching tools". In: *The VLDB Journal*. T. 1. Springer Verlag. 2007, p. 1318-1321 (cf. p. 52).

(Ducq et D. Chen, 2008) Y. Ducq et D. Chen. "How to measure interoperability: Concept and Approach". In: *2008 IEEE International Technology Management Conference (ICE)*. IEEE. 2008, p. 1-8 (cf. p. 22).

(Ducq, D. Chen et Doumeingts, 2012) Y. Ducq, D. Chen et G. Doumeingts. "A contribution of system theory to sustainable enterprise interoperability science base". In: *Computers in Industry* 63.8 (2012), p. 844-857 (cf. p. 14, 27).

(Durand et al., 2020) G. C. Durand, A. Daur, V. Kumar, S. Suman, A. M. Aftab, S. Karim, P. Diwesh, C. Hegde, D. Setlur, S. M. Ismail et al. "Spread the Good Around! Information Propagation in Schema Matching and Entity Resolution for Heterogeneous Data." In: *DI2KG VLDB*. 2020 (cf. p. 50, 71).

(Dürr et Radermacher, 2013) M. Dürr et K. Radermacher. *Einsatz von Datenbanksystemen: ein Leitfaden für die Praxis*. Springer-Verlag, 2013 (cf. p. 91).

(S. P. Dwivedi, 2020) S. P. Dwivedi. "Some algorithms on exact, approximate and error-tolerant graph matching". In: *arXiv preprint arXiv:2012.15279* (2020) (cf. p. 63).

(Y. Dwivedi et al., 2009) Y. Dwivedi, N. Mustafee, M. D. Williams et B. Lal. "Classification of information systems research revisited: A keyword analysis approach". In: (2009) (cf. p. 7).

(Y. K. Dwivedi et al., 2015) Y. K. Dwivedi, D. Wastell, S. Laumer, H. Z. Henriksen, M. D. Myers, D. Bunker, A. Elbanna, M. Ravishankar et S. C. Srivastava. "Research on information systems failures and successes: Status update and future directions". In: *Information Systems Frontiers* 17.1 (2015), p. 143-157 (cf. p. 10).

(Eichelberg et al., 2005) M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac et G. B. Laleci. "A survey and analysis of electronic healthcare record standards". In: *Acm Computing Surveys (Csur)* 37.4 (2005), p. 277-315 (cf. p. 18).

(EIF-PEGS, 2004) EIF-PEGS. *European interoperability framework for pan-european egovernment services*. 2004. URL: https://elfarchive1718.foi.hr/pluginfile.php/55018/mod_resource/content/1/EIF.pdf. (accessed: 27.08.2021) (cf. p. 17, 19).

(Elmhadhbi et al., 2020) L. Elmhadhbi, M.-H. Karray, B. Archimède, J. N. Otte et B. Smith. "A semantics-based common operational command system for multiagency disaster response". In: *IEEE Transactions on Engineering Management* (2020) (cf. p. 18).

(Elshani, Wortmann et Staab, 2020) D. Elshani, T. Wortmann et S. Staab. "Towards Better Co-Design with Disciplinary Ontologies: Review and Evaluation of Data Interoperability in the AEC Industry". In: (2020) (cf. p. 26, 29).

(Elshwimy et al., 2014) F. A. Elshwimy, A. Algergawy, A. Sarhan et E. A. Sallam. "Aggregation of similarity measures in schema matching based on generalized mean". In: *2014 IEEE 30th International Conference on Data Engineering Workshops*. IEEE. 2014, p. 74-79 (cf. p. 46).

(EnviroTREC, 2020) EnviroTREC. *Interoperability and Industry 4.0*. 2020. URL: https://www.envirotrec.ca/2020/interoperability-and-industry-4-0/. (accessed: 31.08.2021) (cf. p. 17).

(Erol, Sauser et Mansouri, 2010) O. Erol, B. J. Sauser et M. Mansouri. "A framework for investigation into extended enterprise resilience". In: *Enterprise Information Systems* 4.2 (2010), p. 111-136 (cf. p. 10).

(Euzenat, Shvaiko et al., 2007) J. Euzenat, P. Shvaiko et al. *Ontology matching*. T. 18. Springer, 2007 (cf. p. 48).

(Evermann, 2008) J. Evermann. "Theories of meaning in schema matching: A review". In: *Journal of Database Management (JDM)* 19.3 (2008), p. 55-82 (cf. p. 48).

(Fan et al., 2009) W. Fan, X. Jia, J. Li et S. Ma. "Reasoning about record matching rules". In: *Proceedings of the VLDB Endowment* 2.1 (2009), p. 407-418 (cf. p. 42).

(Fan, Y. Wu et J. Xu, 2016) W. Fan, Y. Wu et J. Xu. "Functional dependencies for graphs". In: *Proceedings of the 2016 International Conference on Management of Data*. 2016, p. 1843-1857 (cf. p. 47).

(Faria et al., 2013) D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz et F. M. Couto. "The agreementmakerlight ontology matching system". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2013, p. 527-541 (cf. p. 50).

(Faria et al., 2019) D. Faria, C. Pesquita, T. Tervo, F. M. Couto et I. F. Cruz. "AML and AMLC results for OAEI 2019". In: *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC)*. T. 2536. 2019 (cf. p. 50).

(Farias, Roxin et Nicolle, 2015) T. M. Farias, A. Roxin et C. Nicolle. "FOWLA, a federated architecture for ontologies". In: *International Symposium on Rules and Rule Markup Languages for the Semantic Web*. Springer. 2015, p. 97-111 (cf. p. 29).

(Fernandes et al., 2020) J. Fernandes, F. Ferreira, F. Cordeiro, V. G. Neto et R. Santos. "How can interoperability approaches impact on Systems-of-Information Systems characteristics?" In: *XVI Brazilian Symposium on Information Systems*. 2020, p. 1-8 (cf. p. 2, 21, 23, 26, 27).

(J. Ferreira et al., 2011) J. Ferreira, C. Agostinho, J. Sarraipa et R. Jardim-Goncalves. "Monitoring morphisms to support sustainable interoperability of enterprise systems". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2011, p. 71-82 (cf. p. 19).

(R. Ferreira et al., 2016) R. Ferreira, R. D. Lins, S. J. Simske, F. Freitas et M. Riss. "Assessing sentence similarity through lexical, syntactic and semantic analysis". In: *Computer Speech & Language* 39 (2016), p. 1-28 (cf. p. 77, 79).

(Foerster et al., 2010) T. Foerster, L. Lehto, T. Sarjakoski, L. T. Sarjakoski et J. Stoter. "Map generalization and schema transformation of geospatial data combined in a Web Service context". In: *Computers, Environment and Urban Systems* 34.1 (2010), p. 79-88 (cf. p. 44).

(Foggia, Percannella et Vento, 2014) P. Foggia, G. Percannella et M. Vento. "Graph matching and learning in pattern recognition in the last 10 years". In: *International Journal of Pattern Recognition and Artificial Intelligence* 28.01 (2014), p. 1450001 (cf. p. 63).

(Folino et al., 2012) F. Folino, D. Talia, D. Saccà et G. Manco. "Entity resolution: effective schema and data reconciliation." Thèse de doct. 2012 (cf. p. 44).

(Folmer et Krukkert, 2015) E. Folmer et D. Krukkert. "Linked data for transaction based enterprise interoperability". In: *International IFIP Working Conference on Enterprise Interoperability*. Springer. 2015, p. 113-125 (cf. p. 15).

(Ford et al., 2007) T. C. Ford, J. M. Colombi, S. R. Graham et D. R. Jacques. "Survey on interoperability measurement". In: (2007) (cf. p. 16, 22, 23).

(Fortineau, Paviot et Lamouri, 2013) V. Fortineau, T. Paviot et S. Lamouri. "Improving the interoperability of industrial information systems with description logic-based models—the state of the art". In: *Computers in Industry* 64.4 (2013), p. 363-375 (cf. p. 23).

(Fraga-Lamas, Lopes et Fernández-Caramés, 2021) P. Fraga-Lamas, S. I. Lopes et T. M. Fernández-Caramés. "Green IoT and edge AI as key technological enablers for a sustainable digital transition towards a smart circular economy: An industry 5.0 use case". In: *Sensors* 21.17 (2021), p. 5745 (cf. p. 12).

(Gal, 2005) A. Gal. "On the cardinality of schema matching". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2005, p. 947-956 (cf. p. 47).

(Gal, 2006) A. Gal. "Why is schema matching tough and what can we do about it?" In: *ACM Sigmod Record* 35.4 (2006), p. 2-5 (cf. p. 45).

(Gal et al., 2005) A. Gal, G. Modica, H. Jamil et A. Eyal. "Automatic ontology matching using application semantics". In: *AI magazine* 26.1 (2005), p. 21-21 (cf. p. 52).

(Gal et Sagi, 2010) A. Gal et T. Sagi. "Tuning the ensemble selection process of schema matchers". In: *Information Systems* 35.8 (2010), p. 845-859 (cf. p. 45-47).

(Gal et al., 2012) A. Gal, T. Sagi, M. Weidlich, E. Levy, V. Shafran, Z. Miklós et N. Q. V. Hung. "Making sense of top-k matchings: A unified match graph for schema matching". In: *Proceedings of the Ninth International Workshop on Information Integration on the Web*. 2012, p. 1-6 (cf. p. 46).

(Gali et al., 2019) N. Gali, R. Mariescu-Istodor, D. Hostettler et P. Fränti. "Framework for syntactic string similarity measures". In: *Expert Systems with Applications* 129 (2019), p. 169-185 (cf. p. 47, 65).

(Gallagher, 2006) B. Gallagher. "The state of the art in graph-based pattern matching". In: (2006) (cf. p. 63).

(B. Gallupe, 2001) B. Gallupe. "Knowledge management systems: surveying the landscape". In: *International Journal of Management Reviews* 3.1 (2001), p. 61-77 (cf. p. 10).

(Gao et al., 2010) X. Gao, B. Xiao, D. Tao et X. Li. "A survey of graph edit distance". In: *Pattern Analysis and applications* 13.1 (2010), p. 113-129 (cf. p. 63).

(Garey, 1979) M. R. Garey. "A Guide to the Theory of NP-Completeness". In: *Computers and intractability* (1979) (cf. p. 62).

(Garlapati et Biswas, 2012) R. Garlapati et R. Biswas. *Interoperability in Healthcare: A focus on the Social Interoperability*. 2012 (cf. p. 15, 16).

(Gaynor et al., 2014) M. Gaynor, F. Yu, C. H. Andrus, S. Bradner et J. Rawn. "A general framework for interoperability with applications to healthcare". In: *Health Policy and Technology* 3.1 (2014), p. 3-12 (cf. p. 16).

(Geraci, 1991) A. Geraci. *IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries*. IEEE Press, 1991 (cf. p. 2).

(Getoor et Machanavajjhala, 2013) L. Getoor et A. Machanavajjhala. "Entity resolution for big data". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, p. 1527-1527 (cf. p. 42).

(Ghaleb et al., 2020) F. F. Ghaleb, A. A. Taha, M. Hazman, M. M. Abd Ellatif et M. Abbass. "On Quasi Cycles in Hypergraph Databases". In: *IEEE Access* 8 (2020), p. 147560-147568 (cf. p. 58).

(Ghanshyambhai et Ghodasara, 2020) A. Ghanshyambhai et G. V. Ghodasara. "Innovative Aspects Of Research Advancement In Graph Theory". Thèse de doct. Doctoral Thesis, 2020 (cf. p. 63).

(Goel et P. Gupta, 2020) R. Goel et P. Gupta. "Robotics and industry 4.0". In: *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*. Springer, 2020, p. 157-169 (cf. p. 11).

(Gomaa, Fahmy et al., 2013) W. H. Gomaa, A. A. Fahmy et al. "A survey of text similarity approaches". In: *international journal of Computer Applications* 68.13 (2013), p. 13-18 (cf. p. 43, 65).

(Gomes et al., 2020) M. G. Gomes, V. H. C. da Silva, L. F. R. Pinto, P. Centoamore, S. Digiesi, F. Facchini et G. C. d. O. Neto. "Economic, environmental and social gains of the implementation of artificial intelligence at dam operations toward Industry 4.0 principles". In: *Sustainability* 12.9 (2020), p. 3604 (cf. p. 11).

(Gorecki et al., 2020) S. Gorecki, J. Possik, G. Zacharewicz, Y. Ducq et N. Perry. "A multicomponent distributed framework for smart production system modeling and simulation". In: *Sustainability* 12.17 (2020), p. 6969 (cf. p. 28).

(Gridwise Architecture Council, 2010) Gridwise Architecture Council. *Introduction to interoperability and decision-maker's interoperability checklist version 1.5*. 2010. URL: https://www.gridwiseac.org/pdfs/gwac_decisionmakerchecklist_v1_5.pdf. (accessed: 26.08.2021) (cf. p. 16).

(Grilo et Jardim-Goncalves, 2010) A. Grilo et R. Jardim-Goncalves. "Value proposition on interoperability of BIM and collaborative working environments". In: *Automation in construction* 19.5 (2010), p. 522-530 (cf. p. 12).

(Grozev et Buyya, 2014) N. Grozev et R. Buyya. "Inter-Cloud architectures and application brokering: taxonomy and survey". In: *Software: Practice and Experience* 44.3 (2014), p. 369-390 (cf. p. 30).

(Gruber, 1993) T. R. Gruber. "A translation approach to portable ontology specifications". In: *Knowledge acquisition* 5.2 (1993), p. 199-220 (cf. p. 29).

(Guédria, D. Chen et Naudet, 2009) W. Guédria, D. Chen et Y. Naudet. "A maturity model for enterprise interoperability". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2009, p. 216-225 (cf. p. 22).

(Guédria et Naudet, 2014) W. Guédria et Y. Naudet. "Extending the ontology of enterprise interoperability (ooei) using enterprise-as-system concepts". In: *Enterprise Interoperability VI*. Springer, 2014, p. 393-403 (cf. p. 13).

(Guédria, Naudet et D. Chen, 2008) W. Guédria, Y. Naudet et D. Chen. "Interoperability maturity models–survey and comparison–". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2008, p. 273-282 (cf. p. 23).

(Guédria, Naudet et D. Chen, 2015) W. Guédria, Y. Naudet et D. Chen. "Maturity model for enterprise interoperability". In: *Enterprise Information Systems* 9.1 (2015), p. 1-28 (cf. p. 22).

(Guglielmina et A. Berre, 2005) C. Guglielmina et A. Berre. "Athena,"project a4"(slide presentation)". In: *ATHENA Intermediate Audit, Athens* (2005) (cf. p. 18).

(Guijarro, 2007) L. Guijarro. "Interoperability frameworks and enterprise architectures in e-government initiatives in Europe and the United States". In: *Government Information Quarterly* 24.1 (2007), p. 89-101 (cf. p. 17).

(Guijarro, 2009) L. Guijarro. "Semantic interoperability in eGovernment initiatives". In: *Computer Standards & Interfaces* 31.1 (2009), p. 174-180 (cf. p. 17).

(Gulić, Vrdoljak et Ptiček, 2018) M. Gulić, B. Vrdoljak et M. Ptiček. "Automatically specifying a parallel composition of matchers in ontology matching process by using genetic algorithm". In: *Information* 9.6 (2018), p. 138 (cf. p. 52).

(C. Guo et al., 2013) C. Guo, C. Hedeler, N. W. Paton et A. A. Fernandes. "Matchbench: Benchmarking schema matching algorithms for schematic correspondences". In: *British National Conference on Databases*. Springer. 2013, p. 92-106 (cf. p. 52).

(H. Guo, Y. Liu et Nault, 2021) H. Guo, Y. Liu et B. R. Nault. "Provisioning Interoperable Disaster Management Systems: Integrated, Unified, and Federated Approaches". In: *Management Information Systems Quarterly* 45.1 (2021), p. 4 (cf. p. 26).

(S. Gupta et al., 2007) S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey et W. S. Noble. "Quantifying similarity between motifs". In: *Genome biology* 8.2 (2007), p. 1-9 (cf. p. 81).

(Gürdür et Asplund, 2018) D. Gürdür et F. Asplund. "A systematic review to merge discourses: Interoperability, integration and cyber-physical systems". In: *Journal of Industrial information integration* 9 (2018), p. 14-23 (cf. p. 23, 26).

(Gurobi Optimization, LLC, 2022) Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual*. 2022 (cf. p. 87).

(Hadj Taieb, Zesch et Ben Aouicha, 2020) M. A. Hadj Taieb, T. Zesch et M. Ben Aouicha. "A survey of semantic relatedness evaluation datasets and procedures". In: *Artificial Intelligence Review* 53.6 (2020), p. 4407-4448 (cf. p. 46).

(Haile, Altmann et al., 2015) N. Haile, J. Altmann et al. "Risk-Benefit-Mediated Impact of Determinants on the Adoption of Cloud Federation." In: *PACIS*. 2015, p. 17 (cf. p. 30).

(Haile et Altmann, 2018) N. Haile et J. Altmann. "Evaluating investments in portability and interoperability between software service platforms". In: *Future Generation Computer Systems* 78 (2018), p. 224-241 (cf. p. 23, 26, 30).

(Hakak et al., 2019) S. I. Hakak, A. Kamsin, P. Shivakumara, G. A. Gilkar, W. Z. Khan et M. Imran. "Exact string matching algorithms: Survey, issues, and future research directions". In: *IEEE access* 7 (2019), p. 69614-69637 (cf. p. 43, 65).

(Hankel et Rexroth, 2015) M. Hankel et B. Rexroth. "The reference architectural model industrie 4.0 (rami 4.0)". In: *ZVEI* 2.2 (2015), p. 4-9 (cf. p. 19).

(Haskins et al., 2006) C. Haskins, K. Forsberg, M. Krueger, D. Walden et D. Hamelin. "Systems engineering handbook". In: *INCOSE*. T. 9. 2006, p. 13-16 (cf. p. 21).

(Al-Hassan, H. Lu et J. Lu, 2015) M. Al-Hassan, H. Lu et J. Lu. "A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system". In: *Decision Support Systems* 72 (2015), p. 97-109 (cf. p. 76).

(Hättasch et al., 2022) B. Hättasch, M. Truong-Ngoc, A. Schmidt et C. Binnig. "It's AI Match: A Two-Step Approach for Schema Matching Using Embeddings". In: *arXiv preprint arXiv:2203.04366* (2022) (cf. p. 51, 97).

(Henning, 2018) F. Henning. "A theoretical framework on the determinants of organisational adoption of interoperability standards in Government Information Networks". In: *Government Information Quarterly* 35.4 (2018), S61-S67 (cf. p. 15).

(Heubusch, 2006) K. Heubusch. "Interoperability: what it means, why it matters". In: *Journal of AHIMA* 77.1 (2006), p. 26-30 (cf. p. 16).

(Hevner et Chatterjee, 2010) A. Hevner et S. Chatterjee. *Design research in information systems theory and practice*. Springer, 2010 (cf. p. 7).

(Heyvaert et al., 2017) P. Heyvaert, A. Dimou, R. Verborgh et E. Mannens. "Ontology-based data access mapping generation using data, schema, query, and mapping knowledge". In: *European Semantic Web Conference*. Springer. 2017, p. 205-215 (cf. p. 52).

(Hodapp et Hanelt, 2022) D. Hodapp et A. Hanelt. "Interoperability in the era of digital innovation: An information systems research agenda". In: *Journal of Information Technology* (2022), p. 02683962211064304 (cf. p. 15).

(Hogan et al., 2021) A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier et al. "Knowledge graphs". In: *ACM Computing Surveys (CSUR)* 54.4 (2021), p. 1-37 (cf. p. 44).

(Hong-Minh et D. Smith, 2007) T. Hong-Minh et D. Smith. "Hierarchical approach for datatype matching in xml schemas". In: *24th British National Conference on Databases (BNCOD'07)*. IEEE. 2007, p. 120-129 (cf. p. 76).

(Hung et al., 2019) N. Q. V. Hung, M. Weidlich, N. T. Tam, Z. Miklós, K. Aberer, A. Gal et B. Stantic. "Handling probabilistic integrity constraints in pay-as-you-go reconciliation of data models". In: *Information Systems* 83 (2019), p. 166-180 (cf. p. 52, 53).

(IDEAS, 2002-2003) IDEAS. *List of IDEAS deliverables, INTEROP-VLab, the European Virtual Laboratory for Enterprise Interoperability (I-VLab)*. 2002-2003. URL: %7Bhttp://interop-vlab.eu/ideas/%7D. (accessed: 16.09.2021) (cf. p. 18).

(IEEE, 1990) IEEE. "IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries". In: *Institute of Electrical and Electronics Engineers* (1990) (cf. p. 16).

(1998). *ISO 14258:* Standard. Geneva, CH: International Organization for Standardization, 1998 (cf. p. 2, 19).

(2000). *ISO 15704:* Standard. Geneva, CH: International Organization for Standardization, 2000 (cf. p. 6).

(INTEROP, 2003-2007) INTEROP. *List of INTEROP deliverables, , INTEROP-VLab, the European Virtual Laboratory for Enterprise Interoperability (I-VLab)*. 2003-2007. URL: %7Bhttp://interop-vlab.eu/interop/%7D. (accessed: 16.09.2021) (cf. p. 19).

(Iroju et al., 2013) O. Iroju, A. Soriyan, I. Gambo et J. Olaleke. "Interoperability in healthcare: benefits, challenges and resolutions". In: *International Journal of Innovation and Applied Studies* 3.1 (2013), p. 262-270 (cf. p. 18).

(Ivanov et Voigt, 2011) P. Ivanov et K. Voigt. "Schema, ontology and metamodel matching-different, but indeed the same?" In: *International Conference on Model and Data Engineering*. Springer. 2011, p. 18-30 (cf. p. 51).

(Izquierdo et al., 2018) Y. T. Izquierdo, G. M. García, E. S. Menendez, M. A. Casanova, F. Dartayre et C. H. Levy. "QUIOW: a keyword-based query processing tool for RDF datasets and relational databases". In: *International Conference on Database and Expert Systems Applications*. Springer. 2018, p. 259-269 (cf. p. 91).

(Jabin, Dimyadi et Amor, 2019) J. Jabin, J. Dimyadi et R. Amor. *Systematic literature review on interoperability measurement models*. Rapp. tech. Technical report]. doi: 10.13140/RG. 2.2. 33957.35047, 2019 (cf. p. 23).

(Jaccard, 1912) P. Jaccard. "The distribution of the flora in the alpine zone. 1". In: *New phytologist* 11.2 (1912), p. 37-50 (cf. p. 72).

(Jandyal et al., 2022) A. Jandyal, I. Chaturvedi, I. Wazir, A. Raina et M. I. U. Haq. "3D printing–A review of processes, materials and applications in industry 4.0". In: *Sustainable Operations and Computers* 3 (2022), p. 33-42 (cf. p. 11).

(Jardim-Goncalves et al., 2013) R. Jardim-Goncalves, A. Grilo, C. Agostinho, F. Lampathaki et Y. Charalabidis. "Systematisation of Interoperability Body of Knowledge: the foundation for Enterprise Interoperability as a science". In: *Enterprise Information Systems* 7.1 (2013), p. 7-32 (cf. p. 4, 14, 16).

(Jaro, 1989) M. A. Jaro. "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida". In: *Journal of the American Statistical Association* 84.406 (1989), p. 414-420 (cf. p. 71).

(Javaid et al., 2022) M. Javaid, A. Haleem, R. P. Singh et R. Suman. "Artificial intelligence applications for industry 4.0: A literature-based study". In: *Journal of Industrial Integration and Management* 7.01 (2022), p. 83-111 (cf. p. 11).

(Jayashree et Marthandan, 2010) S. Jayashree et G. Marthandan. "Government to E-government to E-society". In: *Journal of Applied Sciences(Faisalabad)* 10.19 (2010), p. 2205-2210 (cf. p. 17).

(Jazdi, 2014) N. Jazdi. "Cyber physical systems in the context of Industry 4.0". In: *2014 IEEE international conference on automation, quality and testing, robotics*. IEEE. 2014, p. 1-4 (cf. p. 11).

(Jeong et al., 2008) B. Jeong, D. Lee, H. Cho et J. Lee. "A novel method for measuring semantic similarity for XML schema matching". In: *Expert Systems with Applications* 34.3 (2008), p. 1651-1658 (cf. p. 43).

(Jha et al., 2008) A. K. Jha, D. Doolan, D. Grandt, T. Scott et D. W. Bates. "The use of health information technology in seven nations". In: *International journal of medical informatics* 77.12 (2008), p. 848-854 (cf. p. 16).

(G. Jiang, Cybenko et Hendler, 2006) G. Jiang, G. Cybenko et J. A. Hendler. "Semantic interoperability and information fluidity". In: *International Journal of Cooperative Information Systems* 15.01 (2006), p. 1-21 (cf. p. 114).

(Jimenez, Solanas et Falcone, 2014) C. E. Jimenez, A. Solanas et F. Falcone. "E-government interoperability: Linking open and smart government". In: *Computer* 47.10 (2014), p. 22-24 (cf. p. 17).

(L. Johnson, 2021) L. Johnson. "What is a System?" In: *Student Works* (2021) (cf. p. 6).

(Di-Jorio, Laurent et Teisseire, 2009) L. Di-Jorio, A. Laurent et M. Teisseire. "Mining frequent gradual itemsets from large databases". In: *International Symposium on Intelligent Data Analysis*. Springer. 2009, p. 297-308 (cf. p. 35, 114).

(Kalb et al., 2013) H. Kalb, P. Lazaridou, E. Pinsent et M. Trier. "Interoperability of web archives and digital libraries: A Delphi study". In: *Proceedings of the 10th International Conference on Preservation of Digital Objects*. Biblioteca Nacional de Portugal. 2013, p. 19-28 (cf. p. 14, 15).

(Kang et Naughton, 2003) J. Kang et J. F. Naughton. "On schema matching with opaque column names and data values". In: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. 2003, p. 205-216 (cf. p. 51).

(Kang et Naughton, 2008) J. Kang et J. F. Naughton. "Schema matching using interattribute dependencies". In: *IEEE Transactions on Knowledge and Data Engineering* 20.10 (2008), p. 1393-1407 (cf. p. 47).

(Kapil, Agrawal et R. Khan, 2016) G. Kapil, A. Agrawal et R. Khan. "A study of big data characteristics". In: *2016 International Conference on Communication and Electronics Systems (ICCES)*. IEEE. 2016, p. 1-4 (cf. p. 43).

(Kasemsap, 2018) K. Kasemsap. "The role of information system within enterprise architecture and their impact on business performance". In: *Global Business Expansion: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2018, p. 1078-1102 (cf. p. 7).

(Kaur, Sharma et Kahlon, 2017) K. Kaur, D. S. Sharma et D. K. S. Kahlon. "Interoperability and portability approaches in inter-connected clouds: A review". In: *ACM Computing Surveys (CSUR)* 50.4 (2017), p. 1-40 (cf. p. 17).

(Keele et al., 2007) S. Keele et al. *Guidelines for performing systematic literature reviews in software engineering*. Rapp. tech. Citeseer, 2007 (cf. p. 117).

(Keevash et Mycroft, 2014) P. Keevash et R. Mycroft. *A geometric theory for hypergraph matching*. American Mathematical Soc., 2014 (cf. p. 60).

(Kiadehi et Mohammadi, 2012) E. F. Kiadehi et S. Mohammadi. "Cloud ERP: Implementation of enterprise resource planning using cloud computing technology". In: *Journal of Basic and Applied Scientific Research* 2.11 (2012), p. 11422-11427 (cf. p. 9).

(Kierkegaard, 2015) P. Kierkegaard. "Interoperability after deployment: persistent challenges and regional strategies in Denmark". In: *Int J Qual Health Care* 27.2 (2015), p. 147-153 (cf. p. 16).

(Kim et al., 2011) J. Kim, Y. Peng, N. Ivezik, J. Shin et al. "An optimization approach for semantic-based XML schema matching". In: *International Journal of Trade, Economics, and Finance* 2.1 (2011), p. 78-86 (cf. p. 52).

(Kiourtis et al., 2017) A. Kiourtis, A. Mavrogiorgou, D. Kyriazis et M. Themistocleous. "Acquiring the ontological representation of healthcare data through metamodeling techniques". In: *European, Mediterranean, and Middle Eastern Conference on Information Systems*. Springer. 2017, p. 324-336 (cf. p. 26).

(Kitchenham et al., 2009) B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey et S. Linkman. "Systematic literature reviews in software engineering–a systematic literature review". In: *Information and software technology* 51.1 (2009), p. 7-15 (cf. p. 117).

(Klann et al., 2016) J. G. Klann, A. Abend, V. A. Raghavan, K. D. Mandl et S. N. Murphy. "Data interchange using i2b2". In: *Journal of the American Medical Informatics Association* 23.5 (2016), p. 909-915 (cf. p. 29).

(Klaus, Rosemann et Gable, 2000) H. Klaus, M. Rosemann et G. G. Gable. "What is ERP?" In: *Information systems frontiers* 2.2 (2000), p. 141-162 (cf. p. 8).

(Kok et Domingos, 2009) S. Kok et P. Domingos. "Learning Markov logic network structure via hypergraph lifting". In: *Proceedings of the 26th annual international conference on machine learning*. 2009, p. 505-512 (cf. p. 114).

(Kolaitis, 2005) P. G. Kolaitis. "Schema mappings, data exchange, and metadata management". In: *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2005, p. 61-75 (cf. p. 43).

(Konstantas et al., 2006) D. Konstantas, N. Boudjlida, J.-P. Bourrières et M. Léonard. *Interoperability of enterprise software and applications*. Springer, 2006 (cf. p. 16).

(Kotzé et Neaga, 2010) P. Kotzé et I. Neaga. "Towards an enterprise interoperability framework". In: (2010) (cf. p. 12-15, 23, 27).

(Kouroubali et Katehakis, 2019) A. Kouroubali et D. G. Katehakis. "The new European interoperability framework as a facilitator of digital transformation for citizen empowerment". In: *Journal of biomedical informatics* 94 (2019), p. 103166 (cf. p. 19).

(Koutras et al., 2021) C. Koutras, K. Psarakis, G. Siachamis, A. Ionescu, M. Fragkoulis, A. Bonifati et A. Katsifodimos. "Valentine in action: matching tabular data at scale". In: *Proceedings of the VLDB Endowment* 14.12 (2021), p. 2871-2874 (cf. p. 52).

(Kramer, 2021) D. Kramer. "Research and Development of Interoperability Concepts for IoT Platforms". In: (2021) (cf. p. 23).

(Krishnan, 2013) K. Krishnan. *Data warehousing in the age of big data*. Newnes, 2013 (cf. p. 12, 33).

(Kurilovas, 2009) E. Kurilovas. "Interoperability, standards and metadata for e-Learning". In: *Intelligent distributed computing iii*. Springer, 2009, p. 121-130 (cf. p. 18).

(Kurtzberg, 1962) J. M. Kurtzberg. "On approximation methods for the assignment problem". In: *Journal of the ACM (JACM)* 9.4 (1962), p. 419-439 (cf. p. 77).

(Kurze et al., 2011) T. Kurze, M. Klems, D. Bermbach, A. Lenk, S. Tai et M. Kunze. "Cloud federation". In: *Cloud Computing* 2011 (2011), p. 32-38 (cf. p. 30).

(Labreche et al., 2023a) M. Labreche, X. Lorca, A. Montarnal, S. Weill, J.-P. Adi et T. Sébastien. "A general approach for schema matching problem: case of databases". In: (2023) (cf. p. 113).

(Labreche et al., 2023b) M. Labreche, A. Montarnal, X. Lorca, S. Weill, J.-P. Adi et T. Sébastien. "Towards a general framework of interoperability concepts: the perspective of federated interoperability". In: (2023) (cf. p. 113).

(Labreche et al., 2020) M. Labreche, A. Montarnal, S. Truptil, X. Lorca, S. Weill et J.-P. Adi. "Towards a Framework for Federated Interoperability to Implement an Automated Model Transformation". In: *Working Conference on Virtual Enterprises*. Springer. 2020, p. 143-152 (cf. p. 113).

(Lafi, Feki et Hammoudi, 2014) L. Lafi, J. Feki et S. Hammoudi. "Metamodel matching techniques: Review, comparison and evaluation". In: *International Journal of Information System Modeling and Design (IJISMD)* 5.2 (2014), p. 70-94 (cf. p. 43).

(Lalitha et al., 2019) A. Lalitha, O. C. Kilinc, T. Javidi et F. Koushanfar. "Peer-to-peer federated learning on graphs". In: *arXiv preprint arXiv:1901.11173* (2019) (cf. p. 30).

(Lammari, Comyn-Wattiau et Akoka, 2007) N. Lammari, I. Comyn-Wattiau et J. Akoka. "Extracting generalization hierarchies from relational databases: A reverse engineering approach". In: *Data & Knowledge Engineering* 63.2 (2007), p. 568-589 (cf. p. 114).

(Lampathaki et al., 2012) F. Lampathaki, S. Koussouris, C. Agostinho, R. Jardim-Goncalves, Y. Charalabidis et J. Psarras. "Infusing scientific foundations into Enterprise Interoperability". In: *Computers in Industry* 63.8 (2012), p. 858-866 (cf. p. 14, 18).

(Land, 1985) F. Land. "Is an information theory enough?" In: *The Computer Journal* 28.3 (1985), p. 211-215 (cf. p. 7).

(Lasi et al., 2014) H. Lasi, P. Fettke, H.-G. Kemper, T. Feld et M. Hoffmann. "Industry 4.0". In: *Business & information systems engineering* 6.4 (2014), p. 239-242 (cf. p. 11).

(K. C. Laudon et J. P. Laudon, 2004) K. C. Laudon et J. P. Laudon. *Management information systems: Managing the digital firm*. Pearson Educación, 2004 (cf. p. 7).

(K. C. Laudon et J. P. Laudon, 2018) K. C. Laudon et J. P. Laudon. *Management information systems: Managing the digital firm*. T. Fifteenth Edition. Pearson Educación, 2018 (cf. p. 8, 9).

(Laura, Wesarg et Sakas, 2022) C. O. Laura, S. Wesarg et G. Sakas. "Graph matching survey for medical imaging: On the way to deep learning". In: *Methods* 202 (2022), p. 3-13 (cf. p. 63).

(Lauras et al., 2014) M. Lauras, F. Bénaben, S. Truptil, J. Lamothe, G. Macé-Ramète et A. Montarnal. "A meta-ontology for knowledge acquisition and exploitation of collaborative social systems". In: *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014)*. IEEE. 2014, p. 1-7 (cf. p. 29).

(Lauras et al., 2021) M. Lauras, R. Oger, J. Li, B. Montreuil, M. Kohl, A. Habl et J. Lesbegueries. "Towards a Collaborative and Open Supply Chain Management Operating Services Platform". In: *Working Conference on Virtual Enterprises*. Springer. 2021, p. 611-620 (cf. p. 9).

(Leacock et Chodorow, 1998) C. Leacock et M. Chodorow. "Combining local context and WordNet similarity for word sense identification". In: *WordNet: An electronic lexical database* 49.2 (1998), p. 265-283 (cf. p. 74).

(G. d. S. S. Leal, Guédria et Panetto, 2019) G. d. S. S. Leal, W. Guédria et H. Panetto. "Interoperability assessment: A systematic literature review". In: *Computers in Industry* 106 (2019), p. 111-132 (cf. p. 22, 23).

(G. S. S. Leal et al., 2016) G. S. S. Leal, W. Guédria, H. Panetto et M. Lezoche. "Towards a comparative analysis of interoperability assessment approaches for collaborative enterprise systems". In: *23rd IPSE International Conference on Transdisciplinary Engineering*. Sous la dir. IPSE. T. 4. Advances in Transdiciplinary Engineering. Best student paper award. Curitiba, Brazil: IOS Press, oct. 2016, p. 45-54 (cf. p. 22).

(G. S. Leal, Guédria et Panetto, 2020) G. S. Leal, W. Guédria et H. Panetto. "Enterprise interoperability assessment: a requirements engineering approach". In: *International Journal of Computer Integrated Manufacturing* 33.3 (2020), p. 265-286 (cf. p. 21).

(J. Lee, Cho et K. M. Lee, 2011) J. Lee, M. Cho et K. M. Lee. "Hyper-graph matching via reweighted random walks". In: *CVPR 2011*. IEEE. 2011, p. 1633-1640 (cf. p. 38).

(Lemrabet, 2012) Y. Lemrabet. "Proposition d'une méthode de spécification d'une architecture orientée services dirigée par le métier dans le cadre d'une collaboration inter-organisationnelle". Thèse de doct. Ecole centrale de Lille, 2012 (cf. p. 26).

(Leng et al., 2022) J. Leng, W. Sha, B. Wang, P. Zheng, C. Zhuang, Q. Liu, T. Wuest, D. Mourtzis et L. Wang. "Industry 5.0: Prospect and retrospect". In: *Journal of Manufacturing Systems* 65 (2022), p. 279-295 (cf. p. 11).

(Levenshtein et al., 1966) V. I. Levenshtein et al. "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet physics doklady*. T. 10. 8. Soviet Union. 1966, p. 707-710 (cf. p. 72).

(Y. Levy et Ellis, 2006) Y. Levy et T. J. Ellis. "A systems approach to conduct an effective literature review in support of information systems research." In: *Informing Science* 9 (2006) (cf. p. 118).

(D. Li, Laurent et Teisseire, 2007) D. Li, A. Laurent et M. Teisseire. "On transversal hypergraph enumeration in mining sequential patterns". In: *11th International Database Engineering and Applications Symposium (IDEAS 2007)*. IEEE. 2007, p. 303-307 (cf. p. 57).

(J.-Q. Li et al., 2017) J.-Q. Li, F. R. Yu, G. Deng, C. Luo, Z. Ming et Q. Yan. "Industrial internet: A survey on the enabling technologies, applications, and challenges". In: *IEEE Communications Surveys & Tutorials* 19.3 (2017), p. 1504-1526 (cf. p. 12).

(L. Li et al., 2020) L. Li, Y. Fan, M. Tse et K.-Y. Lin. "A review of applications in federated learning". In: *Computers & Industrial Engineering* (2020), p. 106854 (cf. p. 30).

(T. Li et al., 2020) T. Li, A. K. Sahu, A. Talwalkar et V. Smith. "Federated learning: Challenges, methods, and future directions". In: *IEEE Signal Processing Magazine* 37.3 (2020), p. 50-60 (cf. p. 30).

(Y. Li, D.-B. Liu et W.-M. Zhang, 2005) Y. Li, D.-B. Liu et W.-M. Zhang. "Schema matching using neural network". In: *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE. 2005, p. 743-746 (cf. p. 52).

(Liao, Y. Xu et Ling, 2021) X. Liao, Y. Xu et H. Ling. "Hypergraph Neural Networks for Hypergraph Matching". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, p. 1266-1275 (cf. p. 60).

(Lin et al., 1998) D. Lin et al. “An information-theoretic definition of similarity.” In: *Icml*. T. 98. 1998. 1998, p. 296-304 (cf. p. 74).

(Lisboa et Soares, 2014) A. Lisboa et D. Soares. “E-Government interoperability frameworks: a worldwide inventory”. In: *Procedia Technology* 16 (2014), p. 638-648 (cf. p. 18).

(H. Liu, 2011) H. Liu. “Integration of model driven engineering and ontology approaches for solving interoperability issues”. Thèse de doct. Ecole Centrale de Lille, 2011 (cf. p. 25, 29).

(L. Liu et al., 2020) L. Liu, W. Li, N. R. Aljohani, M. D. Lytras, S.-U. Hassan et R. Nawaz. “A framework to evaluate the interoperability of information systems–Measuring the maturity of the business process alignment”. In: *International Journal of Information Management* 54 (2020), p. 102153 (cf. p. 34).

(X. Liu et al., 2021) X. Liu, Q. Tong, X. Liu et Z. Qin. “Ontology Matching: State of the Art, Future Challenges, and Thinking Based on Utilized Information”. In: *IEEE Access* 9 (2021), p. 91235-91243 (cf. p. 48).

(Y. Liu et al., 2015) Y. Liu, C.-J. Sun, L. Lin, X. Wang et Y. Zhao. “Computing semantic text similarity using rich features”. In: *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. 2015, p. 44-52 (cf. p. 44).

(Y. Liu et X. Xu, 2017) Y. Liu et X. Xu. “Industry 4.0 and cloud manufacturing: A comparative analysis”. In: *Journal of Manufacturing Science and Engineering* 139.3 (2017) (cf. p. 11).

(Livi et Rizzi, 2013) L. Livi et A. Rizzi. “The graph matching problem”. In: *Pattern Analysis and Applications* 16.3 (2013), p. 253-283 (cf. p. 61, 63).

(Lorca, 2014) X. Lorca. “Éléments de flexibilité et d’efficacité en programmation par contraintes”. Thèse de doct. Université de Nantes, 2014 (cf. p. 39).

(Lytras et García, 2008) M. D. Lytras et R. García. “Semantic Web applications: a framework for industry and business exploitation–What is needed for the adoption of the Semantic Web from the market and industry”. In: *International Journal of Knowledge and Learning* 4.1 (2008), p. 93-108 (cf. p. 31).

(Lyytinen et Newman, 2006) K. Lyytinen et M. Newman. “Punctuated equilibrium, process models and information system development and change: towards a socio-technical process analysis”. In: *Sprouts: Working papers on information environments, systems and organizations* 6.1 (2006), p. 1-48 (cf. p. 7).

(Lyytinen et Rose, 2006) K. Lyytinen et G. M. Rose. “Information system development agility as organizational learning”. In: *European Journal of Information Systems* 15.2 (2006), p. 183-199 (cf. p. 7).

(Maatuk, Ali et Rossiter, 2008) A. Maatuk, A. Ali et N. Rossiter. “Relational database migration: A perspective”. In: *International Conference on Database and Expert Systems Applications*. Springer. 2008, p. 676-683 (cf. p. 42).

(Maciel et al., 2017) R. S. P. Maciel, J. M. N. David, D. Claro et R. Braga. “Full interoperability: Challenges and opportunities for future information systems”. In: *Sociedade Brasileira de Computação* (2017) (cf. p. 19, 21).

(Maddikunta et al., 2022) P. K. R. Maddikunta, Q.-V. Pham, B. Prabadevi, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby et M. Liyanage. “Industry 5.0: A survey on enabling technologies and potential applications”. In: *Journal of Industrial Information Integration* 26 (2022), p. 100257 (cf. p. 11).

(Madhavan, Bernstein et Rahm, 2001) J. Madhavan, P. A. Bernstein et E. Rahm. “Generic schema matching with cupid”. In: *vldb*. T. 1. 2001. 2001, p. 49-58 (cf. p. 52, 76).

(Madi, 2016) K. Madi. “Inexact graph matching: application to 2D and 3D Pattern Recognition”. Thèse de doct. Université de Lyon, 2016 (cf. p. 63).

(Mahanti, 2021) R. Mahanti. “Data Governance and Data Management Functions and Initiatives”. In: *Data Governance and Data Management*. Springer, 2021, p. 83-143 (cf. p. 42).

(Mahmood, Farooq et Ferzund, 2017) A. Mahmood, H. Farooq et J. Ferzund. "Large Scale Graph Matching (LSGM): Techniques, Tools, Applications and Challenges". In: *International Journal of Advanced Computer Science and Applications* 8.4 (2017) (cf. p. 52).

(Malhotra et Temponi, 2010) R. Malhotra et C. Temponi. "Critical decisions for ERP integration: Small business issues". In: *International Journal of Information Management* 30.1 (2010), p. 28-37 (cf. p. 9).

(Mallek, Daclin et Chapurlat, 2011) S. Mallek, N. Daclin et V. Chapurlat. "An approach for interoperability requirements specification and verification". In: *International IFIP Working Conference on Enterprise Interoperability*. Springer. 2011, p. 89-102 (cf. p. 21).

(Mallek, Daclin et Chapurlat, 2012) S. Mallek, N. Daclin et V. Chapurlat. "The application of interoperability requirement specification and verification to collaborative processes in industry". In: *Computers in industry* 63.7 (2012), p. 643-658 (cf. p. 21).

(Mallek et al., 2015) S. Mallek, N. Daclin, V. Chapurlat et B. Vallespir. "Enabling model checking for collaborative process analysis: from bpmn to 'network of timed automata'". In: *Enterprise Information Systems* 9.3 (2015), p. 279-299 (cf. p. 21).

(Manso et Wachowicz, 2009) M.-Á. Manso et M. Wachowicz. "GIS design: A review of current issues in interoperability". In: *Geography Compass* 3.3 (2009), p. 1105-1124 (cf. p. 12).

(Mansuri et Sarawagi, 2006) I. R. Mansuri et S. Sarawagi. "Integrating unstructured data into relational databases". In: *22nd International Conference on Data Engineering (ICDE'06)*. IEEE. 2006, p. 29-29 (cf. p. 114).

(Mantzana, Koumaditis et Themistocleous, 2011) V. Mantzana, K. Koumaditis et M. Themistocleous. "HEALTHCARE IS INTEROPERABILITY-Challenges and Solutions". In: *International Conference on Health Informatics*. T. 2. SCITEPRESS. 2011, p. 559-562 (cf. p. 17).

(Margariti et al., 2022) V. Margariti, T. Stamati, D. Anagnostopoulos, M. Nikolaidou et A. Papastilianou. "A holistic model for assessing organizational interoperability in public administration". In: *Government Information Quarterly* (2022), p. 101712 (cf. p. 22, 23).

(Margariti et al., 2020) V. Margariti, D. Anagnostopoulos, A. Papastilianou, T. Stamati et S. Angeli. "Assessment of organizational interoperability in e-Government: a new model and tool for assessing organizational interoperability maturity of a public service in practice". In: *Proceedings of the 13th international conference on theory and practice of electronic governance*. 2020, p. 298-308 (cf. p. 23).

(Martinez-Gil et Aldana-Montes, 2011) J. Martinez-Gil et J. F. Aldana-Montes. "Evaluation of two heuristic approaches to solve the ontology meta-matching problem". In: *Knowledge and Information Systems* 26.2 (2011), p. 225-247 (cf. p. 80).

(Masmoudi et al., 2021) M. Masmoudi, S. B. A. B. Lamine, H. B. Zghal, B. Archimede et M. H. Karray. "Knowledge hypergraph-based approach for data integration and querying: Application to Earth Observation". In: *Future Generation Computer Systems* 115 (2021), p. 720-740 (cf. p. 57).

(Massmann et al., 2011) S. Massmann, S. Raunich, D. Aumüller, P. Arnold, E. Rahm et al. "Evolution of the COMA match system". In: *Ontology Matching* 49 (2011), p. 49-60 (cf. p. 48).

(May, 1999) W. May. *Information Extraction and Integration with* FLORID*: The* MONDIAL *Case Study*. Rapp. tech. 131. Available from `http://dbis.informatik.uni-goettingen.de/Mondial`. Universität Freiburg, Institut für Informatik, 1999 (cf. p. 91).

(Mazilu et al., 2022) L. Mazilu, N. W. Paton, A. A. Fernandes et M. Koehler. "Schema mapping generation in the wild". In: *Information Systems* 104 (2022), p. 101904 (cf. p. 51).

(Megdiche, Teste et Trojahn, 2016) I. Megdiche, O. Teste et C. Trojahn. "An extensible linear approach for holistic ontology matching". In: *International Semantic Web Conference*. Springer. 2016, p. 393-410 (cf. p. 53).

(Meijer et Bolívar, 2016) A. Meijer et M. P. R. Bolívar. "Governing the smart city: a review of the literature on smart urban governance". In: *Revue Internationale des Sciences Administratives* 82.2 (2016), p. 417-435 (cf. p. 26).

(Meinadier, 2003) J.-P. Meinadier. *Le métier d'intégration de systèmes*. Hermes Science Publications, 2003 (cf. p. 13).

(Melnik, 2004) S. Melnik. *Generic model management: concepts and algorithms*. T. 2967. Springer Science & Business Media, 2004 (cf. p. 50).

(Melnik, Garcia-Molina et Rahm, 2002) S. Melnik, H. Garcia-Molina et E. Rahm. "Similarity flooding: A versatile graph matching algorithm and its application to schema matching". In: *Proceedings 18th international conference on data engineering*. IEEE. 2002, p. 117-128 (cf. p. 50).

(Memon et al., 2014) M. Memon, S. R. Wagner, C. F. Pedersen, F. H. A. Beevi et F. O. Hansen. "Ambient assisted living healthcare frameworks, platforms, standards, and quality attributes". In: *Sensors* 14.3 (2014), p. 4312-4341 (cf. p. 19).

(Merriam-Webster, 2022) Merriam-Webster. *System*. In: *Merriam-Webster.com dictionary* (cf. p. 6).

(Meystre et al., 2017) S. M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis et C. U. Lehmann. "Clinical data reuse or secondary use: current status and potential future progress". In: *Yearbook of medical informatics* 26.01 (2017), p. 38-52 (cf. p. 17).

(Mezgár et Rauschecker, 2014) I. Mezgár et U. Rauschecker. "The challenge of networked enterprises for cloud computing interoperability". In: *Computers in Industry* 65.4 (2014), p. 657-674 (cf. p. 10).

(Michelbacher, 2013) L. Michelbacher. "Multi-word tokenization for natural language processing". In: (2013) (cf. p. 43).

(Mikolov et al., 2013) T. Mikolov, K. Chen, G. Corrado et J. Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013) (cf. p. 74).

(Mikolov et al., 2017) T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch et A. Joulin. "Advances in pre-training distributed word representations". In: *arXiv preprint arXiv:1712.09405* (2017) (cf. p. 44).

(G. A. Miller, 1995) G. A. Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), p. 39-41 (cf. p. 43, 74).

(R. Miller, 2013) R. Miller. *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20 22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department*. Springer Science & Business Media, 2013 (cf. p. 62).

(Minami et Dawson, 2008) C. Minami et J. Dawson. "The CRM process in retail and service sector firms in Japan: Loyalty development and financial return". In: *Journal of Retailing and Consumer Services* 15.5 (2008), p. 375-385 (cf. p. 9).

(Missikoff, 2009) M. Missikoff. "On the scientific basis of Enterprise Interoperability". In: *2009 IEEE International Technology Management Conference (ICE)*. IEEE. 2009, p. 1-9 (cf. p. 14).

(Mondorf et Wimmer, 2016) A. Mondorf et M. A. Wimmer. "Requirements for an architecture framework for Pan-European e-government services". In: *International Conference on Electronic Government*. Springer. 2016, p. 135-150 (cf. p. 17).

(Montarnal, 2015) A. Montarnal. "Deduction of inter-organizational collaborative business processes within an enterprise social network". Thèse de doct. Ecole des Mines d'Albi-Carmaux, 2015 (cf. p. 3, 29).

(Morley et al., 2005) C. Morley, J. Hugues, B. Leblanc et O. Hugues. "Processus métiers et SI: évaluation, modélisation, mise en oeuvre, éditions DUNOD". In: *mars* (2005) (cf. p. 7).

(Morris, 2012) J. Morris. *Practical data migration*. BCS, The Chartered Institute, 2012 (cf. p. 42).

(Mothukuri et al., 2021) V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha et G. Srivastava. "A survey on security and privacy of federated learning". In: *Future Generation Computer Systems* 115 (2021), p. 619-640 (cf. p. 30).

(R. C. Motta, Oliveira et Travassos, 2019) R. C. Motta, K. M. de Oliveira et G. H. Travassos. "A conceptual perspective on interoperability in context-aware software systems". In: *Information and Software Technology* 114 (2019), p. 231-257 (cf. p. 15).

(Mu et al., 2017) W. Mu, F. Benaben, N. Boissel-Dallier et H. Pingaud. "collaborative knowledge framework for mediation information system engineering". In: *Scientific Programming* 2017 (2017) (cf. p. 3).

(Mu et al., 2011) W. Mu, F. Benaben, H. Pingaud, N. Boissel-Dallier et J.-P. Lorré. "A model-driven BPM approach for SOA mediation information system design in a collaborative context". In: *2011 IEEE International Conference on Services Computing*. IEEE. 2011, p. 747-748 (cf. p. 3).

(Mu, Bénaben et Pingaud, 2015) W. Mu, F. Bénaben et H. Pingaud. "A methodology proposal for collaborative business process elaboration using a model-driven approach". In: *Enterprise Information Systems* 9.4 (2015), p. 349-383 (cf. p. 3).

(Muller et al., 2019) M. F. Muller, F. Esmanioto, N. Huber, E. R. Loures et O. C. Junior. "A systematic literature review of interoperability in the green Building Information Modeling lifecycle". In: *Journal of cleaner production* 223 (2019), p. 397-412 (cf. p. 18).

(Murtuza, 2022) S. Murtuza. "Internet of Everything: Application and Various Challenges Analysis a Survey". In: *2022 1st International Conference on Informatics (ICI)*. IEEE. 2022, p. 250-252 (cf. p. 12).

(Mykkänen et Tuomainen, 2008) J. A. Mykkänen et M. P. Tuomainen. "An evaluation and selection framework for interoperability standards". In: *Information and Software Technology* 50.3 (2008), p. 176-197 (cf. p. 18).

(NABİYEV et al., 2016) V. NABİYEV, Ü. Çakiroğlu, H. Karal, A. K. ERÜMİT et Ç. Ayça. "Application of graph theory in an intelligent tutoring system for solving mathematical word problems". In: *Eurasia Journal of Mathematics, Science and Technology Education* 12.4 (2016), p. 687-701 (cf. p. 35).

(Naudet et al., 2010) Y. Naudet, T. Latour, W. Guedria et D. Chen. "Towards a systemic formalisation of interoperability". In: *Computers in Industry* 61.2 (2010), p. 176-185 (cf. p. 12, 13, 28).

(Naudet et al., 2006) Y. Naudet, T. Latour, K. Hausmann, S. Abels, A. Hahn et P. Johannesson. "Describing Interoperability: the OoI Ontology." In: *EMOI-INTEROP*. 2006 (cf. p. 13).

(Naumann et Herschel, 2010) F. Naumann et M. Herschel. "An introduction to duplicate detection". In: *Synthesis Lectures on Data Management* 2.1 (2010), p. 1-87 (cf. p. 42).

(Nayak et Tran, 2007) R. Nayak et T. Tran. "A progressive clustering algorithm to group the XML data by structural and semantic similarity". In: *International Journal of Pattern Recognition and Artificial Intelligence* 21.04 (2007), p. 723-743 (cf. p. 76).

(Neiva et al., 2016) F. W. Neiva, J. M. N. David, R. Braga et F. Campos. "Towards pragmatic interoperability to support collaboration: A systematic review and mapping of the literature". In: *Information and Software Technology* 72 (2016), p. 137-150 (cf. p. 18).

(Ngo et Bellahsene, 2016) D. Ngo et Z. Bellahsene. "Overview of YAM++—(not) Yet Another Matcher for ontology alignment task". In: *Journal of Web Semantics* 41 (2016), p. 30-49 (cf. p. 44, 50).

(Nguyen, 2014) Q. V. H. Nguyen. *Reconciling schema matching networks*. Rapp. tech. EPFL, 2014 (cf. p. 53).

(Nikolov, Uren et E. Motta, 2010) A. Nikolov, V. S. Uren et E. Motta. "Data linking: Capturing and utilising implicit schema-level relations". In: *LDOW*. 2010 (cf. p. 44).

(Nodehi et al., 2017) T. Nodehi, R. Jardim-Goncalves, A. Zutshi et A. Grilo. "ICIF: an inter-cloud interoperability framework for computing resource cloud providers in factories of the future". In: *International Journal of Computer Integrated Manufacturing* 30.1 (2017), p. 147-157 (cf. p. 19).

(Nof et al., 2008) S. Y. Nof, F. G. Filip, A. Molina, L. Monostori et C. E. Pereira. "Advances in e-Manufacturing, e-Logistics, and e-Service Systems Milestone report prepared by IFAC Coordinating Committee on Manufacturing & Logistics Systems". In: *IFAC Proceedings Volumes* 41.2 (2008), p. 5742-5750 (cf. p. 14).

(Nogueira et al., 2016) E. Nogueira, A. Moreira, D. Lucrédio, V. Garcia et R. Fortes. "Issues on developing interoperable cloud applications: definitions, concepts, approaches, requirements, characteristics and evaluation models". In: *Journal of Software Engineering Research and Development* 4.1 (2016), p. 1-23 (cf. p. 23).

(Noran et Bernus, 2011) O. Noran et P. Bernus. "Effective disaster management: an interoperability perspective". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2011, p. 112-121 (cf. p. 18).

(Noran et Zdravković, 2014) O. Noran et M. Zdravković. "Interoperability as a property: enabling an agile disaster management approach". In: *Proceedings of the 4th International Conference on Information Society and Technology (ICIST 2014)*. T. 1. 2014, p. 248-255 (cf. p. 26, 31).

(Noumeir, 2012) R. Noumeir. "Requirements for interoperability in healthcare information systems". In: *Journal of Healthcare Engineering* 3.2 (2012), p. 323-346 (cf. p. 18).

(O'brien et Marakas, 2006) J. A. O'brien et G. M. Marakas. *Management information systems*. T. 6. McGraw-Hill Irwin, 2006 (cf. p. 7).

(OAEI, 2022) OAEI. *Ontology Alignment Evaluation Initiative*. 2022. URL: http://oaei.ontologymatching.org. (accessed: 02.12.2022) (cf. p. 52).

(Ochieng et Kyanda, 2018) P. Ochieng et S. Kyanda. "Large-scale ontology matching: State-of-the-art analysis". In: *ACM Computing Surveys (CSUR)* 51.4 (2018), p. 1-35 (cf. p. 48).

(Öhlund, 2017) S.-E. Öhlund. "Interoperability Capability to interoperate in a shared work practice using information infrastructures: studies in ePrescribing". Thèse de doct. Linköping University Electronic Press, 2017 (cf. p. 13).

(Okano, 2017) M. T. Okano. "IOT and industry 4.0: the industrial new revolution". In: *International Conference on Management and Information Systems*. T. 25. 2017, p. 26 (cf. p. 11).

(Okoli et Pawlowski, 2004) C. Okoli et S. D. Pawlowski. "The Delphi method as a research tool: an example, design considerations and applications". In: *Information & management* 42.1 (2004), p. 15-29 (cf. p. 14).

(Otero-Cerdeira, Rodríguez-Martínez et Gómez-Rodríguez, 2015) L. Otero-Cerdeira, F. J. Rodríguez-Martínez et A. Gómez-Rodríguez. "Ontology matching: A literature review". In: *Expert Systems with Applications* 42.2 (2015), p. 949-971 (cf. p. 48).

(Otjacques, Hitzelberger et Feltz, 2007) B. Otjacques, P. Hitzelberger et F. Feltz. "Interoperability of e-government information systems: Issues of identification and data sharing". In: *Journal of management information systems* 23.4 (2007), p. 29-51 (cf. p. 17).

(Oyeyemi et Scott, 2018) A. Oyeyemi et P. Scott. "Interoperability in health and social care: organisational issues are the biggest challenge". In: *BMJ Health & Care Informatics* 25.3 (2018) (cf. p. 17).

(Paepcke et al., 1998) A. Paepcke, C.-C. K. Chang, T. Winograd et H. García-Molina. "Interoperability for digital libraries worldwide". In: *Communications of the ACM* 41.4 (1998), p. 33-42 (cf. p. 16).

(Palfrey et Gasser, 2012) J. Palfrey et U. Gasser. *Interop*. 2012 (cf. p. 27).

(Panetto, 2007) H. Panetto. "Towards a classification framework for interoperability of enterprise applications". In: *International Journal of Computer Integrated Manufacturing* 20.8 (2007), p. 727-740 (cf. p. 18).

(Panetto et Molina, 2008) H. Panetto et A. Molina. "Enterprise integration and interoperability in manufacturing systems: Trends and issues". In: *Computers in industry* 59.7 (2008), p. 641-646 (cf. p. 12).

(Panetto, Scannapieco et Zelm, 2004) H. Panetto, M. Scannapieco et M. Zelm. "INTEROP NoE: Interoperability research for networked enterprises applications and software". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2004, p. 866-882 (cf. p. 19).

(Panetto et al., 2016) H. Panetto, M. Zdravkovic, R. Jardim-Goncalves, D. Romero, J. Cecil et I. Mezgár. "New perspectives for the future interoperable enterprise systems". In: *Computers in industry* 79 (2016), p. 47-63 (cf. p. 14, 16).

(Papenbrock et al., 2015) T. Papenbrock, J. Ehrlich, J. Marten, T. Neubert, J.-P. Rudolph, M. Schönberg, J. Zwiener et F. Naumann. "Functional dependency discovery: An experimental evaluation of seven algorithms". In: *Proceedings of the VLDB Endowment* 8.10 (2015), p. 1082-1093 (cf. p. 60).

(Pardo, Nam et Burke, 2012) T. A. Pardo, T. Nam et G. B. Burke. "E-government interoperability: Interaction of policy, management, and technology dimensions". In: *Social Science Computer Review* 30.1 (2012), p. 7-23 (cf. p. 18).

(Parundekar, Knoblock et Ambite, 2014) R. Parundekar, C. A. Knoblock et J. L. Ambite. *Aligning Ontologies of Linked Data.* 2014 (cf. p. 43).

(Patel, Debnath et Bhushan, 2022) A. Patel, N. C. Debnath et B. Bhushan. *Semantic Web Technologies: Research and Applications*. CRC Press, 2022 (cf. p. 52).

(Pather, 2017) S. Pather. "Contextualising Information Systems Evaluation Research: Towards a Classification of Approaches". In: *The European Conference on Information Systems Management*. Academic Conferences International Limited. 2017, p. 252-261 (cf. p. 7).

(Paul, 2010) R. J. Paul. "What an information system is, and why is it important to know this". In: *Journal of Computing and Information Technology* 18.2 (2010), p. 95-99 (cf. p. 7).

(Pearlson, Saunders et Galletta, 2016) K. E. Pearlson, C. S. Saunders et D. F. Galletta. *Managing and using information systems: A strategic approach*. John Wiley & Sons, 2016 (cf. p. 9).

(X. Peng et al., 2013) X. Peng, Z. Xing, X. Tan, Y. Yu et W. Zhao. "Improving feature location using structural similarity and iterative graph mapping". In: *Journal of Systems and Software* 86.3 (2013), p. 664-676 (cf. p. 71).

(Pennington, Socher et Manning, 2014) J. Pennington, R. Socher et C. D. Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, p. 1532-1543 (cf. p. 74).

(Perafán, Correa et Buitron, 2020) D. E. P. Perafán, F. J. P. Correa et S. L. Buitron. "A model for the elicitation of organizational system interoperability requirements". In: *Ingeniería Solidaria* 16.3 (2020), p. 1-36 (cf. p. 21).

(Peristeras, Tarabanis et Goudos, 2009) V. Peristeras, K. Tarabanis et S. K. Goudos. "Model-driven eGovernment interoperability: A review of the state of the art". In: *Computer Standards & Interfaces* 31.4 (2009), p. 613-628 (cf. p. 17).

(Peschl et Del Fabro, 2015) G. Peschl et M. D. Del Fabro. "Restricted metamodel-based similarity propagation: a comparative study." In: *CIbSE*. 2015, p. 25 (cf. p. 50).

(Petcu, 2011) D. Petcu. "Portability and interoperability between clouds: challenges and case study". In: *European conference on a service-based internet*. Springer. 2011, p. 62-74 (cf. p. 21).

(Petter, DeLone et McLean, 2012) S. Petter, W. DeLone et E. R. McLean. "The past, present, and future of "IS success"". In: *Journal of the Association for Information Systems* 13.5 (2012), p. 2 (cf. p. 10).

(Peukert, Eberius et Rahm, 2011) E. Peukert, J. Eberius et E. Rahm. "Rule-based construction of matching processes". In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011, p. 2421-2424 (cf. p. 52).

(Piest, Iacob et Sinderen, 2020) J. P. S. Piest, M.-E. Iacob et M. van Sinderen. "A Federated Interoperability Approach for Data Driven Logistic Support in SMEs." In: *I-ESA Workshops*. 2020 (cf. p. 29).

(Piest et al., 2020) J. P. S. Piest, L. O. Meertens, J. Buis, M. E. Iacob et M. J. van Sinderen. "Smarter interoperability based on automatic schema matching and intelligence amplification". In: *10th I-ESA SIFAI Workshop. URL http://ceur-ws. org*. T. 2900. 2020 (cf. p. 29).

(Pillai et al., 2021) S. G. Pillai, K. Haldorai, W. S. Seo et W. G. Kim. "COVID-19 and hospitality 5.0: Redefining hospitality operations". In: *International Journal of Hospitality Management* 94 (2021), p. 102869 (cf. p. 17).

(Pires et al., 2019) F. Pires, A. Cachada, J. Barbosa, A. P. Moreira et P. Leitão. "Digital twin in industry 4.0: Technologies, applications and challenges". In: *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*. T. 1. IEEE. 2019, p. 721-726 (cf. p. 11).

(Poinet, Stefanescu et Papadonikolaki, 2020) P. Poinet, D. Stefanescu et E. Papadonikolaki. "Collaborative workflows and version control through open-source and distributed common data environment". In: *International Conference on Computing in Civil and Building Engineering*. Springer. 2020, p. 228-247 (cf. p. 29).

(Poppe et al., 2015) K. Poppe, J. Wolfert, C. Verdouw et A. Renwick. "A European perspective on the economics of big data". In: *Farm Policy Journal* 12.1 (2015), p. 11-19 (cf. p. 12).

(Popplewell, 2011) K. Popplewell. "Towards the definition of a science base for enterprise interoperability: a European perspective". In: *Journal of Systemics, Cybernetics, and Informatics* 9.5 (2011), p. 6-11 (cf. p. 14).

(Portisch, Hladik et Paulheim, 2021) J. Portisch, M. Hladik et H. Paulheim. "Background knowledge in schema matching: Strategy vs. data". In: *International Semantic Web Conference*. Springer. 2021, p. 287-303 (cf. p. 44).

(Pouliot et al., 2018) J. Pouliot, S. Larrivée, C. Ellul et A. Boudhaim. "Exploring schema matching to compare geospatial standards: application to underground utility networks". In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives* 42.4/W10 (2018), p. 157-164 (cf. p. 51).

(Pradhan, Gyanchandani et Wadhvani, 2015) N. Pradhan, M. Gyanchandani et R. Wadhvani. "A Review on Text Similarity Technique used in IR and its Application". In: *International Journal of Computer Applications* 120.9 (2015), p. 29-34 (cf. p. 51, 65).

(Prakoso, Abdi et Amrit, 2021) D. W. Prakoso, A. Abdi et C. Amrit. "Short text similarity measurement methods: a review". In: *Soft Computing* 25.6 (2021), p. 4699-4723 (cf. p. 43, 65).

(Quoc Viet Nguyen et al., 2013) H. Quoc Viet Nguyen, X. H. Luong, Z. Miklós, T. T. Quan et K. Aberer. "Collaborative schema matching reconciliation". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2013, p. 222-240 (cf. p. 44).

(Rada et al., 1989) R. Rada, H. Mili, E. Bicknell et M. Blettner. "Development and application of a metric on semantic nets". In: *IEEE transactions on systems, man, and cybernetics* 19.1 (1989), p. 17-30 (cf. p. 74).

(Rahm, 2011) E. Rahm. "Towards large-scale schema and ontology matching". In: *Schema matching and mapping*. Springer, 2011, p. 3-27 (cf. p. 45, 48).

(Rahm et Bernstein, 2001) E. Rahm et P. A. Bernstein. "A survey of approaches to automatic schema matching". In: *the VLDB Journal* 10.4 (2001), p. 334-350 (cf. p. 43, 48, 49).

(Rajsiri et al., 2010) V. Rajsiri, J.-P. Lorré, F. Benaben et H. Pingaud. "Knowledge-based system for collaborative process specification". In: *Computers in Industry* 61.2 (2010), p. 161-175 (cf. p. 3).

(Ram et Park, 2004) S. Ram et J. Park. "Semantic Conflict Resolution Ontology (SCROL): An ontology for detecting and resolving data and schema-level semantic conflicts". In: *IEEE Transactions on Knowledge and Data engineering* 16.2 (2004), p. 189-202 (cf. p. 47).

(Ramapantulu, Teo et Chang, 2017) L. Ramapantulu, Y. M. Teo et E.-C. Chang. "A conceptural framework to federate testbeds for cybersecurity". In: *2017 winter simulation conference (WSC)*. IEEE. 2017, p. 457-468 (cf. p. 28).

(Ramète et al., 2012) G. M. Ramète, J. Lamothe, M. Lauras et F. Benaben. "A road crisis management metamodel for an information decision support system". In: *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. IEEE. 2012, p. 1-5 (cf. p. 3).

(Rashid, Hossain et Patrick, 2002) M. A. Rashid, L. Hossain et J. D. Patrick. "The evolution of ERP systems: A historical perspective". In: *Enterprise resource planning: Solutions and management*. IGI global, 2002, p. 35-50 (cf. p. 9).

(Ratcliff et Metzener, 1988) J. W. Ratcliff et D. E. Metzener. "Pattern-matching-the gestalt approach". In: *Dr Dobbs Journal* 13.7 (1988), p. 46 (cf. p. 73).

(Rauch, 2020) E. Rauch. "Industry 4.0+: The next level of intelligent and self-optimizing factories". In: *Design, Simulation, Manufacturing: The Innovation Exchange*. Springer. 2020, p. 176-186 (cf. p. 11).

(Ray, Gulla et Dash, 2007) D. Ray, U. Gulla et S. S. Dash. "Interoperability of e-government information systems: a survey". In: *Towards next generation e-government* (2007), p. 12-25 (cf. p. 17).

(Rebhi et al., 2017) W. Rebhi, N. B. Yahia, N. B. B. Saoud et C. Hanachi. "Towards contextualizing community detection in dynamic social networks". In: *International and Interdisciplinary Conference on Modeling and Using Context*. Springer. 2017, p. 324-336 (cf. p. 35).

(Reegu, Daud et S. Alam, 2021) F. Reegu, S. M. Daud et S. Alam. "Interoperability Challenges in Healthcare Blockchain System-A Systematic Review". In: *Annals of the Romanian Society for Cell Biology* (2021), p. 15487-15499 (cf. p. 18).

(Resnik, 1995) P. Resnik. "Using information content to evaluate semantic similarity in a taxonomy". In: *arXiv preprint cmp-lg/9511007* (1995) (cf. p. 74).

(Rezaei, T. K. Chiew et S. P. Lee, 2014) R. Rezaei, T. K. Chiew et S. P. Lee. "A review on E-business Interoperability Frameworks". In: *Journal of Systems and Software* 93 (2014), p. 199-216 (cf. p. 18).

(Rezaei et al., 2014a) R. Rezaei, T. K. Chiew, S. P. Lee et Z. S. Aliee. "A semantic interoperability framework for software as a service systems in cloud computing environments". In: *Expert Systems with Applications* 41.13 (2014), p. 5751-5770 (cf. p. 23).

(Rezaei et al., 2014b) R. Rezaei, T. K. Chiew, S. P. Lee et Z. S. Aliee. "Interoperability evaluation models: A systematic review". In: *Computers in Industry* 65.1 (2014), p. 1-23 (cf. p. 21, 23).

(Rezaei, T.-k. Chiew et S.-p. Lee, 2013) R. Rezaei, T.-k. Chiew et S.-p. Lee. "A review of interoperability assessment models". In: *Journal of Zhejiang University SCIENCE C* 14.9 (2013), p. 663-681 (cf. p. 23).

(Riesen, X. Jiang et Bunke, 2010) K. Riesen, X. Jiang et H. Bunke. "Exact and inexact graph matching: Methodology and applications". In: *Managing and Mining Graph Data* (2010), p. 217-247 (cf. p. 62).

(Riley, 2020) C. Riley. "Unpacking interoperability in competition". In: *Journal of Cyber Policy* 5.1 (2020), p. 94-106 (cf. p. 15).

(Robertson et Walker, 1994) S. E. Robertson et S. Walker. "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval". In: *SIGIR'94*. Springer. 1994, p. 232-241 (cf. p. 74).

(Rochwerger et al., 2009) B. Rochwerger, D. Breitgand, E. Levy, A. Galis, K. Nagin, I. M. Llorente, R. Montero, Y. Wolfsthal, E. Elmroth, J. Caceres et al. "The reservoir model and architecture for open federated cloud computing". In: *IBM Journal of Research and Development* 53.4 (2009), p. 4-1 (cf. p. 30).

(Romero et F. Vernadat, 2016) D. Romero et F. Vernadat. "Enterprise information systems state of the art: Past, present and future trends". In: *Computers in Industry* 79 (2016), p. 3-13 (cf. p. 27).

(Ronoh, Omieno et Mutua, 2018) H. Ronoh, K. Omieno et S. Mutua. "An interoperability framework for E-government heterogeneous information systems". In: *IJARCCE* 7.10 (2018), p. 115-126 (cf. p. 10).

(Roque et Chapurlat, 2009) M. Roque et V. Chapurlat. "Interoperability in collaborative processes: Requirements characterisation and proof approach". In: *Working Conference on Virtual Enterprises*. Springer. 2009, p. 555-562 (cf. p. 21).

(Roth et al., 2018) F. M. Roth, C. Becker, G. Vega et P. Lalanda. "XWARE—a customizable interoperability framework for pervasive computing systems". In: *Pervasive and mobile computing* 47 (2018), p. 13-30 (cf. p. 21).

(Rouen, 2013) I. Rouen. "Simulation and optimization of interoperability planning". In: *Journal of Theoretical and Applied Information Technology* 52.3 (2013) (cf. p. 23).

(Roy et al., 2019) A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab et C. Wachinger. "Braintorrent: A peer-to-peer environment for decentralized federated learning". In: *arXiv preprint arXiv:1905.06731* (2019) (cf. p. 30).

(Ruokolainen, 2009) T. Ruokolainen. "Modelling framework for interoperability management in collaborative computing environments". In: *Licentiate thesis, University of Helsinki, Department of Computer Science* (2009) (cf. p. 26).

(Saaksvuori et Immonen, 2008) A. Saaksvuori et A. Immonen. *Product lifecycle management systems*. Springer, 2008 (cf. p. 10).

(Sachsenmeier, 2016) P. Sachsenmeier. "Industry 5.0—The relevance and implications of bionics and synthetic biology". In: *Engineering* 2.2 (2016), p. 225-229 (cf. p. 11).

(Saenz de Ugarte, Artiba et Pellerin, 2009) B. Saenz de Ugarte, A. Artiba et R. Pellerin. "Manufacturing execution system–a literature review". In: *Production planning and control* 20.6 (2009), p. 525-539 (cf. p. 10).

(Saha, Stanoi et Clarkson, 2010) B. Saha, I. Stanoi et K. L. Clarkson. "Schema covering: a step towards enabling reuse in information integration". In: *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*. IEEE. 2010, p. 285-296 (cf. p. 52, 53).

(Sahay, Mehta et Jadon, 2020) T. Sahay, A. Mehta et S. Jadon. "Schema matching using machine learning". In: *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE. 2020, p. 359-366 (cf. p. 52).

(Saïs, 2019) F. Saïs. "Knowledge Graph Refinement: Link Detection, Link Invalidation, Key Discovery and Data Enrichment". Thèse de doct. Université Paris Sud, 2019 (cf. p. 114).

(Sakka, 2012) O. Sakka. "Alignement sémantique entre référentiels d'entreprise: Application aux systèmes d'exécution de la fabrication (MES)". Thèse de doct. INSA de Lyon, 2012 (cf. p. 10).

(Salton et Buckley, 1988) G. Salton et C. Buckley. "Term-weighting approaches in automatic text retrieval". In: *Information processing & management* 24.5 (1988), p. 513-523 (cf. p. 74).

(Sanchez, Terlizzi et al., 2017) O. P. Sanchez, M. A. Terlizzi et al. "Cost and time project management success factors for information systems development projects". In: *International Journal of Project Management* 35.8 (2017), p. 1608-1626 (cf. p. 7).

(Sanfeliu et Fu, 1983) A. Sanfeliu et K.-S. Fu. "A distance measure between attributed relational graphs for pattern recognition". In: *IEEE transactions on systems, man, and cybernetics* 3 (1983), p. 353-362 (cf. p. 62).

(E. M. d. Santos et Reinhard, 2012) E. M. d. Santos et N. Reinhard. "Electronic government interoperability: Identifying the barriers for frameworks adoption". In: *Social Science Computer Review* 30.1 (2012), p. 71-82 (cf. p. 17).

(K. S. S. Santos, Pinheiro et Maciel, 2021) K. S. S. Santos, L. B. L. Pinheiro et R. S. P. Maciel. "Interoperability Types Classifications: A Tertiary Study". In: *XVII Brazilian Symposium on Information Systems*. 2021, p. 1-8 (cf. p. 18, 118).

(Al-Sayed, Hassan et Omara, 2020) M. M. Al-Sayed, H. A. Hassan et F. A. Omara. "CloudFNF: An ontology structure for functional and non-functional features of cloud services". In: *Journal of Parallel and Distributed Computing* 141 (2020), p. 143-173 (cf. p. 26).

(Scholl et Klischewski, 2007) H. J. Scholl et R. Klischewski. "E-government integration and interoperability: framing the research agenda". In: *International Journal of Public Administration* 30.8-9 (2007), p. 889-920 (cf. p. 17).

(Schultz, Spivak et Wisnesky, 2016) P. Schultz, D. I. Spivak et R. Wisnesky. "Algebraic model management: A survey". In: *International Workshop on Algebraic Development Techniques*. Springer. 2016, p. 56-69 (cf. p. 42).

(Schultz et Wisnesky, 2017) P. Schultz et R. Wisnesky. "Algebraic data integration". In: *Journal of Functional Programming* 27 (2017) (cf. p. 42).

(Scopus, 2004) Scopus. *Expertly curated abstract & citation database*. 2004. URL: https://blog.scopus.com/. (accessed: 09.11.2021) (cf. p. 117).

(Sellami, 2009) S. Sellami. "Méthodologie de Matching à Large Echelle pour des schémas XML". Thèse de doct. université lyon 2, 2009 (cf. p. 50).

(Serrano et al., 2017) M. Serrano, A. Gyrard, M. Boniface, P. Grace, N. Georgantas, R. Agarwal, P. Barnagu, F. Carrez, B. Almeida, T. Teixeira et al. "Cross-domain interoperability using federated interoperable semantic IoT/Cloud testbeds and applications: The FIESTA-IoT approach". In: (2017) (cf. p. 30).

(Shehzad et al., 2021) H. M. F. Shehzad, R. B. Ibrahim, A. F. Yusof, K. A. M. Khaidzir, M. Iqbal et S. Razzaq. "The role of interoperability dimensions in building information modelling". In: *Computers in Industry* 129 (2021), p. 103444 (cf. p. 18).

(F. Shi et al., 2018) F. Shi, Q. Li, T. Zhu et H. Ning. "A survey of data semantization in internet of things". In: *Sensors* 18.1 (2018), p. 313 (cf. p. 57).

(Y. Shi et al., 2021) Y. Shi, G. Cheng, T.-K. Tran, J. Tang et E. Kharlamov. "Keyword-Based Knowledge Graph Exploration Based on Quadratic Group Steiner Trees." In: *IJCAI*. T. 2021. 2021, p. 1555-1562 (cf. p. 91).

(Shraga et Gal, 2022) R. Shraga et A. Gal. "PoWareMatch: a Quality-aware Deep Learning Approach to Improve Human Schema Matching". In: *ACM Journal of Data and Information Quality (JDIQ)* 14.3 (2022), p. 1-27 (cf. p. 52).

(Shrestha et al., 2019) M. Shrestha, T. X. Tran, B. Bhattarai, M. L. Pusey et R. S. Aygun. "Schema matching and data integration with consistent naming on protein crystallization screens". In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.6 (2019), p. 2074-2085 (cf. p. 42).

(Shvaiko et Euzenat, 2011) P. Shvaiko et J. Euzenat. "Ontology matching: state of the art and future challenges". In: *IEEE Transactions on knowledge and data engineering* 25.1 (2011), p. 158-176 (cf. p. 43, 44, 48).

(Sinaci et Erturkmen, 2013) A. A. Sinaci et G. B. L. Erturkmen. "A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains". In: *Journal of biomedical informatics* 46.5 (2013), p. 784-794 (cf. p. 28).

(Singh et V. Gupta, 2016) J. Singh et V. Gupta. "Text stemming: Approaches, applications, and challenges". In: *ACM Computing Surveys (CSUR)* 49.3 (2016), p. 1-46 (cf. p. 43).

(Siriginidi, 2000) S. R. Siriginidi. "Enterprise resource planning in reengineering business". In: *Business Process Management Journal* (2000) (cf. p. 8).

(Sisodia et Jindal, 2021) A. Sisodia et R. Jindal. "A meta-analysis of industry 4.0 design principles applied in the health sector". In: *Engineering Applications of Artificial Intelligence* 104 (2021), p. 104377 (cf. p. 11).

(Smiljanić, Keulen et Jonker, 2005) M. Smiljanić, M. v. Keulen et W. Jonker. "Formalizing the XML schema matching problem as a constraint optimization problem". In: *International Conference on Database and Expert Systems Applications*. Springer. 2005, p. 333-342 (cf. p. 52).

(Sorensen, 1948) T. A. Sorensen. “A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons”. In: *Biol. Skar.* 5 (1948), p. 1-34 (cf. p. 72).

(Sorrentino et al., 2009) S. Sorrentino, S. Bergamaschi, M. Gawinecki et L. Po. “Schema normalization for improving schema matching”. In: *International Conference on Conceptual Modeling*. Springer. 2009, p. 280-293 (cf. p. 51, 114).

(Speer, Chin et Havasi, 2017) R. Speer, J. Chin et C. Havasi. “Conceptnet 5.5: An open multilingual graph of general knowledge”. In: *Thirty-first AAAI conference on artificial intelligence*. 2017 (cf. p. 74).

(Stroetmann et al., 2006) K. A. Stroetmann, T. Jones, A. Dobrev et V. N. Stroetmann. “eHealth is Worth it”. In: *The economic benefits of implemented eHealth solutions at ten European sites* (2006) (cf. p. 16).

(Struijk et al., 2022) M. Struijk, C. X. Ou, R. M. Davison et S. Angelopoulos. *Putting the IS back into IS research*. 2022 (cf. p. 10).

(Studer, Benjamins et Fensel, 1998) R. Studer, V. R. Benjamins et D. Fensel. “Knowledge engineering: principles and methods”. In: *Data & knowledge engineering* 25.1-2 (1998), p. 161-197 (cf. p. 29).

(Subaeki et al., 2019) B. Subaeki, A. Rahman, S. Putra et C. Alam. “Success model for measuring information system implementation: Literature review”. In: *Journal of Physics: Conference Series*. T. 1402. 7. IOP Publishing. 2019, p. 077015 (cf. p. 10).

(Suchanek, Abiteboul et Senellart, 2011) F. M. Suchanek, S. Abiteboul et P. Senellart. “Paris: Probabilistic alignment of relations, instances, and schema”. In: *arXiv preprint arXiv:1111.7164* (2011) (cf. p. 44).

(Suh, 2021) H. Suh. “An integrative model for information system success: Explaining is success by strategic alignment synthesis is investments and is maturity”. In: (2021) (cf. p. 7).

(Suleman, 2002) H. Suleman. *Open digital libraries*. Virginia Polytechnic Institute et State University, 2002 (cf. p. 16).

(Sun, Zhou et Fei, 2020) H. Sun, W. Zhou et M. Fei. “A survey on graph matching in computer vision”. In: *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2020, p. 225-230 (cf. p. 63).

(Sutanta et al., 2016) E. Sutanta, R. Wardoyo, K. Mustofa et E. Winarko. “Survey: Models and Prototypes of Schema Matching.” In: *International Journal of Electrical & Computer Engineering (2088-8708)* 6.3 (2016) (cf. p. 48).

(Symons, 1991) V. Symons. “Impacts of information systems: four perspectives”. In: *Information and Software Technology* 33.3 (1991), p. 181-190 (cf. p. 7).

(2015). *ISO 9001:* Standard. Geneva, CH: International Organization for Standardization, 2015 (cf. p. 7).

(Szejka et al., 2014) A. L. Szejka, A. Aubry, H. Panetto, O. C. Júnior et E. R. Loures. “Towards a conceptual framework for requirements interoperability in complex systems engineering”. In: *OTM Confederated International Conferences” On the Move to Meaningful Internet Systems”*. Springer. 2014, p. 229-240 (cf. p. 21).

(Taghavi et al., 2018) M. Taghavi, J. Bentahar, K. Bakhtiyari et C. Hanachi. “New insights towards developing recommender systems”. In: *The computer journal* 61.3 (2018), p. 319-348 (cf. p. 44).

(Tayur et Suchithra, 2017) V. M. Tayur et R. Suchithra. “Review of interoperability approaches in application layer of Internet of Things”. In: *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE. 2017, p. 322-326 (cf. p. 21).

(Tchoffa et al., 2021) D. Tchoffa, N. Figay, P. Ghodous, H. Panetto et A. El Mhamedi. "Alignment of the product lifecycle management federated interoperability framework with internet of things and virtual manufacturing". In: *Computers in Industry* 130 (2021), p. 103466 (cf. p. 19).

(Thiéblin et al., 2020) E. Thiéblin, O. Haemmerlé, N. Hernandez et C. Trojahn. "Survey on complex ontology matching". In: *Semantic Web* 11.4 (2020), p. 689-727 (cf. p. 48).

(Thomas et al., 2022) J. M. Thomas, A. Moallemy-Oureh, S. Beddar-Wiesing et C. Holzhüter. "Graph Neural Networks Designed for Different Graph Types: A Survey". In: *arXiv preprint arXiv:2204.03080* (2022) (cf. p. 114).

(Thuy, Y.-K. Lee et S. Lee, 2013) P. T. T. Thuy, Y.-K. Lee et S. Lee. "Semantic and structural similarities between XML Schemas for integration of ubiquitous healthcare data". In: *Personal and ubiquitous computing* 17.7 (2013), p. 1331-1339 (cf. p. 76).

(Tolk et Muguira, 2003) A. Tolk et J. A. Muguira. "The levels of conceptual interoperability model". In: *Proceedings of the 2003 fall simulation interoperability workshop*. T. 7. Citeseer. 2003, p. 1-11 (cf. p. 18, 19).

(Toosi, Calheiros et Buyya, 2014) A. N. Toosi, R. N. Calheiros et R. Buyya. "Interconnected cloud computing environments: Challenges, taxonomy, and survey". In: *ACM Computing Surveys (CSUR)* 47.1 (2014), p. 1-47 (cf. p. 30).

(Touzi et al., 2009) J. Touzi, F. Benaben, H. Pingaud et J. P. Lorré. "A model-driven approach for collaborative service-oriented architecture design". In: *International journal of production economics* 121.1 (2009), p. 5-20 (cf. p. 3).

(Truptil et al., 2008) S. Truptil, F. Bénaben, P. Couget, M. Lauras, V. Chapurlat et H. Pingaud. "Interoperability of information systems in crisis management: Crisis modeling and metamodeling". In: *Enterprise Interoperability III*. Springer, 2008, p. 583-594 (cf. p. 3).

(Tu, Zacharewicz et D. Chen, 2016) Z. Tu, G. Zacharewicz et D. Chen. "A federated approach to develop enterprise interoperability". In: *Journal of Intelligent Manufacturing* 27.1 (2016), p. 11-31 (cf. p. 14, 28).

(Tversky, 1977) A. Tversky. "Features of similarity." In: *Psychological review* 84.4 (1977), p. 327 (cf. p. 72).

(Ullberg, D. Chen et P. Johnson, 2009) J. Ullberg, D. Chen et P. Johnson. "Barriers to enterprise interoperability". In: *IFIP-International Workshop on Enterprise Interoperability*. Springer. 2009, p. 13-24 (cf. p. 20).

(Unal et Afsarmanesh, 2010) O. Unal et H. Afsarmanesh. "Semi-automated schema integration with SASMINT". In: *Knowledge and information systems* 23.1 (2010), p. 99-128 (cf. p. 50).

(Usov et al., 2010) A. Usov, C. Beyel, E. Rome, U. Beyer, E. Castorini, P. Palazzari et A. Tofani. "The DIESIS approach to semantically interoperable federated critical infrastructure simulation". In: *2010 Second International Conference on Advances in System Simulation*. IEEE. 2010, p. 121-128 (cf. p. 28, 29).

(Valle, Garcés et Nakagawa, 2021) P. H. D. Valle, L. Garcés et E. Y. Nakagawa. "Architectural strategies for interoperability of software-intensive systems: practitioners' perspective". In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 2021, p. 1399-1408 (cf. p. 15).

(Van Overeem, Witters et Peristeras, 2007) A. Van Overeem, J. Witters et V. Peristeras. "An interoperability framework for Pan-European e-government services (PEGS)". In: *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE. 2007, p. 7-7 (cf. p. 17, 19).

(Van Rossum et Drake, 2009) G. Van Rossum et F. L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009 (cf. p. 86).

(F. Vernadat, 2006) F. Vernadat. “Interoperable enterprise systems: architectures and methods”. In: *IFAC Proceedings Volumes* 39.3 (2006), p. 13-20 (cf. p. 17).

(F. Vernadat, 1996) F. Vernadat. *Enterprise modeling and integration*. Boom Koninklijke Uitgevers, 1996 (cf. p. 13).

(F. B. Vernadat, 2009) F. B. Vernadat. “Enterprise integration and interoperability”. In: *Springer handbook of automation*. Springer, 2009, p. 1529-1538 (cf. p. 14, 17).

(F. B. Vernadat, 2010) F. B. Vernadat. “Technical, semantic and organizational issues of enterprise interoperability and networking”. In: *Annual Reviews in Control* 34.1 (2010), p. 139-144 (cf. p. 17).

(Vijaymeena et Kavitha, 2016) M. Vijaymeena et K. Kavitha. “A survey on similarity measures in text mining”. In: *Machine Learning and Applications: An International Journal* 3.2 (2016), p. 19-28 (cf. p. 72).

(Voigt, 2011) K. Voigt. “Structural graph-based metamodel matching”. Thèse de doct. Dresden, Technische Universität Dresden, Diss., 2011, 2011 (cf. p. 44, 50).

(Voigt, Ivanov et Rummler, 2010) K. Voigt, P. Ivanov et A. Rummler. “Matchbox: combined meta-model matching for semi-automatic mapping generation”. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. 2010, p. 2281-2288 (cf. p. 52, 114).

(Wanden-Berghe et Sanz-Valero, 2012) C. Wanden-Berghe et J. Sanz-Valero. “Systematic reviews in nutrition: standardized methodology”. In: *British journal of nutrition* 107.S2 (2012), S3-S7 (cf. p. 118).

(B. Wang et al., 2019) B. Wang, R. Shin, X. Liu, O. Polozov et M. Richardson. “Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers”. In: *arXiv preprint arXiv:1911.04942* (2019) (cf. p. 44).

(J. Wang, B. Guo et L. Chen, 2022) J. Wang, B. Guo et L. Chen. “Human-in-the-loop Machine Learning: A Macro-Micro Perspective”. In: *arXiv preprint arXiv:2202.10564* (2022) (cf. p. 52).

(J. Wang et Y. Dong, 2020) J. Wang et Y. Dong. “Measurement of text similarity: a survey”. In: *Information* 11.9 (2020), p. 421 (cf. p. 43, 65, 71).

(J. Wang et al., 2004) J. Wang, J.-R. Wen, F. Lochovsky et W.-Y. Ma. “Instance-based schema matching for web databases by domain-specific query probing”. In: *VLDB*. T. 4. 2004, p. 408-419 (cf. p. 52).

(L. Wang, Egorova et Mokryakov, 2018) L. Wang, E. Egorova et A. Mokryakov. “Development of hypergraph theory”. In: *Journal of Computer and Systems Sciences International* 57.1 (2018), p. 109-114 (cf. p. 61).

(N. Wang et al., 2016) N. Wang, H. Liang, Y. Jia, S. Ge, Y. Xue et Z. Wang. “Cloud computing research in the IS discipline: A citation/co-citation analysis”. In: *Decision Support Systems* 86 (2016), p. 35-47 (cf. p. 10).

(Q. Wang et X. Wen, 2015) Q. Wang et X. Wen. “Propagating Dependencies under Schema Mappings: A Graph-based Approach”. In: *Proceedings of the 19th International Database Engineering & Applications Symposium*. 2015, p. 126-135 (cf. p. 50).

(T. Wang, Truptil et Benaben, 2017) T. Wang, S. Truptil et F. Benaben. “An automatic model-to-model mapping and transformation methodology to serve model-based systems engineering”. In: *Information Systems and e-Business Management* 15.2 (2017), p. 323-376 (cf. p. 4).

(Y. Wang et al., 2020) Y. Wang, Y. Hou, W. Che et T. Liu. “From static to dynamic word representations: a survey”. In: *International Journal of Machine Learning and Cybernetics* 11.7 (2020), p. 1611-1630 (cf. p. 44).

(Weichhart, Guédria et Naudet, 2016) G. Weichhart, W. Guédria et Y. Naudet. “Supporting interoperability in complex adaptive enterprise systems: A domain specific language approach”. In: *Data & Knowledge Engineering* 105 (2016), p. 90-106 (cf. p. 13, 28-30).

(Weichhart et Naudet, 2014) G. Weichhart et Y. Naudet. "Ontology of enterprise interoperability extended for complex adaptive systems". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2014, p. 219-228 (cf. p. 13).

(Weichhart et Stary, 2015) G. Weichhart et C. Stary. "A domain specific language for organisational interoperability". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2015, p. 117-126 (cf. p. 13, 28).

(Weichhart et Stary, 2017) G. Weichhart et C. Stary. "Interoperable process design in production systems". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2017, p. 26-35 (cf. p. 27).

(W. Wen et al., 2020) W. Wen, D. D. Zeng, J. Bai, K. Zhao et Z. Li. "Learning Embeddings Based on Global Structural Similarity in Heterogeneous Networks". In: *IEEE Intelligent Systems* 36.6 (2020), p. 13-22 (cf. p. 71).

(Wimmer, Boneva et Di Giacomo, 2018) M. A. Wimmer, R. Boneva et D. Di Giacomo. "Interoperability governance: a definition and insights from case studies in Europe". In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. 2018, p. 1-11 (cf. p. 17).

(Winkler, 1999) W. E. Winkler. "The state of record linkage and current research problems". In: *Statistical Research Division, US Census Bureau*. Citeseer. 1999 (cf. p. 71).

(Z. Wu et Palmer, 1994) Z. Wu et M. Palmer. "Verb semantics and lexical selection". In: *arXiv preprint cmp-lg/9406033* (1994) (cf. p. 74).

(Xiao et Watson, 2019) Y. Xiao et M. Watson. "Guidance on conducting a systematic literature review". In: *Journal of Planning Education and Research* 39.1 (2019), p. 93-112 (cf. p. 117).

(H. Xu et al., 2020) H. Xu, L. Xiang, Y. Le, X. Gan, Y. Jia, L. Fu et X. Wang. "High-order relation construction and mining for graph matching". In: *arXiv preprint arXiv:2010.04348* (2020) (cf. p. 38).

(J. Xu et al., 2021) J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian et F. Wang. "Federated learning for healthcare informatics". In: *Journal of Healthcare Informatics Research* 5.1 (2021), p. 1-19 (cf. p. 30).

(L. D. Xu, E. L. Xu et L. Li, 2018) L. D. Xu, E. L. Xu et L. Li. "Industry 4.0: state of the art and future trends". In: *International journal of production research* 56.8 (2018), p. 2941-2962 (cf. p. 11).

(X. Xu et al., 2021) X. Xu, Y. Lu, B. Vogel-Heuser et L. Wang. "Industry 4.0 and Industry 5.0—Inception, conception and perception". In: *Journal of Manufacturing Systems* 61 (2021), p. 530-535 (cf. p. 11).

(Yan et al., 2016) J. Yan, X.-C. Yin, W. Lin, C. Deng, H. Zha et X. Yang. "A short survey of recent advances in graph matching". In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. 2016, p. 167-174 (cf. p. 63).

(Q. Yang et al., 2019) Q. Yang, Y. Liu, T. Chen et Y. Tong. "Federated machine learning: Concept and applications". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019), p. 1-19 (cf. p. 30).

(Y. Yang et al., 2019) Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung et al. "Multilingual universal sentence encoder for semantic retrieval". In: *arXiv preprint arXiv:1907.04307* (2019) (cf. p. 74).

(Yin, Zhu et J. Hu, 2021) X. Yin, Y. Zhu et J. Hu. "A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), p. 1-36 (cf. p. 30).

(J. Youssef, 2017) J. Youssef. "Developing an enterprise operating system for the monitoring and control of enterprise operations". Thèse de doct. Université de Bordeaux, 2017 (cf. p. 27).

(J. R. Youssef et al., 2016) J. R. Youssef, D. Chen, G. Zacharewicz et T. Zhiying. "Developing and Enterprise Operating System (EOS) with the Federated Interoperability Approach". In: *I3M Conference*. 2016 (cf. p. 28).

(J. R. Youssef et al., 2018a) J. R. Youssef, G. Zacharewicz, D. Chen et Z. Tu. "Enterprise operating system framework: federated interoperability based on HLA." In: *Int. J. Simul. Process. Model.* 13.4 (2018), p. 337-354 (cf. p. 28).

(J. R. Youssef et al., 2018b) J. R. Youssef, G. Zacharewicz, D. Chen et F. Vernadat. "EOS: enterprise operating systems". In: *International Journal of Production Research* 56.8 (2018), p. 2714-2732 (cf. p. 28).

(Yu et Jagadish, 2006) C. Yu et H. Jagadish. "Schema summarization". In: *Proceedings of the 32nd international conference on Very large data bases*. Citeseer. 2006, p. 319-330 (cf. p. 47).

(Zacharewicz, D. Chen et Vallespir, 2009) G. Zacharewicz, D. Chen et B. Vallespir. "Short-lived ontology approach for agent/HLA federated enterprise interoperability". In: *2009 International Conference on Interoperability for Enterprise Software and Applications China*. IEEE. 2009, p. 329-335 (cf. p. 28, 29).

(Zacharewicz et al., 2020) G. Zacharewicz, N. Daclin, G. Doumeingts et H. Haidar. "Model driven interoperability for system engineering". In: *Modelling* 1.2 (2020), p. 94-121 (cf. p. 2, 27).

(Zacharewicz et al., 2017) G. Zacharewicz, S. Diallo, Y. Ducq, C. Agostinho, R. Jardim-Goncalves, H. Bazoun, Z. Wang et G. Doumeingts. "Model-based approaches for interoperability of next generation enterprise information systems: state of the art and future challenges". In: *Information Systems and e-Business Management* 15.2 (2017), p. 229-256 (cf. p. 12-14, 23, 25).

(Zacharewicz et al., 2009) G. Zacharewicz, O. Labarthe, D. Chen et B. Vallespir. "HLA Multi Agent/Short-Lived Ontology Platform for Enterprise Interoperability". In: *IFIP International Conference on Advances in Production Management Systems*. Springer. 2009, p. 350-357 (cf. p. 28, 30).

(Zaim, Muhammed et Tarim, 2019) H. Zaim, S. Muhammed et M. Tarim. "Relationship between knowledge management processes and performance: critical role of knowledge utilization in organizations". In: *Knowledge Management Research & Practice* 17.1 (2019), p. 24-38 (cf. p. 9).

(Zass et Shashua, 2008) R. Zass et A. Shashua. "Probabilistic graph and hypergraph matching". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, p. 1-8 (cf. p. 60).

(Zdravković et al., 2017) M. Zdravković, F. Luis-Ferreira, R. Jardim-Goncalves et M. Trajanović. "On the formal definition of the systems' interoperability capability: an anthropomorphic approach". In: *Enterprise Information Systems* 11.3 (2017), p. 389-413 (cf. p. 13).

(Zdravković et al., 2015) M. Zdravković, O. Noran, H. Panetto et M. Trajanović. "Enabling interoperability as a property of ubiquitous systems for disaster management". In: *Computer Science and Information Systems* 12.3 (2015), p. 1009-1031 (cf. p. 21, 26).

(Zeb et al., 2022) S. Zeb, A. Mahmood, S. A. Khowaja, K. Dev, S. A. Hassan, N. M. F. Qureshi, M. Gidlund et P. Bellavista. "Industry 5.0 is Coming: A Survey on Intelligent NextG Wireless Networks as Technological Enablers". In: *arXiv preprint arXiv:2205.09084* (2022) (cf. p. 12).

(C. J. Zhang et al., 2013) C. J. Zhang, L. Chen, H. V. Jagadish et C. C. Cao. "Reducing uncertainty of schema matching via crowdsourcing". In: *Proceedings of the VLDB Endowment* 6.9 (2013), p. 757-768 (cf. p. 52).

(P. Zhang, Portillo et Kezunovic, 2006) P. Zhang, L. Portillo et M. Kezunovic. "Compatibility and interoperability evaluation for all-digital protection system through automatic application test". In: *2006 IEEE Power Engineering Society General Meeting*. IEEE. 2006, 7-pp (cf. p. 17).

(X. Zhang, Mao et Cambria, 2022) X. Zhang, R. Mao et E. Cambria. “A survey on syntactic processing techniques”. In: *Artificial Intelligence Review* (2022), p. 1-84 (cf. p. 71).

(Z. Zhang et al., 2008) Z. Zhang, P. Shi, H. Che et J. Gu. “An algebraic framework for schema matching”. In: *Informatica* 19.3 (2008), p. 421-446 (cf. p. 44).

(Z. Zhang, C. Wu et Cheung, 2013) Z. Zhang, C. Wu et D. W. Cheung. “A survey on cloud interoperability: taxonomies, standards, and practice”. In: *ACM SIGMETRICS Performance Evaluation Review* 40.4 (2013), p. 13-22 (cf. p. 18).

(B. Zhao, 2017) B. Zhao. “Web scraping”. In: *Encyclopedia of big data* (2017), p. 1-3 (cf. p. 43).

Table des matières

RÉSUMÉ

Implémentation d'une interopérabilité fédérée supportée par la transformation automatisée à la volée de modèles de données hétérogènes : application aux problèmes d'appariement des schémas

Aujourd'hui, le monde industriel connaît des avancées des technologies de l'information et de la communication sous l'ombrelle de l'industrie 4.0 et 5.0. Les organisations sont à la recherche continue d'adaptabilité et de flexibilité dans un contexte compétitif et dynamique. Ces technologies sont souvent sollicitées par ces organisations pour améliorer leurs performances et leurs capacités à atteindre leurs objectifs. L'adoption de ces technologies n'est pas uniforme et entraîne une augmentation de la quantité de données hétérogènes. L'interopérabilité est alors cruciale, car elle permet l'échange et l'utilisation de ces données. Les travaux menés dans le cadre de cette thèse s'intéressent à la mise en œuvre d'une approche générique et adaptable sur les principes d'une interopérabilité fédérée entre des bases de données. L'approche fédérée n'impose aucun modèle ou format de données. L'interopérabilité est mise en œuvre à la volée et est la plus à même à répondre aux exigences d'adaptabilité, de flexibilité dans des environnements dynamiques. La thèse présente et analyse l'état de l'art actuel sur la notion de l'interopérabilité. La transposition de cette analyse a permis de proposer un cadre général qui rassemble la littérature et décrit les différents concepts de l'interopérabilité. Ce cadre a motivé le développement d'une approche fédérée qui vise à exploiter au maximum les informations disponibles. Pour cela, la thèse s'inspire des techniques issues du traitement automatique des langues, l'optimisation et la théorie des graphes. L'approche a été appliquée sur deux cas d'étude pour des problèmes d'appariement des schémas.

MOTS-CLÉS : Interopérabilité, Théorie des Graphes, Traitement Automatique des Langues, Optimisation.

ABSTRACT

Implementation of a federate interoperability supported by on-the-fly heterogeneous models transformation : application to schema matching problems

Today, the industrial world is experiencing advances in information and communication technologies under the umbrella of Industry 4.0 and 5.0. Organisations are continuously looking for adaptability and flexibility in a competitive and dynamic context. These technologies are often called upon by these organisations to improve their performance and their ability to achieve their objectives. The adoption of these technologies is not uniform and leads to an increase in the amount of heterogeneous data. Interoperability is therefore crucial, as it enables the exchange and use of this data. Interoperability is an essential mechanism when it comes to software upgrades, technical migrations or connector implementations to ensure communication with third party services, software or companies. The work carried out within the framework of this thesis focuses on the implementation of a generic and adaptable approach based on the principles of federated interoperability between databases. The federated approach does not impose any data model or format. Interoperability is implemented on-the-fly and is most likely to meet the requirements of adaptability and flexibility in dynamic environments. The thesis presents and analyses the current state of the art on the notion of interoperability. The transposition of this analysis has allowed to propose a general framework that gathers the literature and describes the different concepts of interoperability. This framework has motivated the development of a federated approach that aims to exploit the available information to the maximum. To this end, the thesis draws on techniques from natural language processing, optimisation and graph theory. The approach has been applied to two case studies of schema matching problems.

KEYWORDS: Interoperability, Graph Theory, Natural Language Processing, Optimisation.