RÉSUMÉ SUBSTANTIEL EN FRANÇAIS DE LA THÈSE DE DOCTORAT DE L'UNIVERSITÉ JEAN MOULIN LYON 3

Membre de l'université de Lyon

Discipline : **Informatique**

par

Firas ZOUARI

Approche à base de services et agents intelligents pour la recommandation et la gestion de crise :

application à l'analyse et le management des maladies émergentes

Remerciements

Cette thèse est l'aboutissement d'années de travail acharné, de dévouement et de passion, et elle n'aurait pas été possible sans le soutien et les conseils de nombreuses personnes.

Je profite de cette occasion pour exprimer ma gratitude à ma directrice de thèse, Mme Chirine GHEDIRA GUEGAN, pour son précieux encadrement, son expertise et ses encouragements, mais aussi pour m'avoir fait confiance tout au long de cette thèse. Je tiens à remercier et à souligner les contributions bénéfiques et les conseils de Mme Nadia KABACHI. Je remercie également Madame Khouloud BOUKADI pour sa disponibilité, ses contributions et ses conseils qui m'ont permis de réaliser ce travail.

Je tiens à remercier les membres du jury qui ont évalué mon travail : Mme Nada MATTA et M. Mahdi ZARGAYOUNA pour la relecture de ma thèse, ainsi que M. Elhadj BENKHALIFA, M. Richard CHBEIR, M. Claude DUSSART, et M. Hedi KARRAY pour leur participation au jury et l'intérêt qu'ils ont porté à mon travail.

Je tiens à remercier tout particulièrement mes parents, ma femme, ma sœur, et ma belle-famille pour leur soutien, leur amour et leurs encouragements indéfectibles tout au long de mon parcours universitaire. Je tiens à remercier tous mes amis, qui m'ont soutenu d'innombrables façons.

Je tiens à exprimer une grande pensée à mes grands-parents auxquels j'adresse tout mon amour. Je réserve une pensée particulière à mon grand-père. Que leurs âmes reposent en paix !

Je tiens à remercier tous les chercheurs, enseignants, collègues, amis et personnel du laboratoire LIRIS pour les bons moments passés ensemble.

Résumé

Nous assistons aujourd'hui à une augmentation de plus en plus importante de nombre d'échanges et de flux migratoires. Ces échanges et les catastrophes naturelles sont considérés parmi les facteurs les plus influents sur la propagation et l'émergence de maladies infectieuses. Ce fait est affirmé par la récente pandémie de COVID-19, qui a provoqué une crise sanitaire critique à l'échelle mondiale. Dans ce contexte, les multiples sources de données notamment les données ouvertes, issues de réseaux sociaux, des données des patients et d'IoT jouent un rôle crucial pour la génération desdites données liées à la santé et leur analyse. Ces données sont caractérisées par un aspect très dynamique, hétérogène, la complexes ayant un facteur de croissance élevé. Ces caractéristiques peuvent avoir un impact sur leurs utilités et handicaper le processus d'analyse particulièrement dans les systèmes de gestion des crises sanitaires qui font l'objet de la présente thèse. Malgré les importants progrès technologiques, les systèmes actuels de gestion de crises sanitaires ne sont pas encore capables de traiter cette masse de données en toute autonomie et intelligence véritable, comme ils doivent toujours faire recours à des situations prévisibles et préprogrammées pour générer des recommandations. Par ailleurs, les utilisateurs utilisent souvent ces systèmes de gestion de crises dans différentes situations chaotiques qui impliquent plusieurs contraintes, entre autres le temps restreint pour prendre des décisions efficaces. Par conséquent, les préférences et les exigences des utilisateurs envers la qualité des données et les recommandations souhaitées peuvent être très variables en fonction des rôles des utilisateurs et du contexte de décision. Ainsi, le défi de la présente thèse est de répondre au problème suivant : "Comment générer des recommandations de manière intelligente et autonome sur des données multi-sources, hétérogènes, incertaines et complexes, regroupées dans un lac de données sans avoir connaissance préalable ?".

De ce fait, nous avons identifié deux sous-problèmes concernant les systèmes de recommandation prenant compte des besoins d'une multitude des utilisateurs dans différents contextes. Plus précisément, nous nous sommes concentrés sur les sous-problèmes sous-jacents, à savoir (1) "Comment assurer la gestion de données hétérogènes, et plus spécifiquement, la curation de données collectées en batch et en streaming d'une manière adaptative en considérant les besoins fonctionnels et non fonctionnels de l'utilisateur ?" et (2) "Comment recommander des mesures de santé préventives tout en proposant des explications adaptées aux rôles des utilisateurs dans différents contextes de décision ?". Ainsi, notre objectif principal est de proposer une approche intégrant un système intelligent pour recommander les mesures de santé préventives appropriées en fonction des besoins de l'utilisateur via l'analyse de données provenant de sources multiples. Pour ce faire, nous avons proposé des contributions abordant chaque étape impliquée dans la recommandation des mesures sanitaires. Premièrement, nous avons proposé une approche de composition de services dea curation des données adaptative dans les data lakehouses en tenant compte du rôle de l'utilisateur, de ses préférences, des contraintes et du contexte de décision. En effet, nous nous sommes appuyés sur les data lakehouses comme une solution pratique pour surmonter les défis de l'intégration des données massives. Nous avons donc tiré profit des technologies sémantiques et d'apprentissage par renforcement pour constituer un framework multicouche pour ladite curation des données. Deuxièmement, nous nous sommes concentrés sur les problèmes de prédiction de maladies et de recommandation de mesures de santé en proposant une approche basée sur les technologies sémantiques pour la recommandation de mesures de santé explicables adaptées à de multiples

utilisateurs ayant des besoins différents. Les contributions présentées sont mises en œuvre et expérimentées sur des scénarios du domaine médical.

Abstract

Today, we are witnessing an ever-increasing number of exchanges and migration flows of exchanges and migratory flows. These exchanges and natural disasters are among the most influential factors in spreading infectious diseases. This fact could be affirmed by the recent pandemic of COVID-19, which has caused an acute health crisis worldwide. In this context, we distinguish several sources that are crucial in the generation of health-related data, including open data, social networks, patient data, and IoTs. These data are characterized by a very dynamic aspect, heterogeneity, complexity, and a high growth factor. These characteristics may impact the data usefulness and handicap the data analysis process, especially in health crisis management systems which are the focus of the present thesis. Further, despite the immense technological advances, current health crisis systems cannot still treat such massive data with genuine autonomy and intelligence since they still need to check predictable and pre-programmed situations to generate outcomes. In addition, the users of such systems may use them in different chaotic situations that imply several constraints, like restricting time to make decisions. Accordingly, they may have changing preferences and requirements regarding the data quality and the desired recommendations according to their user roles and decision context. Thus, the challenge of the present thesis is to answer the following problem. "How to generate recommendations intelligently and autonomously on multi-source, heterogeneous, uncertain, and complex data gathered in a data lake without prior knowledge?"

For this purpose, we identified two sub-problems about the recommendation systems considering different users' needs in different contexts. More precisely, we focused on addressing the underlying sub-problems, namely (1) "How to ensure the management of heterogeneous data, and more specifically, the curation of data adaptively collected in batch and streaming while considering the functional and non-functional needs of the user?" and (2) "How to recommend preventive health measures while providing explanations adapted to user roles in different decision contexts?". Therefore, our main objective is to propose an approach integrating an intelligent system to recommend the appropriate preventive health measures according to the user requirements via analyzing data from multi-sources. Hence, we proposed contributions addressing each step involved in the prediction and recommendation to tackle our main objective. First, we proposed a service-based approach for adaptive data curation in data lakehouses by considering the user role, preferences, constraints, and decision context. Indeed, we relied on data lakehouses as a practical solution to overcome the big data integration challenges. Hence, we took advantage of semantic technologies and reinforcement learning techniques to constitute a multi-layered framework for data curation. Subsequently, we focus on disease prediction and health measures recommendation problems by proposing a semantic-based approach for explainable health measures recommendations adapted for multiple users with different needs. The presented contributions are implemented and experimented on medical domain scenarios.

Table des matières

Glossaire

AI Artificial Intelligence
API Application Program Interface
ACUSEC Adaptive CUration SErvice Composition
CNN Convolutional Neural Network
COVID-19 COronaVIrus Disease appeared in 2019
CSV Comma-separated values
DAG Directed Acyclic Graph
DaQ Dataset Quality Ontology
DCat Data Catalog Vocabulary
DQV Data Quality Vocabulary
DUV Data Usage Vocabulary
ELT Extract-Load-Transform
ETL Extract-Transform-Load
GRU Gated Recurrent Unit
HDO Human Disease Ontology
IEEE Institute of Electrical and Electronics Engineers
ICWS International Conference on Web Services
IDEAS International Database Engineering and Applications Symposium
JDBC Java Database Connectivity
JSON JavaScript Object Notation
KNN K Nearest Neighbor
LSTM Long Short-Term Memory
MAE Mean Absolute Error
MAS Multi-Agent System
MDP Markov Decision Process
MERS Middle East Respiratory Syndrome
ML Machine Learning
NLP Natural Language Processing
NMF/NNMF Non-Negative Matrix Factorization

OxCGRT Oxford Covid-19 Government Response Tracker
PROV-O Provenance Ontology
POS Part Of Speech
QoS Quality of Service
REST REpresentational State Transfer
RDBMS Relational Database Management System
RMSE Root-Mean-Square Error
SARS Severe Acute Respiratory Syndrome
SHAP SHapley Additive exPlanations
SOA Service Oriented Architecture
SOAP Simple Object Access Protocol
SQL Structred Query Language
SSN Semantic Sensors Network
SVD Singular Value Decomposition
SWRL Semantic Web Rule Language
URL Uniform Resource Locator
Ux User Experience
URL Uniform Resource Locator
XAI Explainable Artificial Intelligence
XML Extensible Markup Language
W3C World Wide Web Consortium
WSC Web Service Composition
WISE Web Information Systems Engineering

Table des figures

Liste des tableaux

CHAPITRE 1

Introduction

Contexte

Nous assistons depuis des décennies à une croissance positive des échanges économiques et des flux de migration accélérée par la mondialisation. Dans le même temps, les migrations humaines ajoutées aux phénomènes naturelles ont été l'origine de maintes crises que ce soient commerciales, politiques, sociétales ou encore sanitaires suite à l'émergence de nouvelles maladies et la propagation d'infections dans différentes zones géographiques. Pour faire face à ce genre de situations dans certains cas graves, que les efforts se sont orientés vers des solutions innovantes et efficaces pour faciliter la gestion de crise, la détection de phénomènes anormaux et notamment dans le domaine médical suite à des maladies émergentes et des épidémies, tel qu'a été le cas avec la pandémie de COVID-19. Par définition, un système de gestion de crise est une approche structurée de la gestion et de la réponse à des situations d'urgence. Il s'agit d'un ensemble de politiques, de procédures et de protocoles mis en place pour garantir qu'une organisation puisse répondre efficacement à tout type de crise, comme les catastrophes naturelles, les cyber-attaques ou les urgences de santé publique [2]. Un tel système collecte les données de différentes sources entre autres les réseaux sociaux, les objets connectés, les bases institutionnelles, qui jouent un rôle très important dans la génération des données liées à un domaine, notamment la santé dans le cas de la présente thèse. Ces données sont souvent très dynamiques et disposent d'un facteur de croissance très élevé, ce qui rend les infrastructures existantes incapables de traiter une volumétrie énorme. Ainsi, différents types de référentiels de données, comme les lacs de données (Data Lakes) et les Data Lakehouses, ont été proposés pour mieux exploiter des masses de données qui peuvent être collectées en batch et en streaming. Ces référentiels de données facilitent grandement l'intégration desdites données, qu'elles soient structurées, semistructurées ou non-structurées. Par ailleurs, les systèmes de gestion de crise actuels manquent encore d'autonomie et d'intelligence réelles puisqu'ils doivent souvent avoir recours à des situations prévisibles et préprogrammées.

Tenant compte du contexte présenté, notre objectif est de proposer une approche générique permettant d'aider à gérer les crises en s'appuyant sur des données multi-sources.

Pour mieux illustrer le contexte et les verrous du présent travail, nous présentons les deux scénarii suivants : Supposons qu'un expert en stratégie publique de santé souhaite agir suite à une propagande lancée sur Twitter autour d'un nouveau virus. Ainsi, il doit faire appel à un système qui permette la collecte des données à travers différentes sources de données médicales et des réseaux sociaux, entre autres Twitter, Facebook, etc. Ce système doit assurer la vérification de la pertinence des nouvelles informations publiées sur ces derniers et la génération des actions à prendre suite à une analyse desdites nouvelles données. Dans un scenario similaire, un professionnel de santé souhaite consulter des recommandations médicales liées à une maladie inconnue caractérisée par un ensemble de symptômes qui ressemblent à la grippe. Sachant que plusieurs pays sont en situation de crise sanitaire grave, ce dernier veut consulter les dernières recommandations à prendre en compte pour faire face à cette maladie. Dans ce cas, un système permettant l'analyse des bases institutionnelles, des données issues des objets connectés et des dossiers patients sera utile pour générer essentiellement des recommandations médicales afin de gérer la situation de crise. En plus de la génération de recommandations, un tel système doit fournir des explications adaptées pour chaque rôle d'utilisateur. Par exemple, l'expert stratégique souhaite des explications sous forme des statistiques et des arguments (comme la situation dans les pays voisins, des exemples concrets) pour comprendre le choix des recommandations. Par ailleurs, le professionnel de santé sera peut-être intéressé par des explications sous forme d'informations médicales pour mieux comprendre la situation. Sur la base de ces deux scenarii, deux principales questions se posent, à savoir (1) Comment un système peut s'adapter et adapter les actions à la finalité de plusieurs utilisateurs ainsi qu'au contexte global ? (2) Comment gérer et préparer les différentes données collectées en batch et en streaming ?

Le reste de ce chapitre est consacré à l'examen de la thèse sous plusieurs angles, notamment les questions de recherche, les contributions et l'organisation de la thèse.

Problématique de recherche

Dans ce contexte, la problématique de recherche que nous abordons dans cette thèse est la suivante :

> Comment générer des recommandations de manière intelligente et autonome sur des données multi-sources, hétérogènes, incertaines et complexes collectées dans un data lakehouse sans avoir de connaissances préalables ?

Nous décomposons ce problème de recherche en trois sous-problèmes.

RP1 - Identifier les étapes de la gestion des données et évaluer et comparer les approches existantes pour chaque étape.

L'explosion quantitative des données numériques a engendré nouvelles façons de voir et d'analyser les données en découvrant de nouveaux moyens de les capturer, de les récupérer, de les partager, de les stocker et de les présenter. Dans ce contexte, les lacs de données et les entrepôts de données ont été proposés comme un dépôt de données à faible coût qui surmonte les limites des dépôts de données classiques. Cette technologie émergente se distingue par sa capacité à transporter des données massives sous une forme très hétérogène (c'est-à-dire non structurée, semi-structurée et structurée) en même temps [3]. Cette caractéristique précieuse a augmenté de plus en plus l'utilisation des lacs de données et les data lakehouses, en particulier pour la mise en œuvre d'applications en temps réel lorsqu'il y a une contrainte de temps empêchant d'effectuer un processus d'unification des schémas de données avant de charger les données dans un entrepôt. Malheureusement, malgré les avantages du data lakehouse, son adoption se heurte à de nombreuses difficultés, et ce pour plusieurs raisons. Tout d'abord, l'ingestion de données provenant de sources multiples soulève de nombreuses questions quant à la qualité des données ingérées. La qualité des données concerne leur "fitness for use", qui consiste à évaluer la qualité des données en fonction de leur contexte d'utilisation. La qualité des données à un moment donnée peut être appropriée pour un cas d'application, mais peut ne pas être suffisante pour un autre. Outre la qualité des données et celle des sources, la préservation de la confidentialité des données contenues dans le data lakehouse est une question soulevée par les professionnels manipulant des données sensibles (par exemple, dans le domaine de la santé, de l'administration, du social, etc.). Pour relever ces défis, les scientifiques ont trouvé des mécanismes pour curer les données, garantir leur qualité et préserver la vie privée.
Ce travail vise à catégoriser les travaux existants et à identifier les questions ouvertes concernant la gestion des données dans les lacs de données et les data lakehouses.
C'est pourquoi nous envisageons les questions de recherche suivantes :

- Quels sont les critères d'évaluation et de comparaison des processus de gestion des données liés aux lacs de données et aux data lakehouses ?
- Quels sont les limites des approches existantes liées à chaque étape de la gestion des données ?

En examinant les travaux existants sur la gestion des données, nous avons identifié les problèmes de recherche connexes décrits dans les sections suivantes.

RP2 - Gestion de données multistructurées collectées en batch et en streaming

Les données se caractérisent souvent par leur hétérogénéité, leur complexité et leur incertitude en raison de leurs différentes formes de structure (par exemple, les données non structurées, les données semi-structurées et les données structurées) et de la multiplicité des origines des données. Compte tenu de la quantité croissante de données, il est de plus en plus nécessaire de mettre en œuvre un processus de gestion des données adapté aux

données volumineuses afin de traiter les problèmes susmentionnés. Par exemple, le secteur médical est confronté à une explosion du volume de données générées par les progrès de la génomique et de la protéomique, combinées aux données de laboratoire, à l'historique des patients, aux données de recherche clinique et à la blogosphère de la santé. De plus, divers outils tentent d'interagir entre eux pour former des plateformes de plus en plus complexes et multiplier l'hétérogénéité des données. En conséquence, le traitement de ces données nécessite à lui seul un travail colossal que les outils classiques de gestion des données ne pourraient pas réaliser [4]. À cette fin, le processus de gestion des données peut contenir plusieurs étapes, telles que l'intégration des données, l'analyse des données et la curation des données. Cette dernière est une étape cruciale qui repose sur la gestion et la promotion de l'utilisation des données à partir de leur point de création, en les enrichissant ou en les mettant à jour pour qu'elles restent adaptées à un objectif spécifique. Par conséquent, la fiabilité du processus décisionnel dépend de la rigueur de chaque étape de la gestion des données, y compris la curation des données. Cependant, cette dernière peut nécessiter trop de temps et impliquer plus d'efforts, dépassant 50 % de l'effort total et du temps de traitement [5, 6, 7]. D'autre part, les contextes décisionnels critiques, tels que pour la gestion de crises, sont généralement évolutifs et peuvent imposer des exigeances notamment en termes de temps d'exécution et de précision des résultats et réponses du système d'information. Il est donc nécessaire d'adapter l'étape de curation des données en fonction de l'évolution des facteurs de décision afin de ne pas nuire aux performances globales du système et de s'aligner sur les attentes des utilisateurs et leurs contextes de décision. Ainsi, le défi consiste à effectuer la curation des données pour plusieurs types de sources de données en même temps, tout en tenant compte des différents besoins des utilisateurs dans des contextes différents.

C'est pourquoi nous considérons les questions de recherche suivantes :

- Comment effectuer la curation de données pour des données multistructurées collectées en batch et en streaming ?
- Comment prendre en compte les besoins des différents utilisateurs dans différents contextes lors de la curation des données ?

Une fois les données collectées et traitées, il faut les exploiter pour pouvoir gérer les crises. C'est pourquoi nous présentons ci-après le deuxième problème scientifique qui porte sur la recommandation des actions à entreprendre pour aider à ladite gestion.

RP3 - Générer des recommandations pour des utilisateurs multiples ayant des besoins différents

L'intelligence artificielle (IA) est devenue un sujet clé dans nos conversations personnelles et professionnelles, dans les débats politiques, les conférences industrielles et les conférences numériques. Par exemple, l'IA est de plus en plus utilisée dans le domaine médical pour différentes finalités, telles que l'analyse des données des patients et des images médicales, la détection des tumeurs et la détection des épidémies. L'IA a donc montré des performances prometteuses dans diverses tâches, proches de celles des experts, voire meilleures. Néanmoins, bien que les modèles d'IA offrent de bonnes performances, cer-

tains modèles, comme les modèles d'apprentissage profond, agissent comme une boîte noire et sont peu interprétables. Or, les experts de certains domaines sensibles, tels que la santé, ont besoin d'explications sur les résultats d'un modèle de recommandation pour mieux comprendre le choix des décisions et attribuer leur confiance. En outre, les utilisateurs peuvent avoir des besoins et des perspectives différents en termes d'explications. Pour mieux illustrer le défi scientifique, nous reprenons l'exemple d'Alice et de Bob des scénarii suscités, qui utilisent un système pour prévoir et gérer les crises sanitaires. Suite à la prédiction d'une menace, ce système recommande des mesures sanitaires pour traiter la maladie prédite. Le système doit donc générer et adapter des explications à chaque rôle d'utilisateur. Alice peut être intéressée par des informations médicales telles que les symptômes, la nature des examens et les traitements. Dans le même temps, Bob peut être intéressé par d'autres explications, telles que des statistiques et la situation dans les pays voisins.
C'est pourquoi nous envisageons la question de recherche suivante :

— Comment recommander des mesures tout en fournissant des explications adaptées aux différents rôles des utilisateurs dans différents contextes décisionnels ?

RP4 - Expérimentation des différentes étapes de gestion de données dans un système de gestion de crise
Les étapes de gestion de données présentées devraient être intégrées dans un système de gestion de crise permettant d'aider les professionnels à prendre des décisions en analysant différentes sources de données afin de recommander les mesures, dans notre cas sanitaires, préventives appropriées. Il est donc essentiel d'analyser les performances de la solution proposée pour vérifier son efficacité pratique et son coût. C'est pourquoi nous considérons les questions de recherche suivantes :

— Les propositions peuvent-elle être mise en œuvre ?
— Quelles sont les performances du système de test en termes de temps de traitement, d'évolutivité et de coût ?
— La proposition présente-t-elle une solution efficace dans la pratique ?

Synthèse des contributions

Cette thèse comprend quatre contributions, chacune abordant l'un des sous-problèmes de recherche sus-mentionnés.

C1 - Un état de l'art de la gestion des données dans le lac de données/data lakehouse en réponse à RP1.

Pour avoir une vision globale du processus de gestion des données, nous avons introduit une cartographie systématique qui couvre les étapes de la gestion des données, telles que la curation, l'évaluation de la qualité, la préservation de la vie privée et la prédiction, en utilisant les données stockées dans le lac de données/data lakehouse. En effet, il est nécessaire d'appliquer une première analyse complète des mécanismes utilisés pour

gérer les données dans un lac de données/data lakehouse et effectuer des recherches en utilisant des données curées tout en assurant la qualité des données et en préservant la vie privée de leurs fournisseurs. Nous avons défini des critères tels que les domaines, le type de proposition et l'étape de gestion des données afin de fournir un schéma de classification des articles. Ensuite, nous avons fourni une analyse des articles examinés. Cette analyse montre les questions ouvertes liées à la curation des données, à l'évaluation de la qualité, à la préservation de la vie privée et à la prédiction sur les données stockées dans les lacs de données.

Cette contribution fait l'objet de la publication suivante dans la conférence internationale, International Database Engineering Applications Symposium (IDEAS 2021) :

— Firas Zouari, Nadia Kabachi, Khouloud Boukadi, Chirine Ghedira Guegan : Data Management in the Data Lake : A Systematic Mapping. IDEAS 2021 : 280-284 [8]

C2 - Une approche adaptative de la curation basée sur les services en réponse à RP2.

Nous avons proposé ensuite une approche pour la curation de données adaptative mise en œuvre sous la forme d'un framework de curation de données pour les sources de données multistructurées en batch et en streaming tout en tenant compte des besoins du décideur. Le framework global proposé est basé sur des services et englobe les étapes de la curation des données (c'est-à-dire la collecte des sources de données, l'évaluation de la qualité des données, la caractérisation des sources de données et la curation des données). Le processus de curation des données englobe la composition des services de curation et les modules de curation des données qui utilisent une bibliothèque de services de curation dans laquelle chaque service présente une tâche de curation. Le framework de curation s'appuie sur ACUSEC, une approche de composition adaptative des services de curation que nous avons proposée, et compose les services de curation de manière adaptative aux caractéristiques du processus de décision afin d'optimiser le processus de curation des données en termes de temps d'exécution et d'alignement avec les besoins de l'utilisateur. Par la suite, le module de curation des données du framework invoque les services de composition.
En ce qui concerne ACUSEC, notre contribution originale s'appuie sur des techniques d'intelligence artificielle, en particulier l'apprentissage automatique, pour générer des schémas de composition de services de curation de données de manière adaptative. À cette fin, notre approche tient compte des exigences fonctionnelles de l'utilisateur, telles que les caractéristiques de la source de données, et des exigences non fonctionnelles, telles que les préférences et les contraintes de l'utilisateur et le contexte de la décision. Principalement, le framework que nous proposons tire profit de l'apprentissage par renforcement en tant que solution pratique capable d'apprendre au fil du temps à prendre des décisions de plus en plus efficaces dans un environnement dynamique tel que celui du data lakehouse.
Cette contribution est la base de la publication suivante dans la Conférence internationale de l'IEEE sur les services Web (ICWS 2021), la Conférence internationale sur l'ingénierie des systèmes d'information Web (WISE 2022) et le journal Clusterning Computing 2023 :

— Firas Zouari, Chirine Ghedira Guegan, Nadia Kabachi, Khouloud Boukadi : To-

wards an adaptive curation service composition based on machine learning. ICWS 2021 : 73-78 [9]

— Firas Zouari, Chirine Ghedira Guegan, Khouloud Boukadi, Nadia Kabachi : A service-based framework for adaptive data curation in data lakehouses WISE 2022 : 225-240 [10]

— Hela Taktak, Khouloud Boukadi, Firas Zouari, Chirine Ghedira-Guegan, Michael Mrissa, et al. A knowledge-driven service composition framework for wildfire prediction. Cluster Computing, 2023.

C3 - Une approche de recommandation des actions à opérer en réponse à la RP3.

Nous avons proposé un modèle basé sur l'apprentissage profond pour la recommandation des actions à entreprendre pour la gestion des crises. Ainsi, nous avons tiré parti des techniques d'apprentissage profond, pour concevoir et proposer un modèle de recommandation par classification multi-sorties dans lequel chaque sortie prédit le niveau de rigueur de chaque action à prendre. Bien que les performances obtenues par les modèles d'apprentissage profond semblent satisfaisantes, ces modèles (modèles d'apprentissage) agissent comme une boîte noire et sont peu interprétables. Nous nous sommes donc attachés à améliorer l'interprétabilité des résultats de nos modèles. À cette fin, notre contribution est dirigée par l'intelligence artificielle explicable (XAI), un paradigme lié à l'explication des modèles type boîte noire tels que l'apprentissage profond. Plus précisément, nous proposons une approche sémantique pour expliquer les modèles boîte noire associés à notre modèle de recommandation.

C4 - Une évaluation par la mise en œuvre d'un système de test en réponse à la RP4.

Nous avons mis en œuvre nos contributions proposées pour constituer un système de gestion de crise à rôles multiples qui analyse des données hétérogènes collectées à partir de diverses sources (par exemple, des capteurs, des bases de données d'instituts de santé, etc.) Plus précisément, ce système utilise un data lakehouse pour stocker les données collectées. Ensuite, il assure la curation des données et les analyse pour recommander les mesures adéquates pour la gestion de crise. En outre, le système mis en œuvre explique le choix des mesures recommandées d'une manière adaptée à chaque rôle d'utilisateur. En conséquence, nous avons examiné l'efficacité des contributions proposées à travers les différents modules du système (c'est-à-dire la curation des données, la prédiction des crises, la recommandation des mesures et l'explication).

CHAPITRE 2

Concepts Fondamentaux & État de l'art

2.1 Introduction

Les maladies infectieuses peuvent se propager rapidement si les mesures préventives ne sont pas prises à temps grâce à l'intervention de ressources humaines compétentes et de ressources matérielles rapidement mobilisables. En effet, la réponse doit être immédiate, ciblée et coordonnée en cas d'épidémie transfrontalière. De même, les menaces chimiques ou les catastrophes environnementales (par exemple, les éruptions volcaniques) peuvent rapidement se propager au-delà des frontières d'un pays ou des capacités de réaction nationales. Par exemple, la récente pandémie de COVID-19 a constitué une crise sanitaire importante qui nécessite une action coordonnée à grande échelle. À cette fin, les systèmes de gestion des crises doivent analyser différentes données provenant de sources et de fournisseurs divers afin de prévoir les pandémies et de mobiliser les ressources nécessaires pour faire face à la maladie prévue. Ces systèmes peuvent donc traiter des données massives, hétérogènes et complexes. Par conséquent, nous avons adopté les data lakehouses comme référentiel pour surmonter les défis liés à la collecte et au stockage de données hétérogènes massives. Les data lakehouses combinent les aspects clés des lacs et des entrepôts, ce qui permet d'unifier le stockage à l'aide du modèle d'entrepôt de données unique et d'assurer la flexibilité analytique des data lakehouses. Néanmoins, le processus de prévision et de gestion des crises implique encore des défis non résolus en termes de nettoyage, de traitement et de recommandation des données, que nous examinons et abordons tout au long de ce manuscrit.

Dans ce chapitre, nous expliquons les concepts de base liés à nos propositions, tels que les lacs de données, la curation des données, les méthodologies de prévision des crises et l'intelligence artificielle explicable. De même, nous passons en revue les travaux qui ont été proposés concernant la curation des données, la prédiction des maladies et la recommandation et l'explication des mesures recommandée, et nous discutons de leurs limites.

Enfin, nous concluons ce chapitre en résumant notre étude et en discutant du positionnement de nos contributions.

2.2 Data lakehouses

De nos jours, les technologies émergentes ont conduit à l'émergence d'une quantité massive de données. Les big data englobent des ensembles de données trop volumineux pour être traités par les systèmes de base de données traditionnels. Elles nécessitent donc de nouveaux processus et méthodes de stockage, d'intégration, de traitement et d'analyse. En raison de l'aspect hétérogène des big data, les référentiels, comme les entrepôts de données classiques, nécessitent l'exécution d'un processus ETL (Extract - Transformation - Load) avant de stocker les données, ce qui peut s'avérer coûteux pour le stockage des big data. Ainsi, les nouvelles générations des référentiels de données comme les lacs de données et les data lakehouses se basent sur le processus ELT qui inverse l'ordre d'exécution des étapes de chargement et de transformation afin de surmonter les limites de l'ETL. Nous présentons une vue d'ensemble des approches d'intégration de données existantes et des référentiels de données.
Un entrepôt de données est un dépôt central de données provenant de différentes sources internes et externes de l'entreprise. Il recueille des données provenant de diverses sources, tant internes qu'externes, et optimise la recherche de données à des fins commerciales. Contrairement aux bases de données classiques, un entrepôt de données stocke des données historiques, structurées, non volatiles et orientées objet, ce qui le rend fondamentalement conçu pour l'analyse des données dans le contexte de la prise de décision. Néanmoins, les données collectées dans les bases de données et les entrepôts de données doivent être nettoyées et préparées avant d'être stockées, ce qui peut handicaper les opérations dans certains contextes décisionnels où le temps peut être critique. Les lacs de données ont donc été proposés pour résoudre ce problème. Un lac de données est un entrepôt centralisé englobant des données volumineuses (big data) dans un format brut et granulaire provenant de nombreuses sources. En conséquence, les lacs de données s'appuient sur le processus ELT pour stocker les données et optimiser les coûts de stockage. Néanmoins, les lacs de données n'offrent pas les mêmes performances en termes de gestion et d'optimisation que les entrepôts de données. Les data lakehouses sont donc proposés comme une solution qui combine les principaux atouts des deux architectures de données susmentionnées (c'est-à-dire les entrepôts de données et les lacs de données). Plus précisément, les data lakehouses garantissent un stockage à faible coût dans un format ouvert accessible par différents systèmes, ainsi que des fonctions de gestion et d'optimisation robustes. Contrairement aux lacs de données, les data lakehouses sont des espaces de stockage directement accessibles qui offrent des fonctionnalités traditionnelles de SGBD telles que les transactions ACID, le versionnage des données, l'audit, l'indexation, la mise en cache et l'optimisation des requêtes [11]. Nous présentons une comparaison entre les différents référentiels de données dans le tableau 2.1.

TABLE 2.1 – Comparaison de différents référentiels de données

	Entrepôt de donnée	**Data Lake**	**Data Lakehouse**
Format de donnée	Transformée	Brut	Brut
Intégration de donnée	ETL	ELT	ELT
Qualité de donnée	Curée	Non garantie	Non garantie
Schema	Schema-on-write	Schema-on-read	Schema-on-read
Interrogeable	Oui	Non	Oui
Transactions ACID	Oui	Non	Oui
Maturité	Mature	Immature	Immature

2.3 Architecture orientée service

2.3.1 Définition

Une architecture orientée services (SOA) est composée de services découvrables et faiblement couplés. Dans cette architecture, les fournisseurs et les consommateurs de services sont indépendants, et les utilisateurs peuvent composer des services dans un processus métier pour créer de nouveaux services. [12]

2.3.2 Composition des services

La composition de services est le processus de combinaison de services atomiques en services composites à valeur ajoutée. Elle intègre des services pour réaliser une tâche spécifique et ajoute ainsi une plus grande valeur [13]. En outre, elle s'appuie sur la sélection de services qui choisit les services les plus appropriés parmi un ensemble de services disponibles pour répondre aux exigences et contraintes fonctionnelles et non fonctionnelles d'un utilisateur [14].

2.3.3 Types de compositions

Nous distinguons deux types de composition de services :

— **Composition statique de services :** les flux de contrôle et de données sont définis par l'utilisateur au moment de la conception.

— **Composition dynamique de services :** les flux de contrôle et de données sont générés automatiquement au moment de l'exécution.

2.3.4 Orchestration et chorégraphie des services

Pour réaliser la composition de services, nous avons identifié les deux approches suivantes :

— **Orchestration :** il contrôle activement les services et les interactions, comme le font les musiciens d'un orchestre. Néanmoins, l'inconvénient de cette approche est que le contrôleur doit communiquer et attendre la réponse de chaque service, ce qui peut avoir un impact sur les performances du système.

— **Chorégraphie :** la chorégraphie crée un modèle que les services doivent suivre sans supervision. Elle permet donc de créer des systèmes plus rapides, plus cohérents, plus souples et plus efficaces.

2.4 Multi-agent systems

Nous considérons un agent comme une entité ayant des composantes cognitives telles que des croyances, des capacités, des choix et des engagements [15]. En conséquence, les agents intelligents effectuent trois actions en continu :

1. Surveiller de l'environnement pour détecter les conditions dynamiques
2. Influer sur les conditions de l'environnement
3. Tirer des conclusions et des raisonnements pour interpréter des perceptions et résoudre des problèmes

La figure 2.1 présente une vue d'ensemble des différents composants et de leurs interactions dans un système multi-agents.

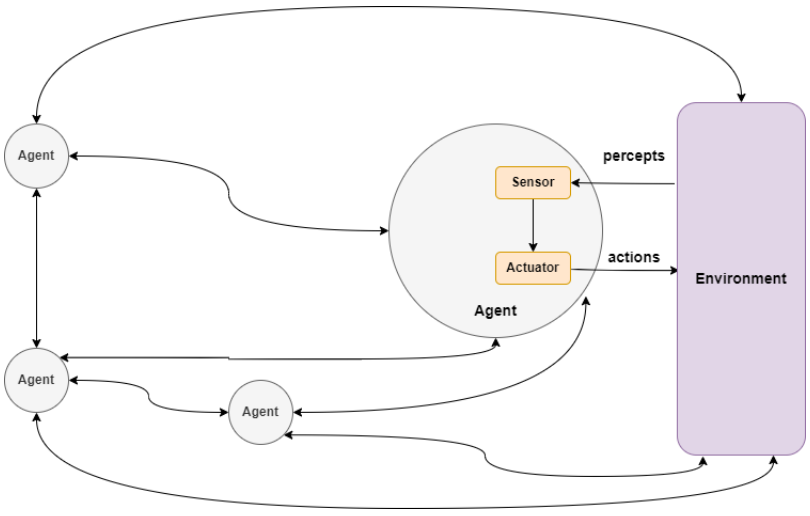

FIGURE 2.1 – Aperçu sur les systèmes multi-agents

2.5 Data curation

Les data lakehouses peuvent être une solution efficace dans les contextes et situations d'urgence actuels, tels que les crises, qui nécessitent la collecte et l'analyse de données en temps réel. Néanmoins, l'hétérogénéité et la complexité des données restent des défis majeurs pour les entrepôts de données. Le traitement des données, qui fait partie de la curation des données, est donc nécessaire pour améliorer la qualité des données avant leur analyse ou leur visualisation. À cette fin, la curation de données assure la gestion et la promotion de l'utilisation des données à partir de leur point de création en les enrichissant ou en les mettant à jour pour qu'elles restent adaptées à un objectif spécifique. En conséquence, la curation des données fournit davantage d'informations sur la provenance des données, le contexte original de mesure et d'utilisation, et l'objet de l'observation pour faciliter la réutilisation des données [16]. Nous présentons dans les sections suivantes les concepts de base et analysons les travaux existants relatifs à la curation.

2.5.1 Champs d'application de la curation

Dans cette section, nous présentons les différents niveaux concernés par la curation.

Données

La curation des données englobe les tâches de curation consacrées au nettoyage et à l'enrichissement des données. En effet, la curation des données couvre les tâches de traitement des données qui englobent la réparation et la déduplication des données, qui préparent les données à être traitées dans le cadre d'opérations ultérieures. Comme la qualité des données d'entrée reflète la qualité des résultats d'un système, la curation des données englobe les tâches d'enrichissement qui assurent l'organisation et la maintenance des ensembles de données auxquels les utilisateurs à la recherche d'informations peuvent accéder et qu'ils peuvent utiliser. Comme indiqué précédemment, les lacs de données/lakehouses contiennent des données collectées à partir de différentes sources qui peuvent être complexes, incertaines et hétérogènes. Par conséquent, l'enrichissement sémantique, tel que la liaison avec des bases de connaissances externes, est nécessaire pour promouvoir la qualité des résultats. D'autre part, la contextualisation via le profilage des données, par exemple, est nécessaire pour organiser le lac de données/data lakehouse et, par conséquent, faciliter son indexation et son catalogage. En conséquence, la curation des données est une étape primordiale qui empêche le lac de donnée/data lakehouses de se transformer en un marécage de données. Un marais de données est un lac de données/data lakehouse non géré, inaccessible aux utilisateurs ou de faible valeur.

Metadonnées

La curation des métadonnées est l'une des opérations les plus nécessaires pour maintenir l'organisation d'un lac de données/data lakehouse. La gestion des métadonnées crée un catalogue pour indexer les ensembles de données regroupés dans le lac de données/data

lakehouse. En conséquence, le système peut identifier les ensembles de données nécessaires à l'analyse lorsqu'un processus d'analyse des données est exécuté. Contrairement aux entrepôts de données dans lesquels la gestion des métadonnées est bien assurée, les lacs de données n'ont pas de gestion des métadonnées et se transforment rapidement en un marécage de données si la gestion des métadonnées n'est pas bien définie. Pour cela, des tâches de curation telles que la modélisation et l'extraction des métadonnées sont nécessaires pour assurer l'organisation d'un lac de données/data lakehouse. La modélisation des métadonnées se concentre sur la définition de modèles pour la représentation et l'identification des ensembles de données. Par exemple, différentes zones d'organisation (données brutes, zones de données conservées, etc.) peuvent être créées dans un lac de données/data lakehouse pour organiser les ensembles de données en fonction des besoins de l'utilisateur final. D'autre part, l'extraction des métadonnées permet d'extraire les métadonnées nécessaires des ensembles de données selon des règles définies afin d'identifier les caractéristiques des ensembles de données (par exemple, le format, l'auteur, etc.). Néanmoins, la gestion des métadonnées est comprise dans les data lakehouses, contrairement aux lacs de données.

Schéma

Comme les lacs de données et les data lakehouses ingèrent des sources de données très hétérogènes provenant de plusieurs sources, ces données peuvent être présentées sous différents formats, comme des sources de données non structurées. Ces dernières doivent être transformées pour être utilisées dans des opérations d'analyse ultérieures. Cependant, les sources de données non structurées n'ont pas de schéma défini et peuvent ne pas respecter un format de représentation des données spécifique. C'est pourquoi la curation des données englobe des tâches dédiées à la curation des schémas. L'extraction de schémas tente d'identifier une représentation de données afin d'extraire un schéma pour l'ensemble de données. La cartographie et l'extraction de schémas recherchent des liens entre un ensemble de données et les ensembles de données connexes regroupés dans le référentiel de données. En outre, comme ces référentiels de données continuent d'ingérer des données en permanence, le contenu et le schéma des ensembles de données peuvent évoluer et changer au fil du temps. Par conséquent, les tâches d'évolution du schéma permettent de maintenir les ensembles de données à jour en faisant évoluer le schéma des ensembles de données à chaque fois. Cette fonctionnalité fait partie intégrante du data lakehouse.

2.5.2 Tâches de curation des données

En analysant la littérature, nous avons identifié plusieurs tâches dédiées à la curation des données, des métadonnées et des schémas. En effet, ces tâches doivent être réorganisées dans un ordre spécifique pour créer un pipeline de curation. Par exemple, un pipeline de curation peut être constitué à partir de tâches organisées dans l'ordre suivant : "Étiquetage POS $\rightarrow$ Identification des racines $\rightarrow$ Liaison avec des classes d'ontologies". L'étiquetage POS consiste à associer les mots d'un texte aux informations grammaticales correspondantes, telles que le genre. Ensuite, ce pipeline identifie les racines des noms.

Ces derniers sont ensuite associés à des classes de l'ontologie afin de procéder à un enrichissement sémantique. Nous avons donc classé les tâches de curation en trois catégories en fonction de l'objectif de la curation. La première catégorie est la curation de données, qui se compose de trois sous-catégories : la contextualisation, la réparation des données et l'établissement de liens sémantiques. Les tâches de contextualisation couvrent les tâches NLP visant à contextualiser ou à effectuer un profilage des données. Nous distinguons des tâches telles que l'étiquetage POS, l'identification de racines, l'identification d'entités nommées, etc. Les tâches de réparation des données sont des tâches dédiées à la manipulation des données, telles que la réparation des données manquantes et l'identification des données erronées. La liaison sémantique englobe les tâches de curation qui visent à réaliser un enrichissement sémantique, comme la mise en correspondance avec des bases de connaissances externes.

Quant à la curation des métadonnées, elle englobe l'extraction et la modélisation des métadonnées. En effet, nous avons identifié des contributions qui analysent les sources de données pour identifier et constituer des métadonnées, tandis que d'autres travaux ont proposé des modèles pour représenter les métadonnées.

En ce qui concerne la curation de schémas, nous avons examiné plusieurs contributions qui extraient, font le matching et le mapping des schémas des sources de données. Les tâches de curation de schémas sont nécessaires pour gérer l'évolution des schémas. Ces tâches sont principalement appliquées aux données semi-structurées et non structurées afin d'identifier le schéma des sources de données. En ce qui concerne le matching et le mapping des schémas, le premier relie un attribut de la source de données A à un autre de la source de données B, tandis que le second identifie les attributs communs ayant des représentations différentes (par exemple, le nom, le type, etc.) dans deux sources de données. Par exemple, nous considérons trois ensembles de données ayant les attributs suivants : Personne (nom, âge, profession), Population (nomP, âgeP, emploi) et Médecins (nom, diplôme, spécialité). En conséquence, nous pouvons effectuer un mapping de schéma pour identifier la correspondance entre les attributs des ensembles de données Personne et Population (par exemple, l'attribut Nom correspond à NomP, et la Profession est représentée par Emploi dans l'ensemble de données Personne). D'autre part, le matching de schéma relie les ensembles de données Personne et Médecins en utilisant l'attribut Nom pour créer un quatrième ensemble de données représentant les informations sur les médecins (c'est-à-dire qu'il fonctionne de manière similaire à l'instruction JOINT dans SQL).

Comme les lacs de données et les lakehouses collectent des données en mode batch et en mode streaming, des pipelines de curation de données fiables doivent être définis pour améliorer la qualité des données avant de les utiliser dans des processus d'analyse de données. En effet, les données en batch et en streaming peuvent nécessiter des tâches de curation de données spécifiques en fonction du format des données, ce qui fait qu'elles ont besoin de pipelines de curation de données différents. En fait, les flux de données peuvent être collectés à partir d'IoT qui regroupent différents capteurs. Par conséquent, des tâches de curation de flux de données dédiées, telles que la normalisation et la standardisation des

données, sont nécessaires pour garantir la curation des données en streaming. Par exemple, nous pouvons avoir deux capteurs qui capturent la température en degrés Celsius et Fahrenheit. La curation des données permet donc de mettre à l'échelle les données afin de créer une représentation unifiée des données.

En analysant les travaux de curation existants, nous avons proposé la taxonomie des principales catégories de tâches de curation de données en batch et en streaming, représentée dans la Figure 2.2. Ces tâches de curation assurent les opérations nécessaires à la curation des données. Nous soulignons que la détection de la dérive des concepts est consacrée à la curation des données en streaming. Les tâches de dérive des concepts détectent la déviation des données capturées en streaming en raison d'une défaillance possible du capteur. Néanmoins, les tâches de curation des autres catégories pourraient être utilisées pour la curation des données collectées en mode batch et en mode streaming.

Nous présentons dans la section suivante les travaux examinés relatifs à la curation des données, des métadonnées et des schémas.

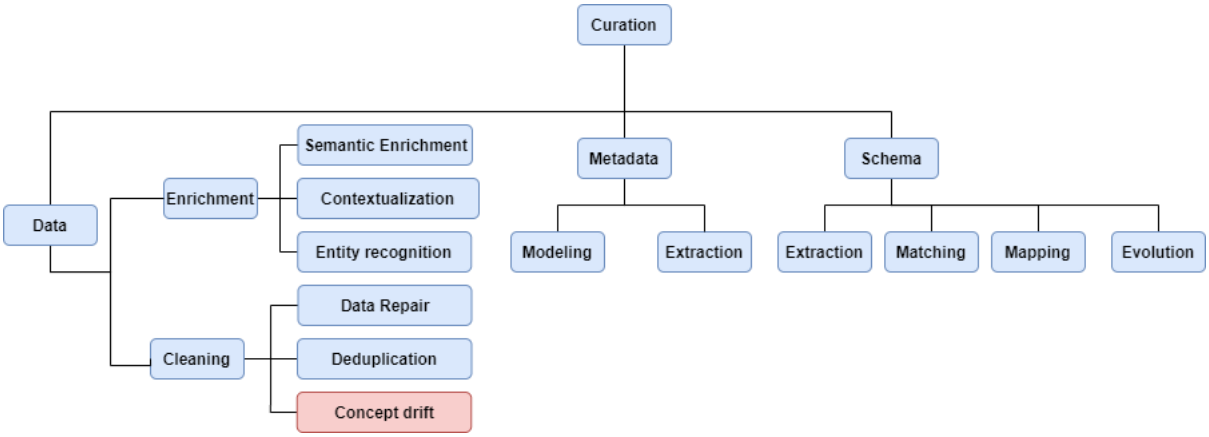

FIGURE 2.2 – Taxonomie des tâches de curation de données

2.5.3 Examen des travaux existants de curation des données

La gestion des données, en général, et la curation des données, en particulier, ont fait l'objet de plusieurs travaux. Nous avons identifié plusieurs approches dans la littérature traitant du processus de curation des données. Dans ce qui suit, nous présentons et examinons les travaux pertinents qui traitent du problème de la curation de données. Pour ce faire, nous nous appuyons sur des critères qui prennent en compte les techniques utilisées, le type de source de données à curer, l'automatisation de l'approche (c'est-à-dire entièrement automatique, semi-automatique, non automatique), le degré d'adaptivité par rapport aux caractéristiques du processus de décision et la nature des données traitées (c'est-à-dire les sources de données par batch/streaming). Le tableau 2.2 décrit les travaux de curation de données examinés. Beheshti et al. [18] ont présenté un pipeline de curation de

TABLE 2.2 – Synthèse des travaux de curation de données examinés
(Cx : Contextualisation, L : Liaison sémantique, Rp : Réparation de données, ML : Apprentissage automatique, S : Techniques sémantiques, G : Techniques à base de graphes, C : Crowdsourcing, R : Techniques à base des règles, U : Données non structurées, SS : Données semi-structurées, S : Données structurées)

Article	**Architecture**	**Tâches de curation**					**Techniques utilisées**					**Données**		
		Donnée			**Metadonnée**	**Schema**								
		Cx	L	Rp			**ML**	**S**	**G**	**C**	**R**	**U**	**SS**	**S**
Skluma, 2017[17]	Pipeline	X			X		X					X	X	
Crowdcorrect, 2018[18]	Pipeline	X					X			X			X	
CoreKG, 2018[19]	Service-based	X					X			X			X	
KAYAK, 2018[20]	Framework		X		X				X				X	
Pomp et al., 2020[21]	Framework		X		X	X			X		X			X
Semlinker, 2020[22]	Framework					X		X					X	
Data synapse, 2018[23]	Service-based		X		X	X		X					X	
Hai et al., 2019	Method				X	X					X		X	
VADA, 2019[24]	Framework	X	X	X		X	X				X			X
Lenses, 2015[25]	Method			X		X					X			X
Simonini, 2019[26]	Method					X			X				X	X
DATAMARAN, 2018[27]	Method					X			X				X	

données (c'est-à-dire une séquence de tâches de curation) qui permet aux analystes d'effectuer le nettoyage et la curation de données sociales. Leur approche vise à préparer les données pour une analyse fiable des données commerciales. L'idée consiste à améliorer la curation des données à chaque étape du pipeline de curation en appliquant d'abord la curation automatique (c'est-à-dire la curation des données à l'aide de services) et la curation semi-automatique (c'est-à-dire le crowdsourcing, l'annotation par des experts) par la suite. L'idée proposée a été employée dans plusieurs autres travaux pour prouver son efficacité dans divers cas d'utilisation, tels que [23] et [19]. Par exemple, Datasynapse [23] est un pipeline de curation qui permet de curer les données pour les contextualiser. En conséquence, les données contextualisées sont stockées dans un lac de données basé sur les connaissances à l'aide de CoreKG, un service dédié proposé par les auteurs et qui s'appuie sur le pipeline de curation pour constituer ce lac de données[19].

Un autre travail proposé par Kanstaninou et al. [24] a présenté une architecture qui contient des composants de préparation de données faiblement couplés. Ces derniers sont constitué à partir des techniques à base des régles et d'apprentissage automatique et peuvent être orchestrés dynamiquement pour exécuter des tâches de préparation des données telles que le matching, le profilage des données, la génération de mapping, la transformation de format et la réparation des données. Malgré la flexibilité de cette approche, elle n'est consacrée qu'à la curation de sources de données structurées.

Dans le contexte de la dynamicité et de la flexibilité, Maccioni et al. ont présenté un framework nommé KAYAK qui encapsule des techniques de graphes pour présenter des pipelines pour la préparation des données [20]. Le framework proposé se situe entre les utilisateurs/applications et les couches du système de fichiers (c'est-à-dire l'emplacement

de stockage des données), et il expose un ensemble de primitives et de tâches pour la préparation des données. Chaque tâche de préparation des données combine un ensemble de primitives représentant une étape de préparation simple (par exemple, insérer un ensemble de données, calculer la joignabilité). Comme pour [18], les tâches de préparation doivent être combinées pour accomplir la tâche de préparation finale des données. Sur la base de ces tâches, les utilisateurs peuvent concevoir leur pipeline de préparation des données (c'est-à-dire l'organisation des tâches de préparation des données) représenté sous la forme d'un DAG (Direct Acyclic Graph). KAYAK contient également une base de connaissances qui englobe des métadonnées sur les sources de données (par exemple, noms d'attributs, noms de relations, etc.), le schéma cible et le flux de travail (par exemple, l'état actuel de la préparation des données, la provenance des informations, etc.) Le framework proposé effectue également le matching et le mapping des schémas afin de maintenir le versionnage des schémas.

Nous avons également examiné Skluma [17], un système automatisé de traitement de grandes quantités de données et de métadonnées profondément imbriquées, de thèmes latents, de relations entre les données et d'extraction de métadonnées contextuelles à partir de documents connexes. Pour ce faire, Skluma utilise l'apprentissage automatique pour extraire les métadonnées des fichiers en fonction de leur contenu et de leur niveau général (nom de fichier, chemin d'accès, taille, somme de contrôle, etc.)

Pomp et al. ont proposé ESKAPE[21], une approche pour l'intégration sémantique des données via l'apprentissage et la représentation évolutive des données. Les parties prenantes peuvent utiliser l'approche proposée pour l'enrichissement sémantique et la réparation des données. Par exemple, certaines plages de valeurs (comme la température) peuvent varier en fonction du contexte. ESKAPE peut donc être utilisé pour unifier la représentation des données et récupérer des informations supplémentaires pour l'enrichissement sémantique. D'autre part, le travail proposé par Pomp et al. décrit des méta-informations supplémentaires requises et intéressantes avec des modèles sémantiques. En ce qui concerne la curation des schémas, ce travail unifie la représentation des données en reliant les concepts sémantiques. Ainsi, il incorpore des autoencodeurs pour une recommandation automatique de classes de données en construisant un graphe de connaissances qui combine les différentes représentations d'une instance de données.

Lenses [25] est un modèle dynamique pour concevoir des tâches de curation de données basées sur des règles et des composants probabilistes appelés lentilles. Les auteurs représentent une lentille comme un composant pour le traitement des données qui peut faire partie d'un pipeline ETL typique. En outre, Lenses s'appuie sur la correspondance des schémas pour faire correspondre le schéma de la source de données à un schéma cible défini par l'utilisateur. Par exemple, la correspondance des schémas pourrait être appliquée pour créer des relations entre des objets JSON ou des tables Web qui peuvent nécessiter des schémas bien définis.

En outre, SemLinker gère l'évolution des schémas en intégrant la technologie sémantique.

En effet, il construit une ontologie globale englobant les différentes versions des schémas de données. En conséquence, les représentations de données de chaque version de schéma de données sont liées sémantiquement. Nous présentons également l'approche proposée par Hai et al.[28] pour l'intégration des données qui détecte les dépendances fonctionnelles pour les données stockées dans des sources hétérogènes et, par conséquent, unifie les différentes représentations de données.

Quant au travail proposé par Simonini et al, [26], il présente une approche pour extraire des informations de schéma lâche, représenté comme graphe contenant un ensemble d'attributs, à partir d'un ensemble de données en utilisant une technique d'induction de match d'attributs. Quant à DATAMARAN [27], il s'agit d'un outil qui extrait automatiquement la structure d'ensembles de données semi-structurés. Pour ce faire, il extrait une vaste collection de modèles de structure à partir d'enregistrements potentiels représenté sous forme de graphe, puis élimine la plupart des candidats. Ensuite, il applique deux techniques de raffinement de la structure pour mettre à jour les modèles de structure.

2.5.4 Analyse et positionnement

La revue de la littérature a montré un nombre infime de travaux traitant de la curation de données dans les data lakehouses, aussi, nous avons considéré des travaux sous-jacents. L'examen desdits travaux met la lumière sur diverses propositions qui abordent la curation selon trois niveaux : les données, les métadonnées et les schémas. Comme nous nous appuyons sur les data lakehouses pour concevoir notre solution, nous nous concentrons sur la curation des données puisque la curation des métadonnées et des schémas est implicitement gérée dans les data lakehouses. Même si certains travaux combinent différents niveaux de curation, nous mettons en lumière les travaux couvrant la curation des métadonnées et des schémas afin d'avoir une vision complète des contributions liées à tous les niveaux de curation, en général, et à la curation des données, en particulier.

En ce qui concerne la curation de données, nous avons remarqué que la plupart des approches proposées ne pouvaient pas être généralisées pour traiter toutes les structures de sources de données après avoir analysé les travaux énumérés ci-dessus. Elles sont donc conçues pour traiter un format de source de données spécifique (c'est-à-dire non structuré, semi-structuré ou structuré). Or, les data lakehouses contiennent différents formats de sources de données, ce qui nécessite des outils sophistiqués pour curer les sources de données. En ce qui concerne l'automatisation du processus de curation, nous avons identifié plusieurs approches qui ne sont pas entièrement automatiques. Dans [18], les auteurs ont proposé divers services pour la curation des données. Par la suite, les services de curation automatisés sont combinés avec le crowdsourcing et l'annotation par des experts pour améliorer le processus de curation. Cependant, leur approche peut être appliquée à la curation de données sociales, généralement présentes sous une forme semi-structurée. Comme [18], le travail présenté par Kanstaninou et al. [24] vise à préparer des sources de données structurées. Skluma [17], quant à lui, vise à contextualiser les sources de données non structurées et semi-structurées, telles que la documentation, les fichiers README,

CSV et les fichiers de texte brut. ESKAPE[21] suggère des concepts sémantiques basés sur des valeurs de sources de données structurées. En ce qui concerne Lenses [25], ce framework est dédié à la curation de sources de données structurées.
En ce qui concerne l'automatisation du processus de curation des données, notre étude révèle que les travaux proposés dans [24], [21], et dans [17] sont entièrement automatiques. En revanche, les autres contributions étudiées sont manuelles, comme [20], ou semi-automatiques, comme [18], et [25]. Néanmoins, comme présenté précédemment, l'intervention de l'acteur humain peut être sujette à des erreurs et prendre du temps.

Notre analyse de la littérature a permis d'identifier plusieurs techniques utilisées pour proposer des contributions de curation, telles que l'apprentissage automatique, les technologies sémantiques, les techniques basées sur les graphes, le crowdsourcing et les techniques basées sur les règles. Nous avons remarqué que l'adoption des techniques d'apprentissage automatique est encore limitée, en particulier pour le nettoyage des données et la mise en correspondance des schémas. En effet, plusieurs tâches de curation telles que la déduplication, le repérage des erreurs, la violation et la réparation des données sont difficiles à automatiser. Par conséquent, la plupart des approches de curation étudiées reposent sur des techniques basées sur des règles, des techniques sémantiques ou l'incorporation de l'apprentissage automatique avec l'une des techniques susmentionnées. Par conséquent, nous avons identifié des contributions basées sur des règles pour les tâches de curation telles que la détection de violations comme [24]. Sinon, les techniques d'apprentissage automatique sont combinées avec d'autres méthodes, telles que le crowdsourcing, pour effectuer la tâche de curation, comme le travail présenté dans [18]. Ce dernier propose une approche semi-automatique qui s'appuie sur la curation automatique via des services de curation et des annotations manuelles via le crowdsourcing et les annotations d'experts. Chaque service des services de curation proposés [29] représente une étape de curation (par exemple, lier un ensemble de données à une base de connaissances) qui peut être employée pour constituer un pipeline de curation. Ces services reposent sur plusieurs techniques telles que la définition de règles, la liaison avec des dictionnaires et des ontologies, et l'apprentissage automatique. Cependant, le pipeline de curation proposé nécessite également une intervention humaine via le crowdsourcing, qui peut parfois être sujette à des erreurs et prendre du temps comme présenté ci-dessus. Néanmoins, l'incorporation de l'apprentissage automatique avec d'autres techniques peut s'expliquer par l'aspect subjectif de certaines tâches de curation, telles que la détection d'erreurs, qui ne peuvent être identifiées à l'aide d'une série de règles et nécessitent une intervention humaine.

Nous avons également étudié la flexibilité des approches examinées. En effet, nous soulignons que toutes (1) les approches étudiées sont statiques en ce qui concerne les caractéristiques du processus de décision. De plus, (2) les travaux présentés dans [24] et [25] assurent un faible niveau d'adaptation en tenant compte des besoins de l'utilisateur final. En revanche, [20] continue d'effectuer la même orchestration des tâches de curation, quelles que soient les exigences de l'utilisateur. En outre, à notre connaissance, (3) toutes les approches examinées ne prennent en compte que les sources de données batch.
En outre, notre étude révèle que l'autonomie totale dans l'exécution de la curation et l'uti-

lisation de l'apprentissage automatique et de la généralisation des règles sont encore des questions ouvertes en matière de curation.
En se basant sur les limites des approches présentées, une solution doit être proposée pour les surmonter à travers la curation de données adaptative. C'est pourquoi nous détaillons dans le chapitre suivant la solution que nous proposons pour la curation de données en batch et en streaming simultanément, de manière adaptative au contexte décisionnel, au profil de l'utilisateur, à ses contraintes et préférences, et au type de source de données à traiter.

Dans ce qui suit, nous présentons les travaux liés à la conception d'un système de recommandation de mesures pour la gestion des crises. En effet, nous visons à analyser les données curées pour prédire les changements de situations qui peuvent provoquer une crise et recommander des mesures utiles pour aider à la gérer.

2.6 Approches de gestion des crises

À l'examen de la revue de la littérature, plusieurs définitions ont été dédiées au concept de crises, de leurs gestions, types et champs d'application. Certains auteurs considèrent qu'il s'agit d'un événement dont les causes peuvent être inconnues, tandis que d'autres considèrent qu'il est causé par une succession d'événements antérieurs. Cependant, la plupart des travaux proposés ont en commun le fait qu'une crise a des conséquences graves. En outre, certains auteurs utilisent des termes spécifiques tels que désastre et catastrophe [30] qui peuvent être davantage liés à des contextes spécifiques. Suite à notre analyse, nous avons identifié que les catastrophes naturelles et d'autres types de crises, comme les crises sanitaires et les récessions économiques, sont considérées comme des crises au niveau macro et représentent l'objet principal de la plupart des travaux existants. Néanmoins, il existe des crises au niveau microéconomique comme les défaillances des services d'une organisation à titre d'exemple. En ce qui concerne la notion de gestion des crises, les crises économiques sont les plus étudiées dans la littérature, suivies par les crises sanitaires (notamment celles causées par des épidémies et pandémies). Par exemple, Hao et al. [31] ont proposé un framework pour la gestion de la pandémie de COVID-19 qui s'adressait au domaine du tourisme. Leur contribution est une base théorique qui englobe quatre principes, à savoir (1) l'évaluation de la catastrophe, (2) la garantie de la sécurité des employés, des clients et des biens, (3) les mesures d'autosauvetage, et (4) l'activation et la revitalisation de l'entreprise. Chacun de ces principes englobe plusieurs stratégies répondant à chaque étape du COVID-19. Dans le même contexte, les auteurs de [32] ont proposé un framework pour la gestion de crise basée sur la connaissance dans l'industrie de l'hôtellerie et du tourisme. En conséquence, le processus de gestion des crises est guidé par les principes de gestion des connaissances afin de garantir l'efficacité et la rigueur de la planification et de l'application des stratégies à chaque phase de la crise. Jung et al. [33] ont proposé un framework conceptuel d'un système intelligent de gestion des catastrophes qui porte sur les catastrophes naturelles. Pour ce faire, ils ont utilisé des techniques d'apprentissage automatique et d'apprentissage profond, telles que les réseaux neuronaux convolutifs, pour

analyser et détecter les incendies dans les vidéos de surveillance. Nous présentons également un travail de gestion de crise au niveau micro, tel que [34], dans lequel les auteurs ont proposé un nouveau modèle qui combine deux approches existantes pour prédire et gérer le risque de la chaîne d'approvisionnement. Les travaux présentés et d'autres ont été dédiés à la planification de ladite gestion à l'image des processus avec différentes phases (pré-crise, milieu de crise et post-crise), où à chacune d'elles, la gestion des crises peut évoluer en termes de besoins et d'exigences. A titre d'exemple, dans les premières phases, les actions sont misent sur l'identification des plans stratégiques et de leurs facteurs influents, ainsi que sur la prévision d'une crise future. Pendant une crise, la gestion au milieu d'une crise couvre les tâches liées à la comparaison de l'efficacité des stratégies de réponse et à la mesure de leurs impacts à court terme. De même, dans la gestion post-crise, les travaux proposés analysent l'impact à long terme de la crise pour ensuite proposer et mettre en œuvre des stratégies de réponse et de récupération. Néanmoins, malgré la variété des travaux existants proposés pour chaque niveau, ils ne prennent pas suffisamment en compte les différents niveaux d'analyse (par exemple, national et/ou international) et l'incorporation des exigences des différents acteurs impliqués dans un tel processus. Par exemple, le travail [35] souligne la nécessité de prendre en considération les attentes, les préférences et les satisfactions des parties prenantes qui peuvent changer au cours des différentes étapes de la gestion de crise (c'est-à-dire avant, au milieu et après la crise). En outre, il souligne la nécessité de prendre en compte les caractéristiques et les différences associées à la gestion multinationale des crises. Ce travail et d'autres comme [36] ont mis en évidence la nécessité de renforcer la capacité cognitive des systèmes et leur adaptabilité aux différents besoins et changements dans la gestion des crises.

2.7 Analyse des approches existantes en matière de prédiction des maladies

Aujourd'hui, la rapidité et l'ampleur des échanges humains et de l'immigration ont fait de la lutte contre les maladies infectieuses un enjeu mondial, aucun État ne pouvant songer à se retrancher derrière ses frontières [37]. En effet, les crises et les urgences sanitaires ont un impact considérable sur les plans économique, social et géopolitique. La prévision des maladies et la gestion des crises et des urgences sanitaires jouent donc un rôle important dans la planification stratégique des actions. En raison de la complexité de la gestion des épédimies, un système de gestion de crise assure l'organisation de la veille et de la sécurité sanitaires [38]. Avec l'apparition de l'épidémie de COVID-19, plusieurs méthodes, plateformes et approches ont été proposées pour prédire et gérer cette maladie. Le tableau 2.3 présente les travaux examinés pour la prédiction des maladies sanitaires. Nous avons identifié deux catégories de contributions, à savoir les approches basées sur les capteurs et les approches basées sur les données. En ce qui concerne les approches basées sur les capteurs, nombre d'entre elles ont utilisé des capteurs et des IoT pour mesurer différents paramètres. En effet, Mir et al. [39] ont proposé un framework qui exploite les paramètres de santé (par exemple, la fièvre, l'essoufflement, la toux, la fatigue, l'historique des voyages, l'oxygène, etc.) collectés à partir de capteurs et d'IoT en temps réel et

calcule la présence du virus COVID-19. Suite à l'identification d'un cas suspect, le framework proposé partage les données avec les centres et professionnels de santé et envoie les patients pour des tests et des consultations. À cette fin, le framework englobe quatre composants principaux : le système utilisateur ou la collecte de données, l'analyse des données, le diagnostic et le système cloud.
Mohammadi et al. [40] ont proposé un système de diagnostic des maladies qui utilise les IoT pour collecter les signaux médicaux d'un patient. Par la suite, un processeur cloud ou local stocke et traite les données collectées. Ainsi, le système de diagnostic permet de prendre des décisions rationnelles concernant la santé personnelle. Si le diagnostic identifie une urgence, une alerte sera émise vers l'hôpital le plus proche pour les urgences médicales.
Al Hossain et al. [41] ont proposé FluSense, une plateforme de surveillance syndromique qui surveille les signaux bio-cliniques des salles d'attente des hôpitaux pour prédire les maladies de type grippal. Pour ce faire, FluSense utilise un microphone et une caméra thermique pour capter le comportement des patients (par exemple, toux, activités vocales, etc.) dans les salles d'attente.
D'autre part, les approches basées sur les données utilisent des données collectées auprès des hôpitaux, des patients, des médias sociaux, etc. En effet, Zhang et al. [42] proposent une solution pour la prédiction de l'analyse du sentiment du coronavirus en utilisant deux composants : une analyse du sentiment hors ligne et un pipeline de prédiction en ligne. Le modèle d'analyse des sentiments hors ligne entraîne et teste les modèles d'apprentissage automatique (arbre de décision, machine à vecteurs de support, forêt aléatoire, régression logistique et voisins les plus proches) pour trouver le modèle optimal utilisé dans le pipeline de prédiction des sentiments en ligne. Ce dernier recueille des tweets en streaming et les introduit dans le modèle d'apprentissage automatique pour prédire l'analyse du sentiment des tweets en temps réel.
Nikparvar [43] a proposé un modèle de mémoire à long terme (LSTM) multi-variable qui est formé en utilisant la mobilité sur des séries temporelles multiples (par exemple, les données collectées à partir des rapports de mobilité de Google). Le modèle proposé prédit les futurs cas d'infection COVID-19, les décès et les schémas de circulation piétonne quotidienne.

TABLE 2.3: Synthèse des travaux de prédiction des maladies examinés

Paper	Category	Technology	Data	Outcome
Mir et al., 2022 [39]	Basé sur des capteurs	Capteurs basés sur l'IoT, linear regression, multilayer perceptron, autoregression	Données basées sur l'IoT (temperature, audio, rythme cardiaque, oxygène, etc.)	COVID-19 Prévision/ Envoi du patient au centre de santé

Table 2.3 **– suite de la page précédente**

Article	**Catégorie**	**Technologie**	**Données**	**Résultat**
Mohammadi et al., 2020 [40]	Basé sur des capteurs	Capteurs portables et adaptables, Réseaux de neurones	Données basées sur l'IoT	Décision concernant la santé du patient
Zhang et al., 2020 [42]	Basé sur des données	Decision-Tree, Support Vector Machine, Random Forest, Logistic regression, K-Nearest neighbor	Données Tweet sur le coronavirus	Analyse des sentiments
Al Hossain et al., 2020 [41]	Basé sur des capteurs	Capteurs basés sur l'IoT, Réseaux de neurones	Données relatives aux patients (toux, images thermiques, etc.)	La grippe
Nikparvar et al., 2021 [43]	Basé sur des données	LSTM (Recurrent Neural Networks)	Rapports sur la mobilité (par exemple, données sur le trafic piétonnier, etc.), données sanitaires (par exemple, cas confirmés COVID-19, décès, etc.), données démographiques (par exemple, densité de population, etc.).	Futurs cas confirmés, décès et trafic piétonnier quotidien

À l'issue de notre analyse, nous avons constaté que de nombreux systèmes de prévision des maladies étaient consacrés à la gestion de l'épidémie de COVID-19 en raison de l'émergence récente du virus SARS-CoV-2. Nous avons également identifié des approches consacrées à d'autres maladies, telles que la grippe, susceptibles de provoquer des épidémies. Les approches examinées génèrent différents résultats, tels que l'analyse des sentiments, les futurs cas confirmés, les décès, les décisions relatives à la santé des patients, etc. Pourtant, la plupart des propositions étudiées ne génèrent qu'un seul type de recommandation pour un seul rôle d'utilisateur. En outre, elles manquent de diversité dans l'élaboration des recommandations et traitent tous les risques de la même manière. Pourtant, certaines

caractéristiques peuvent influencer le processus de recommandation. Par exemple, les stratégies de santé dépendent des caractéristiques du pays, ce qui les rend différentes d'un pays à l'autre. Par exemple, depuis la découverte du premier cas d'infection par le COVID-19, la Chine a adopté des mesures sanitaires strictes pour lutter contre le virus, telles que le confinement de la ville de Wuhan le 23 janvier 2020 et les "quatre mesures précoces" (détection précoce, signalement précoce, isolement précoce et traitement précoce) le 2 février 2020. Par ailleurs, la Corée du Sud a détecté le premier cas de COVID-19 le 27 janvier 2020. En conséquence, la Corée du Sud a adopté des mesures moins strictes, telles que le contrôle aux frontières, le dépistage et les tests, avant d'imposer des contrôles strictes dans la ville de Daegu et la province de Gyeongsang du Nord le 25 février 2020. Contrairement à la Chine et à la Corée du Sud, qui ont mis en œuvre une stratégie de confinement, le Japon a adopté une stratégie d'atténuation à différentes étapes pour réduire la propagation du virus. Toutefois, les deux stratégies sanitaires ont montré des niveaux d'efficacité différents. En effet, l'efficacité de la stratégie sanitaire peut dépendre des caractéristiques du pays (population, indice de développement humain), ainsi que de plusieurs facteurs, dont la situation dans le pays et la gravité de l'épidémie. En outre, les stratégies sanitaires doivent évoluer en tenant compte des facteurs changeants et de l'expérience acquise au cours de la lutte contre le virus. Comme il n'existe pas de recommandation universelle, un système de gestion de crise doit s'adapter à ces facteurs changeants et proposer des mesures sanitaires adaptées à chaque situation. Suite à notre analyse, nous avons également remarqué que les travaux examinés n'ont pas proposé d'explications pour leurs résultats. Pourtant, comme nous l'avons vu plus haut, les experts du domaine peuvent chercher des raisons d'accorder leur confiance au modèle et de comprendre le raisonnement qui sous-tend la recommandation [44]. Par conséquent, nous surmontons les limites des approches mises en évidence grâce à notre contribution, qui est un modèle de recommandation sémantique explicable en matière de santé. Notre objectif est de prendre en compte les caractéristiques du pays et les préférences des utilisateurs pour proposer différentes mesures de gestion de crises adaptées à d'autres pays et à des rôles multi-utilisateurs. En conséquence, nous présentons dans les sections suivantes les concepts de base liés à l'intelligence artificielle explicable et aux systèmes de recommandation, puis les travaux examinés pour expliquer les modèles d'IA à l'aide d'eXplainable Artificial Intelligence.

2.8 Recommandation expliquée par l'intelligence artificielle explicable (XAI)

L'intelligence artificielle explicable (XAI) a récemment fait l'objet d'un grand intérêt de la part des chercheurs. Les premiers systèmes d'intelligence artificielle, tels que les systèmes experts et les systèmes basés sur des règles, étaient faciles à interpréter. Avec l'avènement de l'apprentissage automatique, les systèmes d'intelligence artificielle deviennent de plus en plus opaques. Par exemple, les règles de classification classique offrent un niveau d'interprétabilité plus élevé que les arbres de décision. La figure 2.3 illustre l'interprétabilité et la précision des modèles d'intelligence artificielle les plus courants. Comme le montre cette figure, plus le modèle d'intelligence artificielle est performant, moins il est in-

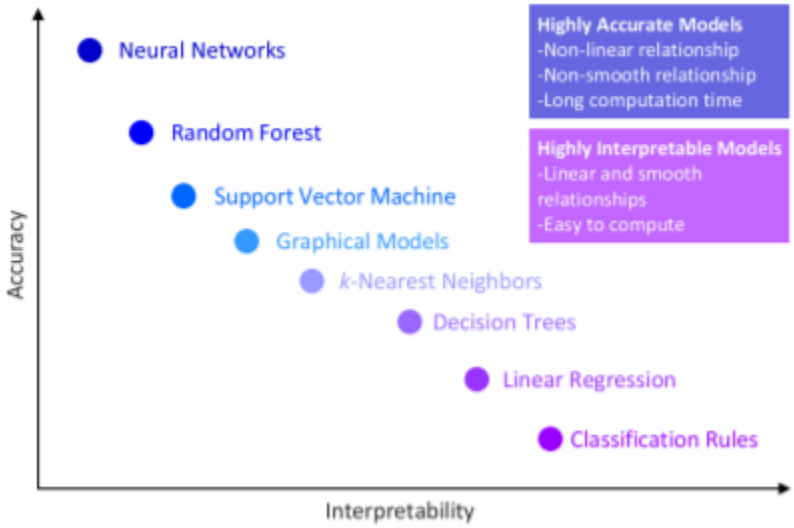

FIGURE 2.3 – Accuracy et interprétabilité des modèles d'IA [1]

terprétable. Malgré les performances et la précision des modèles d'apprentissage profond, ces modèles sont difficiles à interpréter. Pour cette raison, les modèles d'apprentissage profond sont qualifiés de modèles de boîte noire. C'est pourquoi les efforts des chercheurs se sont beaucoup tournés vers l'interprétation et l'explication des résultats générés par ces modèles. Dans cette section, nous présentons une vue d'ensemble du paradigme de l'intelligence artificielle explicative sous différents angles, tels que la terminologie, le champ d'application et les principes de la XAI.

2.8.1 Approches d'explication

Nous avons identifié deux approches permettant d'interpréter et d'expliquer les modèles de boîte noire : la première consiste à expliquer le fonctionnement interne des couches du modèle et à interpréter les processus intermédiaires. Cependant, étant donné la complexité de ces modèles, cette approche est limitée et nécessite un compromis entre l'interprétabilité et la précision du modèle. La seconde approche, "l'explication a posteriori" (post-hoc explanation), n'a pas d'impact sur la performance du modèle profond et garantit l'explication des résultats générés par l'analyse des entrées et des sorties sans affecter la performance du modèle. L'explicabilité post hoc vise les modèles qui ne sont pas facilement interprétables par définition. À cette fin, elle s'appuie sur divers moyens pour améliorer l'interprétabilité des modèles de boîte noire, tels que les explications textuelles, les explications visuelles, les explications locales, les explications par des exemples, les explications par la simplification et les techniques d'explication de la pertinence des caractéristiques. Chacune de ces techniques couvre l'une des façons les plus courantes dont les humains expliquent les systèmes et les processus.

2.8.2 Terminologie

Considérant notre travail, nous adoptons la terminologie proposée dans [1].

— L'intelligibilité est la caractéristique d'un modèle qui permet à un être humain de comprendre sa fonction sans expliquer son fonctionnement interne.
— La compréhensibilité désigne la capacité d'un algorithme d'apprentissage à représenter les connaissances apprises d'une manière compréhensible pour l'homme.
— L'explicabilité est la capacité d'expliquer ou de fournir une signification en termes compréhensibles pour un être humain.
— La transparence désigne le degré de transparence du modèle en lui-même pour qu'il soit compréhensible.

2.8.3 Champs d'application de l'explication

Nous présentons la portée des explications, l'un des critères largement utilisés pour classer les systèmes XAI. En effet, nous avons identifié deux principaux champs d'application des explications : local et global.

— **Explications globales** illustre les opérations du système ou le raisonnement adopté pour générer les résultats. Par exemple, un ensemble de règles imitant le fonctionnement interne d'un système d'intelligence artificielle peut être considéré comme une explication globale.
— **Explications locales** tentent d'expliquer les prédictions individuelles générées par le système plutôt que le fonctionnement du modèle d'IA dans son ensemble. Pour illustrer l'explication locale, nous avons présenté des exemples de résultats générés dans des cas similaires à celui du résultat proposé.

2.8.4 Recommandation et explication pour l'aide à la gestion de crises

Ayant pour objectif de proposer un système de recommandation des actions à entreprendre, nous avons étudié les travaux proposés pour la recommandation et les travaux sous-jacents tels que l'intelligence artificielle explicable, pour pallier à la faible interprétabilité des modèles d'IA. Ainsi, nous avons principalement sélectionné les travaux sur les travaux de l'Intelligence Artificielle eXpliquée dite égaement Interprétable (XAI) qui tirent profit des avantages des technologies sémantiques (comme les ontologies) pour ajouter de la sémantique aux modèles d'IA. Dans ce qui suit, nous détaillons les travaux examinés pour chaque concept.

Approches d'explication basées sur la sémantique : État de l'art

Ces dernières années, l'intelligence artificielle inteerprétable (XAI) a attiré l'attention de nombreux chercheurs. En effet, il existe un compromis entre la performance d'un modèle d'IA et son degré d'interprétabilité. Par exemple, les réseaux de neurones sont plus performants que les arbres de décision. Cependant, ils sont moins interprétables. Dans ce contexte, plusieurs approches et techniques ont été proposées pour favoriser l'interprétabilité des modèles d'IA, en particulier les modèles d'apprentissage profond. Suite à une

analyse de la littérature, nous avons identifié deux catégories d'approches de XAI. La première catégorie englobe les travaux qui visent à expliquer la fonction interne du modèle d'apprentissage profond en ajoutant une sémantique pour chaque couche. Néanmoins, ces travaux favorisent l'interprétabilité des modèles d'apprentissage profond plutôt que leur précision, ce qui peut avoir un impact sur leur performance globale. La deuxième catégorie comprend les méthodes *post-hoc explanation* qui expliquent les modèles d'apprentissage profond en extrayant les relations entre les entrées et les sorties d'un modèle d'IA afin d'apprendre la relation entre eux et de générer des explications [45]. Par conséquent, ces méthodes n'ont pas d'impact sur la performance du modèle. Malgré la diversité des approches proposées pour l'explication post hoc, il existe peu d'approches sémantiques pour expliquer les modèles d'apprentissage profond. Pourtant, les technologies sémantiques telles que les ontologies et les graphes de connaissances permettent de raisonner avec des concepts et des relations d'une manière proche de la façon dont les humains perçoivent les concepts liés. Les technologies sémantiques sont plus cohérentes et facilitent la navigation, l'évolutivité, la flexibilité et l'interopérabilité. Dans cette section, nous présentons et examinons quelques travaux pertinents sur la XAI basée sur la sémantique en examinant les techniques utilisées, la portée de l'explicabilité et les types d'explication.

Le tableau 2.4 présente les travaux examinés pour les approches d'intelligence artificielle explicable basées sur les technologies sémantiques. Les auteurs de [46] étendent TREPAN, un algorithme existant pour la simplification des réseaux de neurones artificiels aux arbres de décision, en incluant des ontologies pour la modélisation de la connaissance du domaine dans la génération d'explications. L'approche Doctor XAI [47] propose une technique d'explicabilité qui traite des données multi-étiquetées et liées à des ontologies. Les auteurs de cet article se concentrent sur l'explication Doctor AI, qui est un modèle de prédiction de la prochaine visite à l'aide des antécédents cliniques du patient.
Le travail [48] a proposé une approche d'explication basée sur DL Learner, un framework pour l'apprentissage de concepts en logique de description à partir d'exemples fournis par l'utilisateur. Les ontologies sont utilisées comme intermédiaire pour expliquer le comportement d'entrée-sortie des réseaux de neurones artificiels formés.
Les auteurs dans [49], ont proposé une approche permettant de regrouper sémantiquement les features des modèles d'IA en utilisant une ontologie pour ajouter une description sémantique pour eux.
De Sosa et al. [50] ont exploité les ontologies et les classifieurs du modèle de réseau de neurones pour relier ses états internes aux concepts de l'ontologie. Ainsi, les auteurs visent à établir le mapping pour comprendre le comportement interne du modèle.
Nous avons également examiné l'ontologie d'explication [51] qui fournit une représentation structurée de plusieurs types d'explication et modélise le rôle des explications, en tenant compte des attributs du système et de l'utilisateur dans le processus.

Analyse & discussion

Après avoir analysé les travaux présentés, nous avons identifié les différentes techniques utilisées dans les travaux examinés. En conséquence, notre analyse révèle que le

TABLE 2.4 – Comparaison des travaux examinés basés sur la sémantique pour une intelligence artificielle eXplicable

Article	**Techniques utilisées**	**Champ d'application**	**Types d'explication**	**Utilisateurs visé**	**Objectif**
TREPAN, 2020 [46]	Modèle de substitution	Globales	Contrastive/ contextuelle	Statique	Une approche qui explique l'algorithme TREPAN
Doctor XAI, 2020 [47]	Extraction de règles et perturbation des features	Locales	Contrefactuel	Statique	Une approche qui explique le modèle Doctor AI
DL Learner, 2017 [48]	Matching	Globales	Quotidienne	Statique	Une approche qui explique les images annotées
Semantic Clustering, 2020 [49]	Clustering	Locales	Contextuel	Statique	Une approche pour le clustering des caractéristiques
De sousa et al., 2021 [50]	Matching	Globales	Basé sur les traces	Statique	Approche qui extrait des règles d'explication
Explanation Ontology, 2020 [51]	Matching	Locales	Contrastive/ quotidienne/ contextuelle	Multi-user	Ontologie qui pourrait être instanciée pour l'explication des modèles d'IA.

matching avec des ontologies externes est la technique la plus utilisée pour l'explication des modèles d'IA. [48] effectue le matching avec le framework DL Learner pour apprendre le comportement entrée-sortie, tandis que [50] s'appuie sur le matching avec l'ontologie externe pour extraire les règles d'explication. Quant à [51], l'ontologie des explications pourrait être utilisée pour faire un matching afin d'enrichir sémantiquement les mécanismes qui génèrent des explications (par exemple, des modèles de substitution pour la simplification). D'autre part, nous avons identifié des travaux qui ont utilisé des ontologies pour dériver un modèle de substitution, comme [46]. TREPAN Reloaded [46] étend l'algorithme TREPAN à l'aide d'ontologies en créant des arbres de décision plus compré-

hensibles. Plus précisément, l'ontologie permet de déterminer les features les plus compréhensibles pour un utilisateur et de leur attribuer une priorité dans le processus de génération de l'arbre. Quant à Doctor XAI [47], ce travail utilise une ontologie pour extraire les règles et la perturbation des caractéristiques de manière sémantique. En ce qui concerne la portée de l'explication, les travaux examinés couvrent différentes portées, telles que locale [47], [49], [51], et globale [46], [48], et [50]. Les explications locales visent à expliquer une instance de modèle, tandis que les explications globales illustrent le fonctionnement du modèle d'IA et tentent de l'imiter. Les travaux portant sur ces deux types d'explications ont prouvé leur efficacité. En ce qui concerne les types d'explication, la plupart des approches examinées n'en proposent qu'un seul. En effet, les types d'explication vont du contrefactuel [47], au quotidien [48], au contextuel [49], [50].

En revanche, seuls TREPAN Reloaded [46] et l'ontologie d'explication [51] proposent différents types d'explication. Le premier a proposé des explications contrastives et contextuelles, tandis que le second combine des explications contrastives, quotidiennes et contextuelles. Notre analyse porte également sur les utilisateurs cibles et révèle que la plupart des travaux examinés fournissent des explications pour un seul utilisateur. Nous avons remarqué que seule l'ontologie d'explication prend en compte plusieurs rôles d'utilisateur pour l'explication. Cependant, l'ontologie d'explication examinée ne fournit pas d'explications dynamiques puisqu'elle prédéfinit à l'avance les liens entre les utilisateurs et les explications. Compte tenu des limites des approches présentées, il est essentiel de proposer une solution qui les surmonte en prenant en compte plusieurs rôles d'utilisateurs et en fournissant des explications différentes de manière dynamique et adaptative en fonction des différents rôles et besoins des utilisateurs.

Dans notre contexte, nous visons à recommander des mesures pour aider à la gestion des crises en tenant compte de plusieurs facteurs contextuels (par exemple, le pays, le taux de reproduction du virus, etc.). De plus, nous considérons les préférences de l'utilisateur pour lui proposer une explication convenable. Notre apport s'appuie sur deux modèles pour la recommandation. En effet, nous nous appuyons sur des techniques d'apprentissage profond pour concevoir un modèle basé sur le contenu pour la recommandation des mesures permettant d'aider à gérer les crises, puisqu'il surpasse les techniques de recommandation traditionnelles (par exemple, le modèle d'espacement des vecteurs)[52]. En outre, nous visons à concevoir un modèle sémantique qui explique les résultats du premier modèle. Ce dernier modèle repose sur le filtrage collaboratif. Nous adoptons le filtrage collaboratif car notre contribution est dédiée à plusieurs rôles d'utilisateurs qui peuvent avoir des intérêts similaires en termes d'explications. Par conséquent, l'identification et l'exploration des intérêts d'utilisateurs similaires permettent d'identifier de nouvelles préférences d'utilisateurs. Comme nos modèles génèrent plusieurs résultats (c'est-à-dire des recommandations et des explications) pour l'utilisateur, nous combinons les deux approches de filterage collaboratif (collaborative filtering) et la recommendation à base de contenu (content-based) pour constituer une approche de recommandation hybride. En effet, les résultats générés par le modèle de recommandation de santé servent d'entrée au modèle d'explication sémantique.

Par conséquent, nous surmontons les faiblesses des approches basées sur le contenu et le filtrage collaboratif en proposant une contribution hybride. Compte tenu du processus de recommandation séquentiel, nous adoptons l'hybridation en pipeline pour concevoir notre contribution à la recommandation des mesures avec une explication sémantique.

Conclusion

Ce chapitre a présenté les concepts et les technologies qui sous-tendent notre recherche en donnant un aperçu sur les approches de prédiction, notamment des épidémies, qui impliquent les différentes technologies d'intelligence artificielle. Ce chapitre couvre également les concepts techniques impliqués dans nos contributions, comme l'intelligence artificielle explicable, les systèmes multi-agents et les data lakehouses, que nous adoptons comme référentiel de données.

Après avoir présenté et survolé les concepts de base, nous avons présenté une exploration et une analyse détaillées de l'état de l'art concernant les sous-problèmes de notre travail de recherche, tels que la curation, la prédiction des crises et la recommandation de mesures pour la gestion des crises. Tout d'abord, nous avons illustré une étude sur la curation sous différents perspectives, à savoir la curation des données, des métadonnées et des schémas. Nous avons identifié plusieurs tâches pour la curation de données, telles que la contextualisation, la liaison avec des bases de connaissances externes, la réparation de données, l'extraction de schémas et de métadonnées, etc. Les contributions examinées s'appuient sur plusieurs techniques, telles que l'apprentissage automatique, les technologies sémantiques (c'est-à-dire les graphes de connaissances, les ontologies, etc.), les techniques basées sur les graphes, le crowdsourcing, etc. En outre, ils traitent différentes structures de données, à savoir des données structurées, semi-structurées et non structurées. Malgré la diversité des techniques et des objectifs des approches examinées, nous avons remarqué que la plupart d'entre elles ne considèrent que les données batch et ne traitent pas les données en streaming. Elles présentent également un comportement statique lors de la curation des données. En effet, elles ont besoin de plus de flexibilité et d'adapter leurs processus de curation en fonction du contexte décisionnel et des exigences de l'utilisateur. Par conséquent, elles ne permettent pas d'organiser et de réorganiser le processus de curation de manière dynamique. Nous avons également constaté que la généralisation des tâches de curation pose un problème, en particulier pour les approches basées sur des règles et sur l'apprentissage automatique. Notre étude montre également que certains travaux de curation ne sont pas totalement autonomes et nécessitent une intervention humaine, qui peut être longue et sujette à des erreurs.

Dans un deuxième temps, le chapitre étudie et met en évidence les différents travaux relatifs à la prévision des épidémies et des maladies et aux recommandations de mesures pour aider à la gestion des crises. Néanmoins, la plupart des travaux examinés se limitent à la prédiction des maladies et n'ont pas proposé de surveiller et de gérer la situation (par exemple, une épidémie) en recommandant en permanence des mesures sanitaires. En

outre, les approches examinées ne tiennent pas compte des différentes préférences ou rôles des utilisateurs afin d'adapter et d'aligner leurs résultats en fonction de leurs besoins. Nous avons également précisé que ces contributions n'ont pas proposé d'explications pour l'utilisateur final, comme nous l'avons souligné dans le chapitre précédent. Cependant, elles revêtent une grande importance dans des domaines critiques tels que celui de la santé. À cette fin, nous avons examiné les travaux existants sur l'intelligence artificielle explicable. Nous avons identifié deux approches principales : l'interprétabilité des modèles d'IA et les méthodes d'explication post-hoc. La première approche impose un compromis entre la précision et l'interprétabilité du modèle d'IA, ce qui peut affecter ses performances. En revanche, la méthode post hoc n'a pas d'incidence sur ces performances. Suite à l'analyse des méthodes post-hoc existantes, nous avons remarqué que seuls quelques travaux s'appuient sur les technologies sémantiques pour générer des explications. Néanmoins, ces dernières englobent plusieurs techniques susceptibles d'améliorer la sémantique des explications fournies à l'utilisateur. Nous avons également constaté que les approches examinées ne fournissent pas plus d'un type d'explication, ce qui peut ne pas être adapté aux besoins des multiples utilisateurs.

Notre travail a pour but de proposer un système qui permet de recommander des mesures sanitaires pour gérer les crises à l'aide de données recueillies auprès de différentes sources, tout en tenant compte des préférences des utilisateurs dans divers contextes décisionnels. Les contributions ci-dessus sont présentées et discutées dans les chapitres suivants.

CHAPITRE 3

Curation adaptative des données pour les données batch et streaming

3.1 Introduction

Ce chapitre présente notre approche pour la curation adaptative des données batch et streaming. L'hétérogénéité et la complexité des données restent des défis critiques pour les big data. Le nettoyage des données est donc nécessaire pour améliorer la qualité des données avant leur analyse ou leur visualisation. À cette fin, la curation des données assure la gestion et la promotion de l'utilisation des données à partir de leur point de création en les enrichissant ou en les mettant à jour pour qu'elles restent adaptées à un usage spécifique. Elle fournit davantage d'informations sur la provenance des données, le contexte original de mesure et d'utilisation, et l'objet de l'observation afin de faciliter la réutilisation des données [16].

Néanmoins, les approches de curation de données existantes ne sont plus suffisantes pour curer les données de big data multi-structurées collectées à partir de sources multiples utilisant différents modes d'ingestion (c.-à-d. les données par lots et en continu) [53]. En outre, le processus de curation des données peut être affecté par des facteurs tels que les caractéristiques de la source de données et le contexte décisionnel. En effet, les contextes décisionnels critiques, tels que les crises, sont généralement évolutifs et imposent des restrictions sur le temps d'exécution et la précision des résultats des systèmes d'information. Il est donc primordial de prendre en compte les caractéristiques des données et le contexte d'utilisation pour identifier et réaliser le processus de curation de données adéquat [54]. En outre, la valeur des données n'est jamais stabilisée, car leur sémantique évolue constamment, ce qui oblige à réorganiser et à modifier le processus de curation des données au fil du temps [55]. Par conséquent, la curation de données doit prendre en compte des caractéristiques changeantes du processus de décision afin d'optimiser la qualité des résultats du

processus de décision et son temps d'exécution, et de s'aligner sur les attentes des utilisateurs. Il est donc difficile d'identifier les tâches de curation de données pratiques et de les réorganiser en fonction des caractéristiques de la source de données, du contexte décisionnel et des attentes de l'utilisateur.

Comme présenté dans le chapitre précédent, la plupart des approches existantes de curation de données sont statiques et ne prennent pas en compte les caractéristiques du contexte décisionnel mentionnées ci-dessus. Ces dernières handicapent les décideurs (c'est-à-dire ceux qui sont confrontés à des situations critiques) qui souhaitent prendre des décisions rapidement et efficacement. En outre, certaines approches existantes nécessitent une intervention humaine, ce qui peut prendre du temps et être source d'erreurs. Néanmoins, l'aspect statique de l'étape de curation des données peut constituer un handicap pour les étapes du processus décisionnel, telles que l'intégration et l'analyse des données. Par exemple, les besoins des parties prenantes peuvent varier en fonction du contexte décisionnel. Nous supposons qu'Alice, haut fonctionnaire adjoint chargé de la défense et de la sécurité au ministère de la santé, et Bob, spécialiste des maladies infectieuses, utilisent un système pour obtenir des recommandations sur la gestion des crises sanitaires. Ce système recueille des données multi-structurées (c'est-à-dire des données structurées, semi-structurées et non structurées) en mode batch et en mode streaming. Supposons qu'Alice utilise le système dans un contexte de crise tandis que Bob l'utilise dans une situation ordinaire. Ils peuvent donc avoir des besoins différents en ce qui concerne la précision des résultats et le temps de réponse du système. En effet, le temps de réponse peut être important pour Alice, alors que la précision des résultats dans le cas de Bob est moins critique. La curation des données dans le cas d'Alice sera donc différente de celle de Bob en termes de contexte décisionnel et de besoins de l'utilisateur. En conséquence, la curation des données doit tenir compte du contexte décisionnel et des exigences fonctionnelles et non fonctionnelles de l'utilisateur qui peuvent avoir un impact sur la qualité des résultats, comme la précision, le temps de réponse, etc. Notre objectif est donc de concevoir une solution permettant d'effectuer une curation de données adaptative pour les données multistructurées en batch et en streaming tout en tenant compte des exigences susmentionnées.

Le présent chapitre est organisé comme suit. Tout d'abord, nous présentons dans la section suivante l'idée générale de l'approche que nous proposons pour la curation adaptative des données en batch et en streaming. En particulier, nous présentons l'ontologie proposée pour la caractérisation des sources de données et l'évaluation de la qualité. Ensuite, nous détaillons notre approche proposée, Adaptive CUration Service Composition (ACUSEC), pour la composition de services de curation adaptatifs. Enfin, nous présentons la mise en œuvre et les expériences élaborées pour évaluer notre proposition.

3.2 Idée générale

Nous proposons une approche basée sur les services, nommée ACUSEC, pour la curation adaptative des données en batch et en streaming en fonction des besoins fonctionnels

et non fonctionnels de l'utilisateur, comme ses préférences et ses exigences, son contexte de décision et la qualité des services de curation. Comme le montre la figure 3.1, nous effectuons la curation des données après avoir évalué la qualité des données et des sources. Pour cela, nous proposons une ontologie qui joue un double rôle en évaluant la qualité des données et des sources de données à travers différentes dimensions et en dégageant les caractéristiques des données qui guident le processus de curation des données. En effet, les caractéristiques de la source de données peuvent avoir un impact sur la sélection des services de curation, ce qui influence, par conséquent, la composition des services globaux. Quant aux données en streaming, nous soulignons le fait que nous traitons les besoins en termes de leur curation (c.-à-d., les caractéristiques identifiées et les services correspondants) qui est différente de la gestion des données en streaming (c.-à-d., organisation du processus de collecte et de contrôle des streams). Ainsi, nous utilisons ACUSEC et l'ontologie proposée pour constituer un nouveau framework de curation de données, que nous détaillons ci-après.

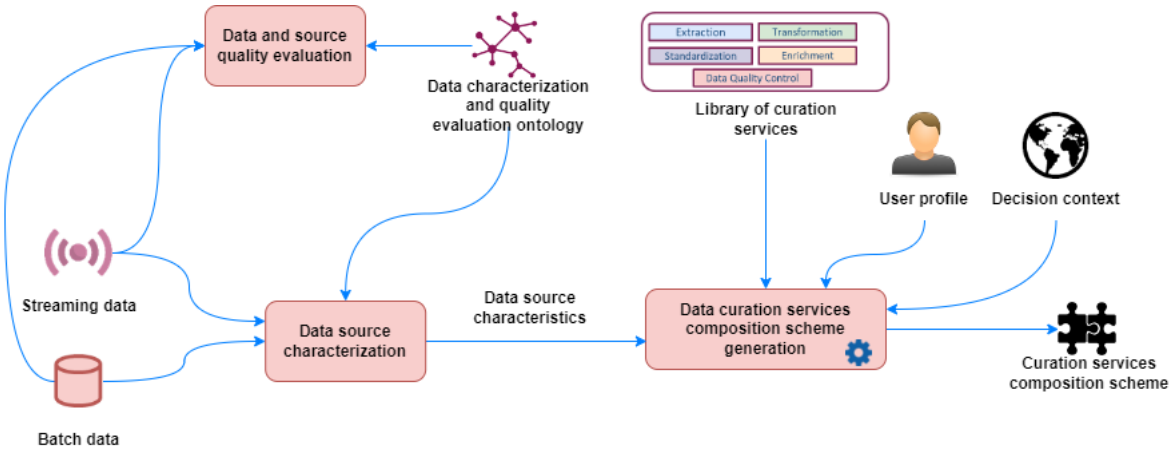

FIGURE 3.1 – Vue d'ensemble de la composition du service de curation adaptative

3.3 Une ontologie pour la caractérisation et l'évaluation de la qualité des sources de données

L'ontologie proposée assure un double rôle. Premièrement, elle évalue la qualité de la source de données pour juger de la nécessité d'une curation des données. Deuxièmement, elle guide l'approche ACUSEC proposée dans le processus de composition des services de curation en identifiant les principales caractéristiques de la source de données qui ont un impact sur la composition. À cette fin, nous avons adopté la méthodologie AOM [56] pour proposer une ontologie modulaire, nommée **DAta chaRacterization Quality evAluation oNtology (DARQAN)**, décrivant les sources de données selon différentes perspectives (par exemple, la provenance, la description technique de la source de données, etc.). Comme le montre la figure 3.2, notre ontologie comprend quatre modules, à savoir

les modules de description des sources de données, de qualité des données, de provenance et de plateforme de collecte de données, que nous détaillons ci-après :

- **Module de description des données :** fournit plusieurs types d'informations, comme des informations sur les données, telles que la période et le lieu d'observation, le système linguistique, les différentes formes de données de la source et des informations sur le fournisseur. Ce module comprend également des informations techniques sur le format de données dans lequel l'ensemble de données est fourni (distribution) (par exemple, un ensemble de données XML, un fichier de texte brut, une base de données SQL, etc.). Le format des données peut être combiné avec des propriétés de données telles que l'URL pour accéder à une base de données MySQL, par exemple, le nom d'utilisateur et le mot de passe. En plus de ces caractéristiques, la curation des données repose principalement sur des informations que nous présentons sous forme de propriétés des classes du module de description de la source de données. Ces éléments caractérisant une source de données sont :
  - Le format de la source de données (structurée (S), semi-structurée (SS) ou non structurée (US))
  - La source de données inclut-elle une URL dans ses valeurs de données ?
  - La source de données doit-elle être convertie dans un autre format ?
  - La source de données doit-elle faire l'objet d'un processus de balisage du PoS ?
  - S'agit-il de données en streaming ?
- **Module d'évaluation de la qualité des données :** évalue la qualité des données et des sources de données par le biais d'un raisonnement basé sur plusieurs facteurs de qualité tels que les dimensions de la qualité, les standards, les certificats, les politiques de qualité et le feedback de l'utilisateur sur la qualité des données. Pour concevoir le module d'évaluation de la qualité des données, nous avons adopté et réutilisé les normes proposées par le W3C, telles que [57, 58, 59, 60, 61]. En étudiant les standards présentés dans [62] et en considérant le contexte du présent travail, nous nous sommes appuyés sur plusieurs dimensions de la qualité des données pour évaluer la qualité de la source et des données sous différentes perspectives. En tenant compte des besoins du domaine et des usages, nous nous concentrons principalement sur l'évaluation de la qualité des données selon les dimensions suivantes :
  - La précision des données et des sources : il est primordial de vérifier la précision des données et la pertinence des sources.
  - Les dimensions de précision liées au temps : Dans notre contexte, l'aspect temporel est crucial pour vérifier la validité temporelle des données, ce qui affecte la fiabilité de la prédiction. Comme les data lakehouses continuent d'ingérer des sources de données brutes, celles-ci peuvent continuer à évoluer et certaines données peuvent devenir obsolètes. Ce qui impose la vérification de ladite validité.
  - La confiance : pour accorder la confiance à une source de données, nous devons vérifier sa réputation en tant que source de données ainsi que la réputation de son éditeur. Nous pouvons poser des questions telles que "Pouvons-nous croire son contenu ?", "Qui est le fournisseur de cette source de données ?", "Quel est

l'avis des utilisateurs ?"

— **Module de provenance :** basé sur la logique proposée dans [58], il retrace les origines d'une donnée. Ce module repose sur trois concepts principaux : entité, activité et agent. La classe Agent représente l'agent qui gère les activités (par exemple, Agent logiciel, Personne et Organisation). Un agent peut agir au nom d'un autre agent et utiliser une entité ou assurer une activité. La classe Activité est conçue pour illustrer les activités qui conduisent à la génération des données. Ces activités peuvent à leur tour être associées à des agents et utiliser des entités. Chaque activité peut avoir des dates de début et de fin et peut transmettre des informations à d'autres activités. Quant à la classe entité, elle décrit les entités traitant des unités de données (par exemple, un capteur). Une entité peut être dérivée d'une autre entité et peut être générée à partir d'une activité. Par exemple, nous pouvons citer la classe Collection qui regroupe des entités (par exemple, la collection Réseau de capteurs regroupe les entités *Sensor*).

— **Module plateforme :** décrit la plateforme à partir de laquelle les mesures (par exemple, la fréquence cardiaque) peuvent être collectées. En effet, une plateforme, telle que les smartphones et les satellites, peut héberger des entités telles que des capteurs. Ainsi, ces derniers sont présentés par des instances de la classe entité (i.e., comme celle présentée dans le module provenance). Chaque capteur est caractérisé par une propriété observable et une caractéristique d'intérêt. Une caractéristique d'intérêt décrit l'élément dont la propriété est estimée ou prise en compte dans l'observation pour obtenir un résultat. D'autre part, la classe des propriétés observables décrit une qualité observable d'une caractéristique d'intérêt.

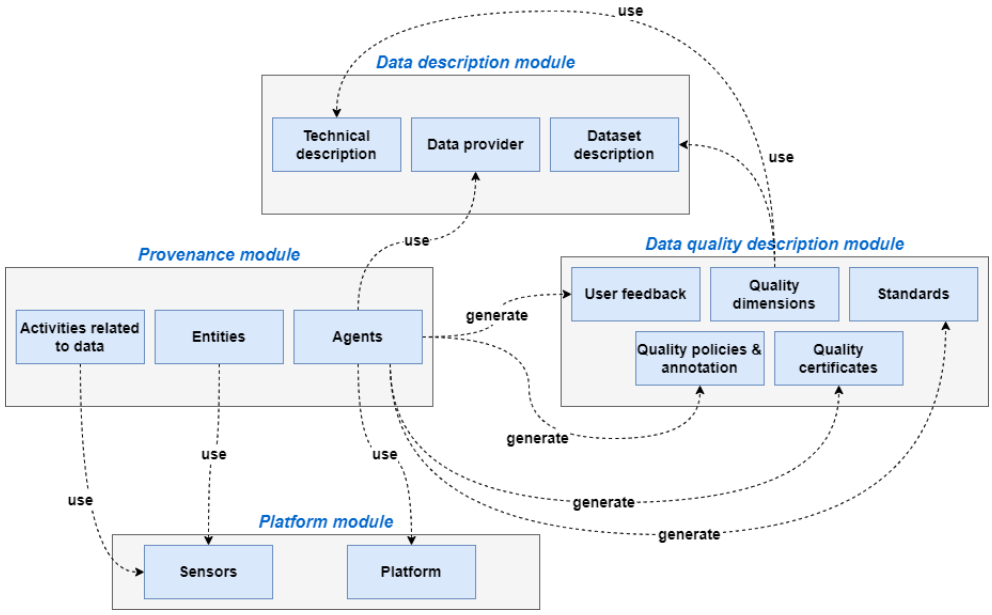

FIGURE 3.2 – Les modules de caractérisation et d'évaluation de la qualité des données

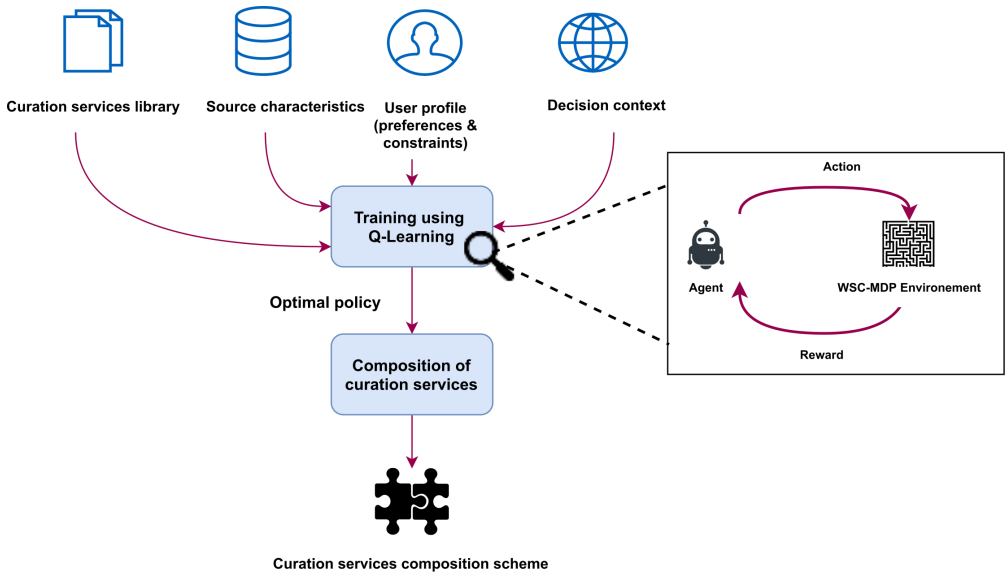

FIGURE 3.3 – Processus d'apprentissage pour la composition de services adaptatifs

3.4 ACUSEC : Adaptive CUration SErvice Composition

Nous présentons dans cette section notre approche permettant de composer les services de curation en impliquant les caractéristiques identifiées. Comme susmentionné, le contexte et les besoins de l'utilisateur doivent être pris en compte tout au long du processus de décision. De ce fait, les tâches de curation de données doivent être organisées dans un ordre bien spécifique en tenant compte de ces facteurs. Ainsi, il est important d'identifier et de sélectionner les services de curation de données utiles pour les données multistructurées collectées en batch et en streaming et de les composer d'une manière adaptative en fonction des caractéristiques de la source de données, des préférences de l'utilisateur, des contraintes et du contexte décisionnel.
À cette fin, nous avons conçu une bibliothèque à partir de services de curation existants [29, 63]. Ensuite, nous nous sommes appuyés sur le paradigme de l'apprentissage par renforcement pour procéder à la composition des services de curation. Comme le montre la figure 3.3, notre contribution, ACUSEC, est une approche en deux étapes. Tout d'abord, ACUSEC utilise l'apprentissage par renforcement pour apprendre les compositions possibles de services de curation en fonction des exigences fonctionnelles et non fonctionnelles de l'utilisateur. En conséquence, la phase d'apprentissage identifie la politique optimale (c'est-à-dire l'ensemble des schémas de composition) pour effectuer la composition des services de curation. La seconde étape consiste à composer les services de curation en utilisant la politique apprise. Dans ce qui suit, nous décrivons les éléments qui constituent notre approche, tels que la bibliothèque de services de curation, les processus d'apprentissage et de composition.

3.4.1 Conception d'une bibliothèque de services de curation

L'approche ACUSEC (Adaptive CUration SErvice Composition) proposée assure une curation de données adaptative, tout en tenant compte du contexte et ciblée sur l'utilisateur. Pour ce faire, elle utilise une bibliothèque de services de curation comme illustré dans la figure 3.4. La bibliothèque proposée regroupe quatre catégories de services définies dans cette bibliothèque : les services d'extraction, d'enrichissement, de contrôle de la qualité des données et de normalisation des données.

— **Catégorie des services d'extraction** englobe les services qui assurent les tâches de NLP (Natural Language Processing) telles que l'extraction d'entités nommées, l'étiquetage POS et l'extraction de stems des mots. L'intégration de tâches de traitement du langage naturel permet d'extraire des caractéristiques qui peuvent être intégrées dans un processus d'enrichissement en vue d'une analyse de données ultérieure.

— **Catégorie des services d'enrichissement** regroupe les services d'enrichissement sémantique des données par le biais de mécanismes d'annotation basés sur les connaissances et d'extraction de similarités entre les termes. L'enrichissement des données d'entrée est pratique pour la gestion des données, car il permet d'améliorer les résultats générés [64].

— **Catégorie des services de contrôle de la qualité des données** contient deux services qui assurent la détection des valeurs manquantes et des anomalies de données (par exemple, l'écart de valeur). Les services de cette catégorie procèdent au nettoyage des données afin d'en améliorer leurs qualité.

— **Catégorie des services de normalisation des données** englobe les services qui unifient les données à l'aide d'une base de connaissances. Cette dernière est utilisée comme modèle de référence contenant des paramètres décrivant les variables avec leurs types et leurs intervalles pour unifier la représentation des données.

3.4.2 Composition de services de curation basée sur l'apprentissage par renforcement

Nous nous appuyons sur l'apprentissage par renforcement pour générer des schémas de compositions tout en tenant compte des changements des besoins qui peuvent impacter la composition. En effet, ce type d'apprentissage permet de s'adapter à la diversité des situations possibles sans reconstruire un modèle d'apprentissage à chaque situation [65]. Plus précisément, nous nous appuyons sur l'algorithme Q-Learning, l'un des algorithmes les plus populaires pour l'apprentissage par renforcement, afin d'apprendre le schéma optimal de composition du service de curation de manière adaptative. L'algorithme Q-Learning est un algorithme de renforcement qui définit un agent interagissant avec un environnement, généralement défini comme un processus de décision de Markov, afin d'apprendre les actions optimales pour effectuer une transition d'un état à un autre. Au cours du processus d'apprentissage, l'agent apprend les poids de transition attribués aux actions de transition. Ces poids représentent les récompenses accumulées après chaque transition. Par conséquent, nous traitons la composition du service de curation comme un problème de maxi-

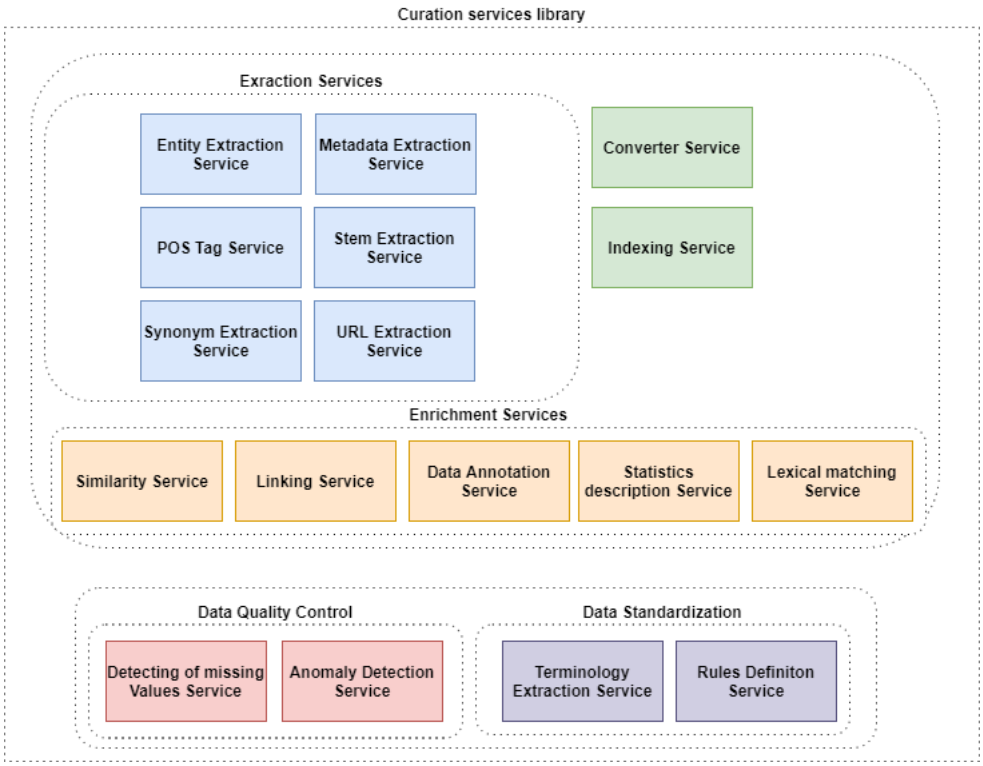

FIGURE 3.4 – Bibliothèque de services de curation

misation du gain qui vise à maximiser la récompense globale. Dans la section suivante, nous décrivons formellement l'environnement conçu comme un processus de décision de Markov et utilisé pour effectuer l'apprentissage par renforcement.

Processus de décision de Markov

Un processus de décision de Markov est un modèle stochastique dans lequel un agent prend des décisions et les résultats de ses actions sont aléatoires. Comme le montre la figure 3.5, nous représentons les services de curation dans un processus de décision de Markov (MDP) dans lequel chaque action de transition présente un service de curation. Ainsi, dans l'environnement du processus de décision de Markov, nous présentons toutes les compositions possibles et valides de tous les services de curation pour tous les types de sources de données, quels que soient les besoins de l'utilisateur et les facteurs environnementaux. Au cours de la phase d'apprentissage, l'agent d'apprentissage explore et exploite l'environnement pour identifier la composition optimale du service de curation en fonction des exigences fonctionnelles et non fonctionnelles.

Processus d'apprentissage basé sur l'apprentissage par renforcement

Nous proposons l'équation 3.1 pour calculer les récompenses de transition. L'équation proposée repose sur la qualité de service des services de curation, les préférences des utilisateurs et les contraintes (par exemple, la valeur du temps de réponse de la qualité de

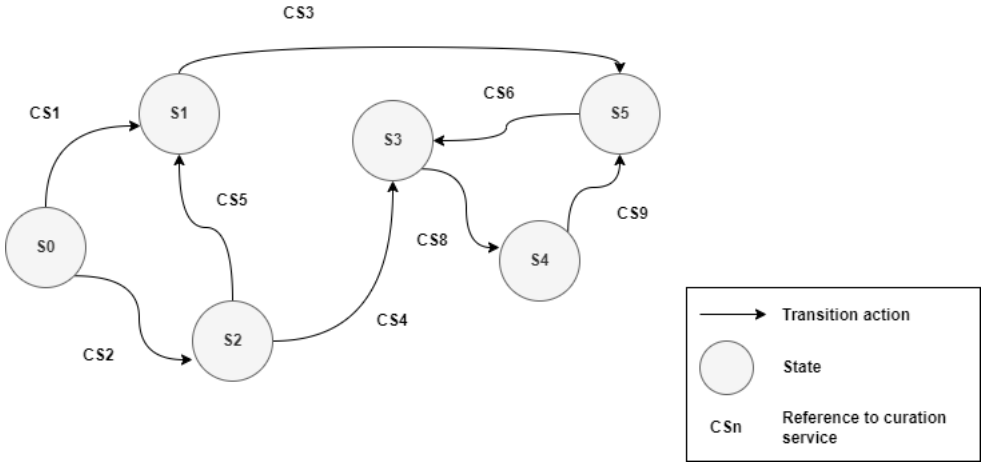

FIGURE 3.5 – Exemple de processus décisionnel de Markov dans lequel chaque action fait référence à un service de curation

service est supérieure à 90 %). La qualité de service (QoS) est une mesure qui permet d'évaluer dans quelle mesure un service est utile à l'utilisateur final. L'algorithme de formation applique ladite équation pour calculer la récompense de transition en fonction des dimensions de la QoS souhaitées.

$$R(s) = \underbrace{\frac{\sum_{i=1}^{m} X(i) - 1 + \phi}{\sum_{i=1}^{m} |X(i) - 1 + \phi|}}_{\text{Part1}} * \underbrace{\sum_{k=1}^{m} w_k * D_k}_{\text{Part2}} \tag{3.1}$$

$$X(k) = \frac{\sum_{i=1}^{m} |D_k - M_k + \phi|}{D_k - M_k + \phi} \tag{3.2}$$

où :

- **w** représente les préférences de l'utilisateur à l'égard d'une qualité de service, définies par un poids allant de 0 à 1
- **D** est une valeur normalisée de l'évaluation de la dimension de la qualité de service, comprise entre 0 et 1
- **M** représente un seuil minimal fixé par l'utilisateur en matière de qualité de service, qui doit être respecté pour que le service puisse être invoqué. La valeur de M est comprise entre 0 et 1.
- ϕ est une valeur de normalisation qui doit être strictement supérieure à 0 et inférieure à 1

Nous avons adopté les préférences de l'utilisateur comme poids pour favoriser une dimension de qualité de service par rapport à une autre. En outre, l'algorithme d'apprentissage prend en compte les contraintes définies sur les valeurs de QoS en fixant un seuil minimum M qui doit être satisfait pour invoquer un service. Compte tenu de la qualité de service, des préférences des utilisateurs et des contraintes, la fonction de récompense renvoie une valeur positive lorsque toutes les contraintes des utilisateurs sont satisfaites. Dans le cas contraire, la fonction renvoie une valeur négative. La composition des services étant

un problème de maximisation des gains, les récompenses négatives empêchent l'agent de choisir des services de curation qui ne correspondent pas aux contraintes de l'utilisateur. La première partie de l'équation 1 calcule la différence entre les contraintes imposées par l'utilisateur et les valeurs de qualité de service. Elle renvoie soit 1 si toutes les contraintes de l'utilisateur sont satisfaites, soit -1 dans le cas contraire. L'équation 3.1 s'appuie sur l'équation 3.2 pour calculer la différence entre une dimension de qualité de service et le seuil M défini par l'utilisateur. Ensuite, la valeur de l'équation 3.2 est normalisée à -1 ou 1 en fonction de la valeur obtenue. La deuxième partie permet d'attribuer les préférences de l'utilisateur aux dimensions de qualité de service. Par conséquent, les préférences sont définies comme des poids pour multiplier les valeurs des dimensions de qualité de service évaluées. Ensuite, la multiplication des deux parties de l'équation permet d'obtenir la valeur de la récompense en fonction des préférences de l'utilisateur, des contraintes et des valeurs de qualité de service.
À la fin du processus de formation, nous obtenons la politique optimale $\pi*$ qui représente la table Q finale. Après avoir réalisé les étapes du processus d'apprentissage et de composition, l'agent d'apprentissage utilise la politique optimale pour retrouver le schéma de composition du service de curation le plus approprié en choisissant la combinaison d'actions qui maximise le gain global. Après avoir identifié le schéma optimal de composition des services de curation, la curation des données est effectuée en invoquant des services dans le schéma de composition.

3.5 ACUSEC : Implémentation

Afin de déployer l'approche, nous avons conçu et developpé un framework en adoptant une architecture orientée services caractérisée par sa fiabilité, évolutivité et faible couplage. Nous visons à optimiser les étapes du processus de gestion des données qui dépendent de la curation de données en termes de temps d'exécution et d'alignement avec les besoins de l'utilisateur. Comme le montre la figure 3.6, notre framework comprend les quatre couches suivantes : la collecte des données, le contrôle de la qualité des données, le traitement des données et la curation des données. **La couche de collecte des données** ingère des sources de données en streaming et en batch ainsi que des informations sur les données en streaming, les fournisseurs de données, l'emplacement et les informations temporelles en tant que métadonnées. Ainsi, notre framework assure la capacité d'adaptation depuis le moment de la collecte des données jusqu'à la génération du pipeline de curation. Ensuite, le framework évalue la qualité des données collectées en batch et en streaming via le module d'évaluation de la qualité des données et le module de surveillance du streaming de données. À cette fin, le module d'évaluation de la qualité des données évalue la qualité des données et la qualité de la source de données. Il évalue donc les dimensions de la qualité, notamment l'exactitude, l'actualité, la crédibilité, la vérifiabilité et la réputation des données. En fonction de la qualité des données, le framework de curation des données détermine si la source de données a besoin d'être curée. La curation des données est effectuée lorsque l'une des dimensions de la qualité des données évaluées est inférieure à un seuil β, que l'utilisateur peut définir. Après l'évaluation des données, les valeurs des

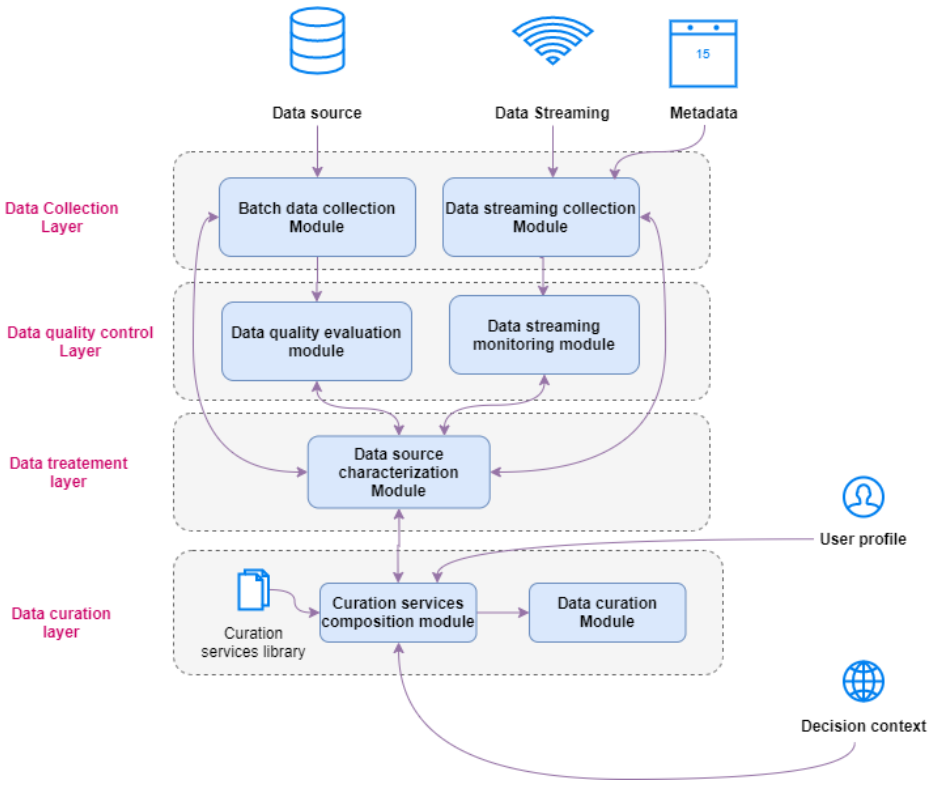

FIGURE 3.6 – Framework de curation de données adaptative

dimensions de qualité des données et la source de données sont transmises au module de caractérisation des données, que nous définissons dans la couche de traitement des données. Le module de caractérisation de la source de données extrait les caractéristiques de la source de données nécessaires au processus de curation des données, telles que le format et le type de la source de données, ainsi que les tâches spécifiques de curation des données.

Sur la base des caractéristiques extraites, du profil de l'utilisateur et du contexte décisionnel, **la couche de curation des données** sélectionne les services de curation les plus appropriés dans la bibliothèque de services de curation pour constituer le schéma de curation des données. Chaque service assure une tâche de curation spécifique (par exemple, suppression des doublons, détection des anomalies, etc.). Ces tâches de curation peuvent également être combinées de manière spécifique pour curer une source de données. Dans la section suivante, nous évaluons les performances et l'efficacité de nos contributions.

3.6 ACUSEC : Evaluation

Afin d'évaluer notre approche et le framework developpé, nous nous sommes focalisés sur l'aspect evolutif en termes de (1) nombre d'utilisateurs, (2) nombre de services, (3) son adaptativité et son alignement sur les besoins de l'utilisateur ainsi que (4) l'efficacité du processus de curation. À cette fin, notre protocole expérimental s'appuie sur trois sources

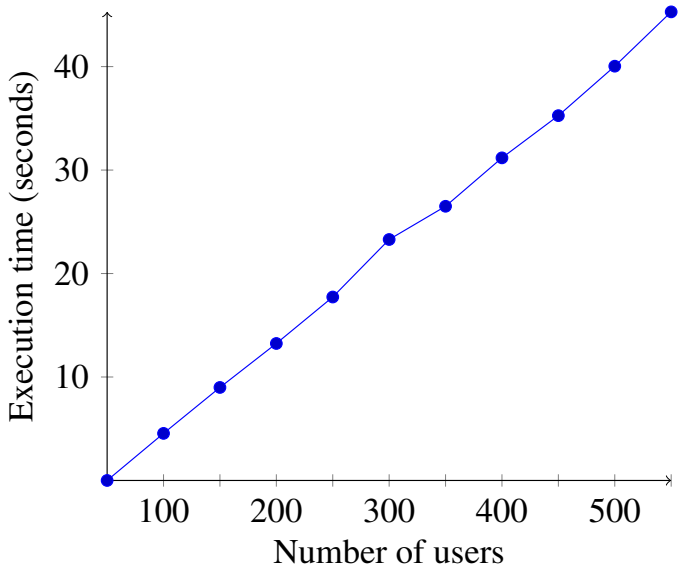

FIGURE 3.7 – Temps d'exécution par nombre d'utilisateurs

de données différentes, à savoir une source de données non structurées [1], une source de données semi-structurées [2] et une source de données structurées [3]. Nous avons comparé les performances de notre méthode de composition de services de curation avec celles des méthodes First-visit Monte Carlo et Temporal-difference Learning [66],
Pour évaluer l'évolutivité de notre méthode de composition de services de curation en fonction du nombre d'utilisateurs qui l'utilisent simultanément et du nombre de services de curation. Nous avons utilisé le multithreading pour créer un environnement de simulation de la composition du service de curation, qui peut être exécuté par plusieurs utilisateurs simultanément. À cette fin, nous avons développé un ensemble de threads, où chacun simule une demande de composition de service de curation par un utilisateur. De plus, afin d'obtenir des conditions similaires pendant l'expérimentation, nous avons défini les mêmes paramètres d'entrée (c'est-à-dire les préférences de l'utilisateur, les contraintes, etc. Ensuite, nous avons progressivement exécuté et augmenté le nombre de threads afin d'examiner le temps de réponse de notre proposition à ces requêtes. La figure 3.7 illustre le temps d'exécution global en fonction du nombre de threads. En ce qui concerne l'évolutivité en fonction du nombre de services, nous avons simulé l'augmentation du nombre de services en augmentant la taille de la table Q. Chaque entrée de la Q-Table correspond à un service. Suivant ces expériences, nous avons remarqué que le temps d'exécution du processus global est inférieur à une seconde en utilisant une Q-Table de taille inférieure à

1. https ://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter : Une source de données qui contient des informations sur la santé provenant de plus de 15 grandes agences d'information sur la santé telles que la BBC.

2. https ://www.ncbi.nlm.nih.gov/datasets/coronavirus/genomes/ : Une source de données fourni par le National Center for Biotechnology Information (NCBI) qui contient des données sur les génomes COVID-19

3. https ://www.google.com/covid19/mobility/ : Un recueil de données qui contient des rapports sur la mobilité communautaire donnant un aperçu des changements dans la réponse aux politiques tout en luttant contre COVID-19.

4000x4000. Dans le cas contraire, le temps d'exécution peut atteindre environ 7 secondes lorsque la taille de la Q-Table est de 12000x12000. En conséquence, les expériences ont montré que le schéma de composition des services est généré en quasi temps réel en utilisant 12000 services. Avec la même configuration, les algorithmes de Monte Carlo et de différence temporelle ne peuvent pas générer un schéma de composition des services en utilisant plus de 200 services. Nous avons étendu ces expériences en évaluant les performances de l'approche de composition de services de curation que nous proposons par rapport à des benchmarks de composition de services et à des baselines qui ont prouvé leur efficacité dans la composition de services, tels que la procédure de recherche adaptative aléatoire et gourmande (GRASP) [67], la composition aléatoire [68], la colonie de fourmis [69], les k plus proches voisins (KNN) [70] et l'algorithme génétique (GA) [71]. Le tableau 3.1 compare le temps d'exécution de chaque méthode de composition de services en fonction du nombre de services.

TABLE 3.1 – Comparaison des performances de différentes méthodes de composition de services

Nombre de services	KNN	GA	GRASP	Random	Ant	Notre proposition
18 Services	2s	64s	1s	0.4s	1s	0.1s
50 Services	17s	840s	27s	0.5s	1s	0.1s
100 Services	2s	>1H	420s	1s	5s	0.1s
200 Services	>1H	-	>1H	1s	7s	0.1s
300 Services	-	-	-	1s	7s	0.2s
1000 Services	-	-	-	1s	12s	0.2s
10000 Services	-	-	-	120s	420s	5.32s
12000 Services	-	-	-	296s	540s	6.58s

Les résultats obtenus et présentés dans le tableau, prouvent l'efficacité de notre proposition. En effet, ses performances surpassent celles des méthodes de compositions existantes qui prennent dans certains cas plus d'une heure pour générer un schéma de composition des services. Egalement, ces résultats affirment la capacité de notre approche en s'adaptant aux changements des besoins sans prendre trop de temps.

3.7 Conclusion

Le chapitre présenté traite de la première question de recherche de cette thèse, qui englobe les deux sous-questions de : "Comment effectuer la curation de données pour plusieurs types de données collectées en batch et en streaming ?" et "Comment prendre en compte les besoins de différents utilisateurs dans différents contextes tout en effectuant la curation de données ?". Nous avons proposé une nouvelle approche pour la curation adaptative des données. Plus précisément, l'approche que nous proposons prend en compte les exigences fonctionnelles et non fonctionnelles de l'utilisateur pour générer un schéma de

composition du service de curation. Pour ce faire, nous avons suivi la méthodologie AOM pour concevoir une ontologie modulaire, nommée DARQAN, qui extrait les caractéristiques des sources de données et évalue leur qualité sous différents perspectives tels que la provenance des données, la plateforme utilisée pour collecter les données, les dimensions de la qualité, les standards, le feedback des utilisateurs, etc. Ensuite, notre approche s'appuie sur les caractéristiques des sources de données extraites ainsi que sur le paradigme de l'apprentissage par renforcement pour composer des services de curation à partir d'une bibliothèque de services conçue en utilisant et intégrant des services de curation existants. Nous avons également validé notre approche par le biais d'expériences évaluant l'évolutivité, l'adaptabilité aux changements et l'alignement sur les besoins des utilisateurs. Ainsi, ces expérimentations ont montré des résultats satisfaisants qui prouvent les bonnes performances et l'efficacité de nos contributions. Dans le chapitre suivant, nous attaquons le deuxième sous-problème lié à la recommandation expliquée des mesures pour la gestion des crises.

CHAPITRE 4

Vers une approche de recommandation expliquée pour la gestion de crise

4.1 Introduction

Dans le chapitre précédent, nous avons présenté l'approche de curation de données que nous proposons, qui tient compte des exigences fonctionnelles et non fonctionnelles de l'utilisateur pour générer de manière adaptative un schéma de composition du service de curation. En conséquence, elle adapte le schéma de composition en fonction de l'évolution de la situation de l'utilisateur et de son contexte décisionnel. Nous avons également proposé un protocole d'évaluation qui évalue l'efficacité de notre proposition en termes de temps d'exécution, d'adaptabilité aux changements et d'alignement sur les besoins de l'utilisateur. Après l'étape de curation des données, les données sont analysées pour prédire les situations inhabituelles qui pourraient déclencher des crises. Par la suite, il sera nécessaire de choisir et appliquer des actions appropriées pour gérer la crise en cas d'identification de maladies. En conséquence, les parties prenantes peuvent adopter ces mesures pour faire face à la situation et éviter qu'elle ne s'aggrave. Pour tackler les limites des approches existantes présentées dans le chapitre 2, ce chapitre présente notre contribution constituée de :

— un modèle basé sur l'apprentissage profond multi-sorties pour les recommandations des mesures adéquates pour répondre aux besoins des différents rôles utilisateurs. Ainsi, le modèle proposé prédit le niveau de rigueur de chaque mesure. Nous appliquons ce modèle au domaine de la santé pour recommander des mesures liées à la fermeture et à la restriction, à l'économie et aux politiques de santé, tout en tenant compte des rôles et des besoins de plusieurs utilisateurs et des caractéristiques de plusieurs pays.

— une approche sémantique pour l'explication adaptative des recommandations en tenant compte des différents rôles, préférences et contextes de décision des utilisa-

teurs. L'approche proposée construit dynamiquement une ontologie d'explication en la mappant avec des ontologies externes. L'ontologie construite englobe plusieurs explications adaptées à différents utilisateurs (pays voisins, contre-exemples, etc.). Ensuite, sur la base de techniques de factorisation matricielle, nous extrayons le sous-graphe d'explication approprié en fonction du rôle et des besoins de l'utilisateur.

4.2 Modèle de recommandation de mesures pour la gestion des crises

Nous présentons ci-après notre contribution, qui couvre un modèle de recommandation de mesures pour la gestion des crises et une approche explicative définie sur la base de ce modèle. Puis, nous détaillons dans la section suivante notre seconde contribution, qui est une approche d'explication basée sur la sémantique permettant d'expliquer les choix de notre modèle de recommandation proposé. Suite à la prédiction de l'apparition et la cause d'une crise (par exemple l'agent pathogène à l'origine du risque), la recommandation doit pouvoir, dans un premier temps, identifier le modèle de recommandation correspondant à cette crise. Ce modèle que nous proposons doit pouvoir préconiser des mesures à plusieurs rôles d'utilisateurs intégrés dans le processus de gestion des crises. Pour ce faire, nous avons adopté les mesures proposées dans l'Oxford Covid-19 Government Response Tracker (OxCGRT) [72], qui s'appuie sur l'expérience de pays antérieurs en termes de gestion de crise, en particulier d'épidémies. L'OxCGRT fournit des mesures et des paramètres normalisés qui évaluent l'efficacité des réponses du gouvernement pendant toute la période de propagation des maladies. Bien que les mesures de l'OxCGRT soient utilisées pour mesurer la rigueur des politiques de santé du gouvernement, nous avons identifié des travaux ayant utilisé ces mêmes mesures dans d'autres domaines et contextes [73], tels que les domaines environnemental [74, 75] et politique [76]. En adoptant ces mesures, nous avons conçu un modèle d'apprentissage profond à multiples sorties qui tire parti de l'expérience antérieure en termes de gestion de crises, notamment les épidémies, pour recommander de futurs actions à entreprendre. Plus précisément, nous avons entraîné ce modèle en utilisant les mesures précédemment adoptées dans plusieurs pays pour prédire l'échelle de rigueur future de chaque mesure sanitaire. En conséquence, notre modèle de recommandation prend en compte les caractéristiques du pays pour recommander des mesures sanitaires appropriées. À notre connaissance, notre modèle de recommandation est le premier modèle transnational à recommander plusieurs actions, ainsi que leur niveau de sévérité pour différents rôles d'utilisateurs. Suite à un processus de sélection des features, nous avons constitué une couche d'entrée regroupant neuf neurones, tels que le pays concerné par les mesures recommandées, l'indice de rigueur actuel, l'indice de santé de confinement actuel et l'espérance de vie, comme le montre la figure 4.1. Le modèle proposé repose sur quatre couches cachées et seize couches de sortie (c'est-à-dire en fonction du nombre de mesures sélectionnées), dans lesquelles chaque couche de sortie prédit le niveau de rigueur de chaque mesure. Par exemple, comme décrit ci-dessus, la mesure "Fermeture d'écoles" est caractérisée par une échelle de gravité à trois niveaux

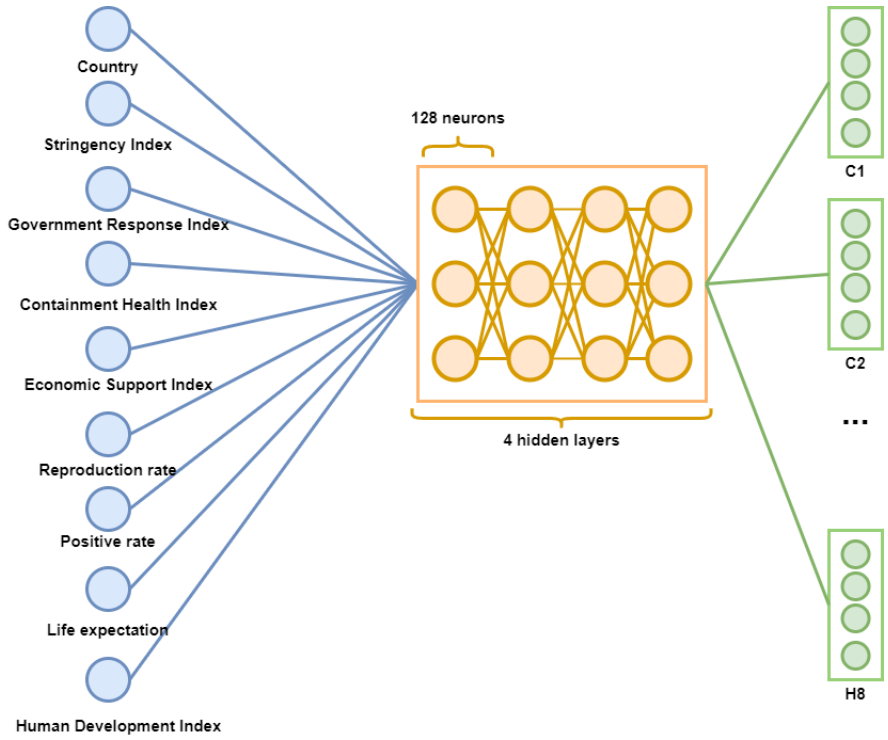

FIGURE 4.1 – Présentation du modèle de recommandation basé sur l'apprentissage profond

allant de la recommandation à l'obligation de fermeture d'une école. Par conséquent, la couche de sortie consacrée à la prédiction à cette mesure prédit quatre classes (c'est-à-dire aucune mesure appliquée, recommandation de fermeture d'école ou de changement d'horaire, etc.).

4.3 Approche sémantique pour l'explication des recommendations

Personne ne peut nier les performances élevées des modèles d'apprentissage profond. Bien qu'efficaces, ils agissent comme des "boîtes noires" car ils sont compliqués et moins interprétables que d'autres modèles d'IA, comme les modèles à boîte de verre (par exemple, les systèmes basés sur des règles). Nous avons ainsi proposé une approche sémantique pour expliquer les résultats du modèle de recommandation et promouvoir l'explicabilité du modèle. Plus précisément, l'approche proposée tire parti des technologies sémantiques pour construire dynamiquement une ontologie d'explication à l'aide d'une mise en correspondance avec des graphes sémantiques, des sources de données et des frameworks externes. Ensuite, notre approche extrait un sous-graphe d'explication de l'ontologie d'explication construit en tenant compte du rôle et des préférences de l'utilisateur. Pour ce faire, nous gardons le processus d'explication transparent pour l'utilisateur en déduisant implicitement ses préférences via l'analyse de ses interactions avec le modèle et, ainsi, en

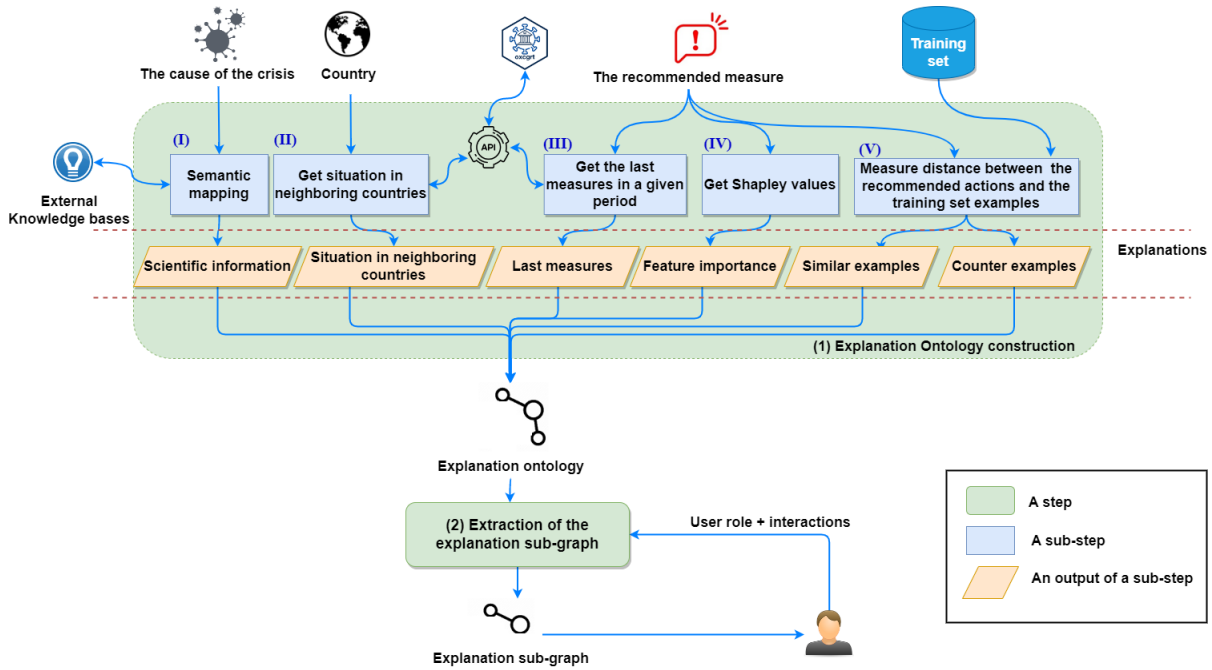

FIGURE 4.2 – Aperçu de l'approche d'explication proposée

adaptant les explications proposées en fonction de ses besoins. À cette fin, nous proposons une approche en deux étapes, à savoir (1) la construction de l'ontologie et (2) l'extraction du sous-graphe approprié. La figure 4.2 présente une vue d'ensemble de notre approche de l'explication. Dans ce qui suit, nous détaillons le raisonnement adopté pour concevoir chaque étape.

4.3.1 Étape de construction de l'ontologie d'explication

Comme indiqué au chapitre 2, les technologies sémantiques donnent un sens aux explications et décrivent les connaissances à l'aide de concepts sémantiques interconnectés par des relations. Ainsi, les ontologies clarifient la structure des connaissances et constituent le cœur de tout système de représentation des connaissances en décrivant les propriétés des objets et les relations entre eux qui peuvent être possibles dans un domaine de connaissance donné. Les ontologies sont utilisées pour représenter la sémantique des concepts, facilitant ainsi l'échange de données entre des systèmes disparates. De ce fait, nous nous appuyons sur cesdites technologies sémantiques, et notamment les ontologies, pour concevoir notre approche d'explication du modèle de recommandation. Nous prenons en compte les besoins des rôles des utilisateurs multiples pour leur proposer une variété d'explications, comme l'explication via les informations scientifiques (I), les statistiques (I,II), les voisins (II), les dernières mesures (III), l'importance des caractéristiques (IV), les pays similaires (V), et les contre-exemples (V) adoptés au cours d'une période donnée. Par exemple, en cas de crise, un scientifique peut être intéressé par des informations scientifiques, tandis qu'un expert en stratégie peut vouloir connaître les statistiques et la

situation dans les pays voisins. En outre, les deux utilisateurs peuvent être intéressés par des exemples similaires et des contre-exemples permettant de comprendre le raisonnement derrière le modèle de recommandation et d'étudier l'efficacité des mesures recommandées. Comme nous visons le domaine de la santé, (I) l'ontologie construite regroupe les informations médicales relatives à la maladie prédite ou la plus proche (c'est-à-dire dans le cas de l'identification d'une maladie inconnue telle que COVID-19 en 2020), telles que les symptômes, les traitements, les synonymes et les liens vers des bibliothèques médicales et gouvernementales externes. Pour ce faire, nous proposons un algorithme permettant de construire l'ontologie d'explication dynamiquement à travers le mapping et importation à partir de graphes de connaissances, de sources de données et de frameworks externes, tels que l'ontologie des maladies humaines (Human Disease Ontology) [77]. Nous recherchons donc dans l'ontologie externe des maladies les concepts qui sont en lien avec la maladie prédite et les présentons (c.-à-d., en faisant l'importation) comme des sous-classes de la classe "Maladie". Nous importons également de cette ontologie des informations médicales sur la maladie prédite, ainsi que des synonymes et des liens vers des bibliothèques scientifiques et gouvernementales externes. D'autre part, notre approche de l'explication utilise les éléments intermédiaires utilisés lors de la phase d'apprentissage du modèle de recommandation, comme l'ensemble de données d'apprentissage et les features du modèle, pour donner lieu à d'autres explications. Par exemple, l'ensemble de formations (c'est-à-dire l'ensemble de données utilisé pour former le modèle de recommandation) est considéré comme un ensemble de données contenant différentes informations susceptibles de servir comme exemples pour les explications. L'ensemble de formation contient l'historique des mesures sanitaires de chaque pays, ainsi que des informations contextuelles pratiques (III) telles que le taux de reproduction des virus, le taux de production, l'indice de développement humain et l'espérance de vie. Nous utilisons ces informations pour expliquer les recommandations à l'utilisateur final. En effet, nous mesurons la distance entre les mesures recommandées et les enregistrements de l'ensemble d'apprentissage à l'aide de la méthode de similarité Cosine [78]. Ensuite, nous extrayons les enregistrements de données relatifs aux pays qui ont appliqué la même mesure recommandée (V). Nous mesurons aussi la distance Cosine à l'aide d'informations contextuelles (par exemple, l'indice de développement humain) pour filtrer les enregistrements de l'ensemble de données. Le processus de filtrage vise à extraire les pays les plus ressemblants au pays concerné en ce qui concerne les informations contextuelles (c.-à-d. ayant des caractéristiques similaires). Ensuite, les enregistrements retirés seront proposés comme exemples similaires (V), tandis que les autres (c'est-à-dire ceux qui n'ont pas été retirés) seront proposés comme contre-exemples (V), et nous les présentons comme des sous-classes des classes "Explanation_by_similar_countries" et "Explanation_by_counter_examples". En plus des explications présentées ci-dessus, notre approche explicative fournit l'importance de la caractéristique comme une explication pour illustrer l'impact de chaque caractéristique sur la génération d'une décision. En conséquence, un utilisateur peut vérifier si le modèle de recommandation est biaisé et comprendre le raisonnement qui sous-tend la recommandation. Pour ce faire, nous utilisons SHAP (SHapley Additive exPlanations), une approche fondée sur la théorie des jeux qui vise à expliquer les résultats des modèles d'apprentissage automatique par le biais d'explications locales des valeurs de Shapley [79].

Ensuite, les valeurs de Shapley de chaque caractéristique sont proposées comme explication à l'utilisateur final. Les figures 4.3-4.6 explicitent le processus entier de construction des explications proposées.

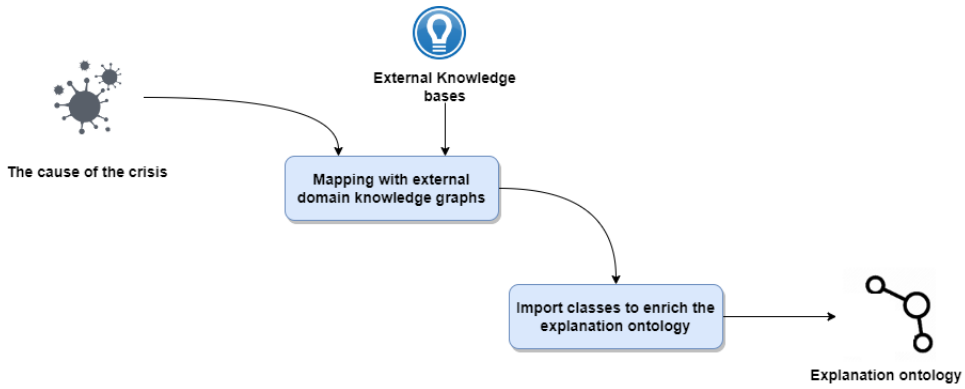

FIGURE 4.3 – Processus de construction d'une explication scientifique

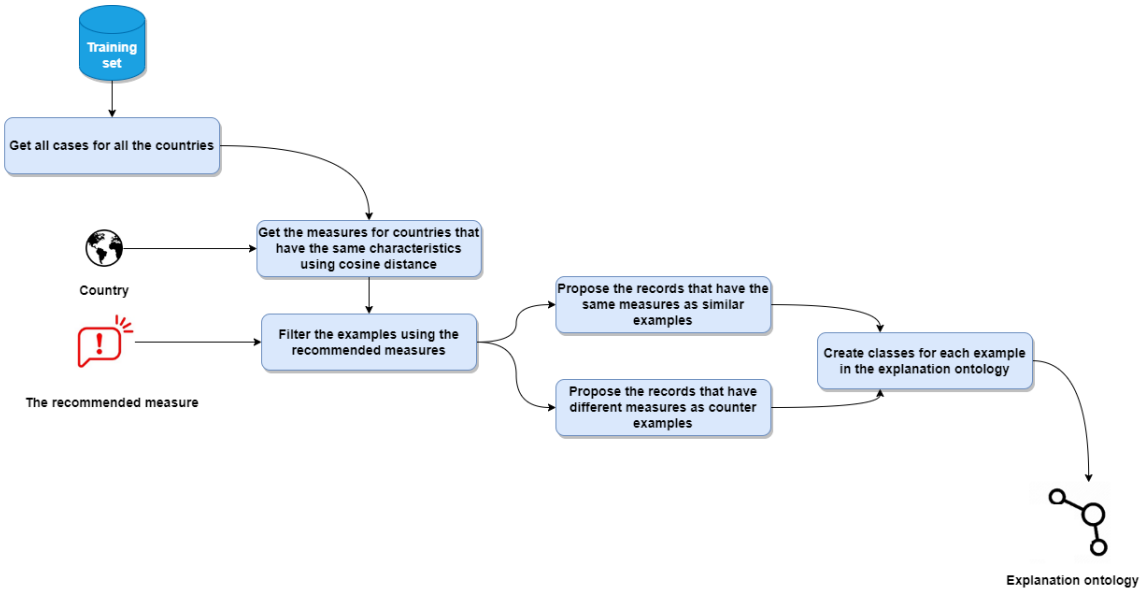

FIGURE 4.4 – Processus de construction d'une explication par des exemples

4.3.2 Étape d'extraction de sous-graphes d'explication

Après la construction dynamique de l'ontologie d'explication, l'étape suivante consiste à extraire et à proposer le sous-graphe approprié pour l'explication en fonction des préférences de l'utilisateur. Comme indiqué précédemment, les utilisateurs finaux peuvent avoir des intérêts différents pour les types d'explications en fonction de leurs rôles et de leurs besoins. Par exemple, les experts en santé peuvent être intéressés par des explications médicales, tandis que les analystes en stratégie peuvent préférer les statistiques, telles que la

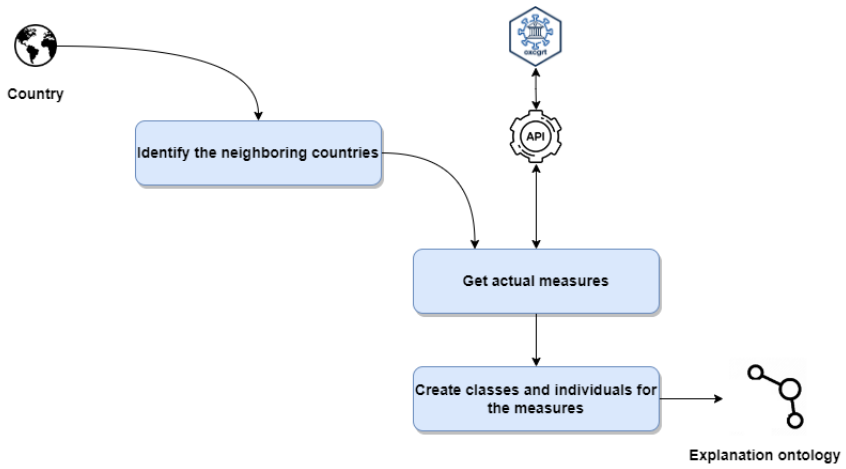

FIGURE 4.5 – Processus de construction d'une explication par la situation dans les pays voisins

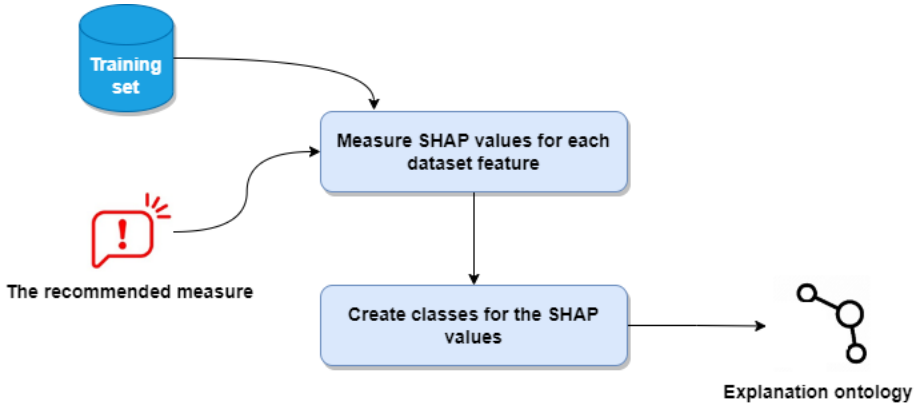

FIGURE 4.6 – Processus de construction d'une explication à l'aide de l'importance des attributs du modèle

situation dans les pays voisins. Nous prenons donc en compte les préférences des utilisateurs au cours du processus d'extraction de sous-graphes à partir de l'ontologie d'explication. À cette fin, notre approche de l'explication s'appuie sur la méthode de factorisation matricielle, l'une des méthodes les plus efficaces utilisées dans les systèmes de recommandation et adoptée par des services en ligne tels que Netflix pour accélérer la recherche de recommandations de contenu utilisateur [80]. De cette manière, nous déduisons implicitement les préférences des utilisateurs en analysant leurs interactions avec le système de recommandation. Nous déduisons ensuite leurs préférences et proposons un sous-graphe d'explication approprié. Plus précisément, l'approche que nous proposons crée une matrice d'explication de l'utilisateur dans laquelle nous avons enregistré les interactions recueillies. Elle s'appuie ensuite sur la factorisation matricielle, à savoir l'algorithme SVD, pour faire correspondre les utilisateurs et les explications à un espace de facteurs latents de dimensionnalité commune. Les interactions entre l'utilisateur et l'explication dans cet espace sont modélisées comme des produits internes, comme décrit dans [81]. Comme nous

avons considéré différents contextes de décision dans notre approche, nous avons défini plusieurs matrices de préférences consacrées à chaque contexte de décision. Ainsi, nous avons déduit les préférences des utilisateurs dans différents contextes afin de recommander des sous-graphes distincts pour chaque utilisateur en fonction de son contexte de décision. Par exemple, un utilisateur peut préférer une explication utilisant la situation dans les pays voisins dans une situation ordinaire. En revanche, en cas de crise, il préférera une explication à l'aide de contre-exemples. Ainsi, l'approche d'explication proposée s'aligne sur les besoins des utilisateurs en fonction de leur contexte de décision. Après avoir présenté l'approche d'explication proposée et détaillé chaque étape, nous nous concentrons dans la section suivante sur la mise en œuvre et les expériences élaborées pour évaluer l'efficacité de notre proposition.

4.4 Mise en œuvre et résultats expérimentaux

Dans cette section, nous présentons et détaillons la mise en œuvre et les expériences élaborées qui se concentrent sur le modèle de recommandation basé sur l'apprentissage profond et l'approche d'explication. Tout d'abord, nous présentons les paramètres expérimentaux adoptés pour concevoir et évaluer nos contributions. Ensuite, nous présentons un extrait des expériences élaborées et discutons de l'objectif et des résultats.
Ainsi, nous avons d'abord évalué la performance du modèle de recommandation en utilisant des mesures de performance dédiées, à savoir le rappel, la précision, l'exactitude et le score F1. Plus précisément, nous avons entraîné et validé le modèle d'apprentissage profond à l'aide de l'ensemble de données OxCGRT, qui englobe les mesures stratégiques prises par les gouvernements pour lutter contre l'épidémie de COVID-19. Nous avons tenu compte des différentes mesures stratégiques, telles que les fermetures d'écoles, les restrictions de voyage et les politiques économiques, collectées entre le 1er janvier 2020 et le 31 janvier 2022, et couvrant plus de 180 pays. Après avoir appliqué les processus de prétraitement et de sélection des caractéristiques, nous avons sélectionné les caractéristiques les plus cohérentes, non redondantes et pertinentes pour construire le modèle d'apprentissage profond. Comme décrit ci-dessus, nous avons choisi neuf caractéristiques sur quinze comme caractéristiques d'entrée, y compris le pays, les indices réels (c'est-à-dire la rigueur, la réponse du gouvernement, le confinement sanitaire et le soutien économique), le taux de reproduction, le taux positif, le taux de population, l'âge médian et l'espérance de vie, ainsi que l'indice de développement humain.
Après l'implémentation du modèle de recommandation, nous présentons un extrait des expérimentations élaborées qui se concentrent sur l'évaluation du modèle de recommandation et l'approche d'explication en évaluant (1) la qualité de l'ontologie d'explication constituée, (2) le sous-graphe extrait, et l'efficacité (3) du mécanisme de recommandation. Dans ce qui suit, nous détaillons les résultats obtenus suite à l'application du protocole expérimental.
Comme nous traitons la recommandation comme un problème de classification, nous avons évalué les performances de notre modèle de recommandation à l'aide des mesures d'évaluation mentionnées précédemment, à savoir le rappel, la précision, l'exactitude et

le score F1.Nous avons divisé le jeu de données OxCGRT en ensembles de formation, de test et de validation pour effectuer les processus de formation et de validation. Nous avons formé le modèle d'apprentissage profond multi-sorties sur 150 époques (c'est-à-dire jusqu'à ce qu'il se stabilise) et utilisé l'optimiseur ADAM pour modifier les hyperparamètres et minimiser la fonction de perte. Ainsi, nous avons évalué les performances de notre modèle de recommandation à l'aide des mesures d'évaluation mentionnées précédemment, à savoir le rappel (95%), la précision (96%), l'accuracy (entre 91% et 99% pour chaque sortie du modèle comme présenté dans la figure 4.7) et le score F1 (95%). Nous avons

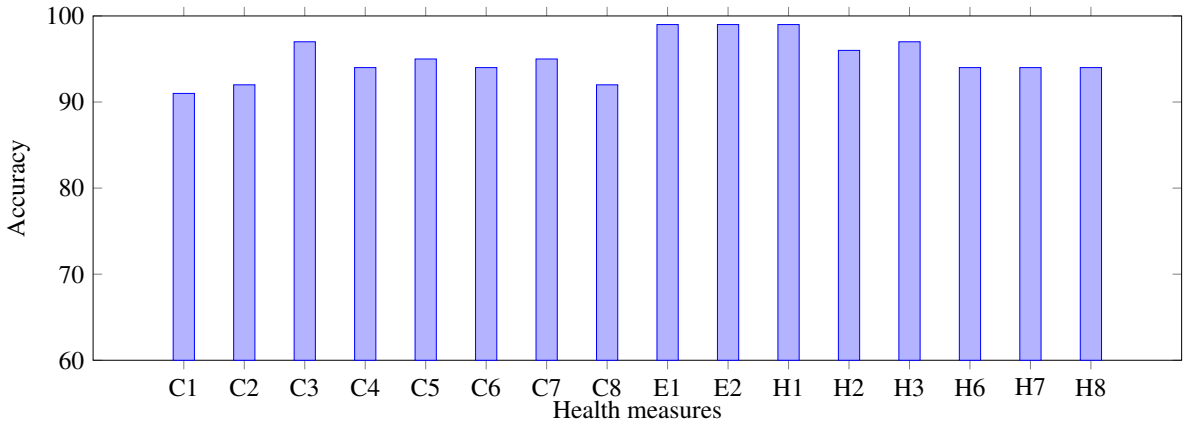

FIGURE 4.7 – Accuracy de la recommandation de chaque mesure de santé

aussi évalué la qualité des types d'explication proposés en examinant les sous-graphes extraits, tout comme nous l'avons fait pour l'ontologie d'explication globale. Pour ce faire, nous avons évalué la qualité de chaque sous-graphe d'explication recommandé à l'aide des mêmes mesures d'évaluation que l'ontologie d'explication (schéma, connaissance, évaluation des graphes). Les résultats présentés dans le tableau 4.1 prouvent la richesse sémantique des sous-graphes d'explication, qui est démontrée par le nombre et la répartition des attributs, des classes et des relations. En outre, nous avons évalué les performances

TABLE 4.1 – Évaluation des sous-graphes d'explication

Catégorie	Métrique	Counter examples	Similar countries	Feature importance	Neighboring countries	Last Measures
Schema metrics	Attribute richness	6.25	6.25	10.2	11.8	3.93
	Inheritance richness	0.5	0.5	0.2	0.2	0.69
	Relationship richness	0.44	0.44	0.8	0.8	0.3
	Attribute class ratio	0.63	0.63	0.6	0.6	0.75
	class/relation ratio	1	1	1	1	1
Knowledge metrics	Average population	2.63	1.5	1	3.4	4.9
	class richness	0.75	0.75	0.8	0.83	0.84
Graph metrics	Absolute root node	1	1	1	1	1
	Absolute leaf cardinality	6	6	4	4	11
	Absolute sibling cardinality	6	6	3	3	11
	Maximal depth	3	3	2	2	3

de l'étape de recommandation des sous-graphes d'explication en examinant différents algorithmes de recommandation. Plus précisément, nous avons évalué les performances de

la factorisation matricielle, de la méthode KNN et d'algorithmes tels que Slope One, Random Recommendation et Co-Clustering. Les performances des algorithmes présentés sont évaluées par rapport à des données synthétiques générées à travers des ensembles de données de référence (Benchmarks). Ainsi, nous avons utilisé ces ensembles de données pour entraîner les algorithmes de recommandation présentés. Nous nous appuyons également sur des mesures telles que la précision, le rappel, l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE). Le tableau 4.2 illustre les résultats de la comparaison et prouve l'efficacité du paradigme de factorisation matricielle que nous avons adopté pour la recommandation d'explication. Nous présentons également dans la Figure 4.8 un exemple de graph d'explication pour la pandémie COVID-19.

TABLE 4.2 – Comparaison des performances des algorithmes de recommandation

		Base de données synthétique 1				**Base de données synthétique 2**			
		P@K	R@K	MAE	RMSE	P@K	R@K	MAE	RMSE
Matrix Factorization	**SVD**	**0.96**	**0.95**	**0.71**	**0.92**	**0.97**	**0.96**	**0.81**	**1.01**
	NMF	0.89	0.89	0.89	1.12	0.93	0.94	1.02	1.28
	SVD++	0.96	0.95	0.73	0.93	0.97	0.96	0.80	1.01
K Nearest Neighbors	KNN-Baseline	0.93	0.93	0.84	1.10	0.95	0.95	0.94	1.16
	KNN with Means	0.92	0.92	0.90	1.14	0.95	0.95	0.89	1.16
	KNN with Z-Score	0.91	0.91	0.93	1.17	0.94	0.94	0.91	1.18
Other	SlopeOne	0.93	0.93	0.81	1.02	0.95	0.95	0.84	1.11
	Random	0.90	0.90	1.11	1.39	0.94	0.94	1.11	1.43
	CoClustering	0.91	0.91	0.87	1.09	0.90	0.90	0.86	1.09

4.5 Conclusion

Dans le présent chapitre, nous avons introduit une nouvelle approche pour la recommandation explicable des actions pour la gestion des crises. Pour ce faire, nous avons conçu un nouveau modèle basé sur l'apprentissage profond. Le modèle proposé génère plusieurs sorties, car chaque couche de sortie prédit la rigueur de chaque mesure à appliquer. En le mettant en oeuvre dans le domaine de la santé, le modèle de recommandation propose des mesures sanitaires liées aux politiques du confinement et de la fermeture, de l'économie et du système de la santé. Comme les algorithmes d'apprentissage profond constituent des modèles de boîte noire, nous avons proposé une approche sémantique qui explique et argumente le choix des recommandations générées par le modèle proposé. L'originalité de nos contributions porte sur plusieurs points, à savoir la diversité des mesures proposées et des types d'explication tels que l'explication par les exemples, les contre-exemples, les pays voisins, l'importance des caractéristiques, et les dernières mesures. Notre approche de l'explication a le mérite de fournir des explications variées en fonction des besoins de l'utilisateur. Pour ce faire, nous nous appuyons sur la conception dynamique d'une ontologie d'explication et utilisons des techniques de factorisation de matrice pour extraire de l'ontologie construite le sous-graphe d'explication adapté à l'uti-

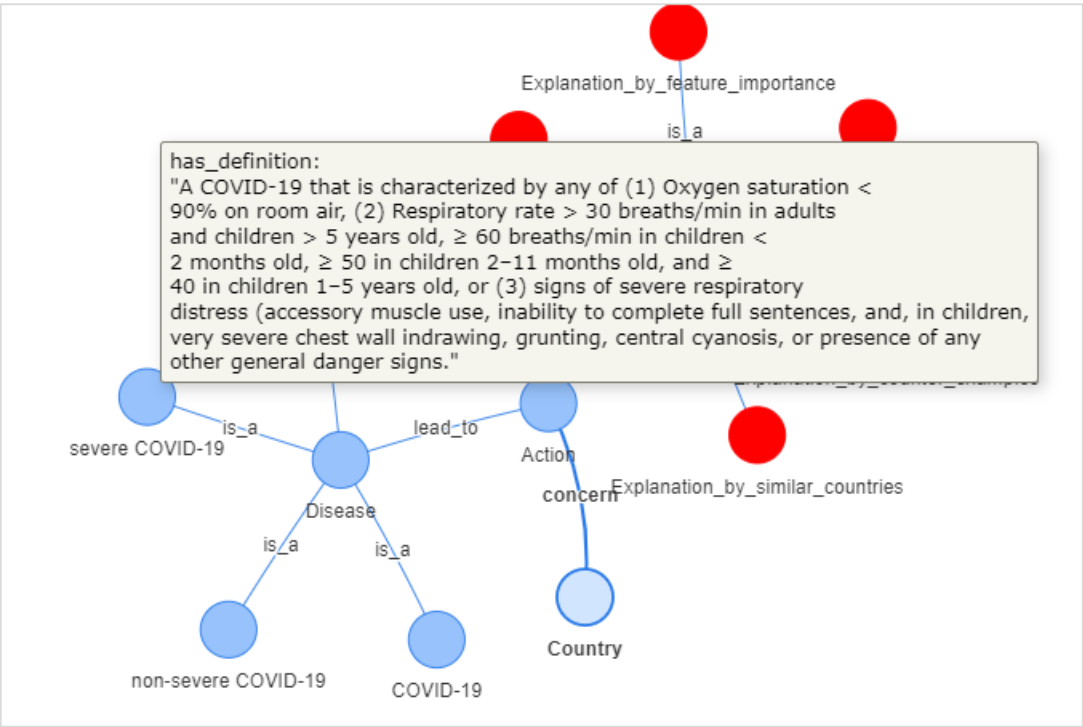

FIGURE 4.8 – Exemple d'explication d'un cas grave de COVID-19

lisateur. Par la suite, nous avons mené plusieurs expériences pour évaluer les performances et prouver l'efficacité de notre proposition en termes de recommandation et d'explication :

- Nous avons évalué les performances du modèle de recommandation en mesurant l'exactitude, la précision, le rappel et le score F1. Les résultats ont montré la robustesse du modèle de recommandation de mesures proposé.
- Nous avons appliqué des mesures d'évaluation des schémas, des connaissances et des graphes pour évaluer l'expressivité et la complexité de l'ontologie d'explication et des sous-graphes extraits. Les résultats ont donc montré la richesse sémantique et la faible complexité des ontologies d'explication et des sous-graphes.
- Nous avons comparé les performances de différents algorithmes de recommandation, tels que les variantes de la factorisation matricielle, les algorithmes basés sur le KNN, SlopeOne, etc. Les résultats ont montré que les algorithmes de factorisation matricielle, à savoir SVD et SVD++, sont plus performants que les autres algorithmes.

CHAPITRE 5

Conclusion and future work

L'humanité a assisté à plusieurs phénomènes qui ont perturbé la vie quotidienne et impacté l'écosystème, comme l'ouragan Katrina aux États-Unis, la crise financière mondiale, et récemment le tremblement de terre en Turquie. À l'échelle mondiale, le dernier quinquennat a été marqué par l'épidémie du COVID-19. Cette épidémie a constitué un exemple concret pour le public en matière de gestion des crises par les différentes communautés concernées, telles que les experts en santé et en stratégie, l'économie, etc. La gestion des crises, en général, exige de surmonter le problème sous plusieurs angles (stratégique, économique, etc.) en analysant des données massives et hétérogènes provenant de diverses sources, telles que les IoT, les réseaux sociaux et les capteurs, collectées en batch et en streaming. En outre, les utilisateurs impliqués dans ce processus peuvent avoir des exigences différentes dans divers contextes concernant les différentes étapes du processus de gestion (c'est-à-dire la curation des données, l'analyse, la prédiction, la recommandation, etc.).

Cette thèse aborde des questions de recherche fondamentales concernant les défis de la curation de données adaptatives et de la recommandation expliquée des actions pour la gestion des crises. Elle a donné lieu à trois contributions principales :

- Une étude de la littérature sur la gestion des données dans les lacs de données (data lakes) et data lakehouses.
- Un framework pour la curation de données multistructurées, fondé sur l'ontologie DARQAN pour la caractérisation des données et l'évaluation de leurs qualités qui guide la composition des services à l'aide de l'approche ACUSEC.
- Un modèle de recommandation multi-cibles basé sur l'apprentissage profond et une approche XAI basée sur la sémantique qui explique les choix du modèle de recommandation pour des rôles d'utilisateurs multiples.

Nous avons validé ces contributions dans le contexte de la curation des données de santé et de la recommandation de mesures sanitaires.

De part les résultats obtenus et leurs efficacités, les contributions améliorent de manière remarquable la gestion des données et la recommandation pour la gestion des crises. Toutefois, il est encore possible d'améliorer le processus de curation des données et de recommandation, et ces améliorations potentielles sont détaillées ci-après.

5.1 Limites

5.1.1 Évolution de la performance de l'approche de la curation des données dans la nouvelle ère des données

Nous avons assisté à un quart de siècle de transformation numérique qui a inauguré l'ère des données, de l'introduction du courrier électronique à l'analyse des données massives, en passant par le stockage en cloud et le SaaS, et nous sommes aujourd'hui à l'aube d'une nouvelle ère. Cette dernière se caractérise par une explosion de la taille et de la quantité des données et des services qui les traitent. En fait, nous considérons qu'une bibliothèque de services de curation ne pourrait pas atteindre plus de 12 000 services. Par conséquent, notre approche de composition reste valable et efficace pour la curation de données. En effet, notre approche de la curation de données basée sur les services repose sur le Q-Learning, qui a prouvé son efficacité et son efficience par rapport aux algorithmes de Monte Carlo à première visite et de la différence temporelle, ainsi qu'à d'autres algorithmes de composition, comme l'illustrent les expériences. Cependant, la nouvelle ère des données peut avoir des exigences différentes, ce qui peut nous amener à revoir et améliorer notre approche de curation.

5.1.2 Évolution de la taille de l'ontologie d'explication

Nous avons proposé un modèle multicibles basé sur l'apprentissage profond pour la recommandation de mesures expliquées via une approche sémantique. Notre approche construit une ontologie d'explication de manière dynamique en important des fragments à partir d'ensembles de données, de frameworks et de graphes de connaissances. Cependant, la construction de l'ontologie peut devenir gourmande en termes de classes importées, ce qui peut entraîner un risque d'augmentation de la taille du graphe structurel. Bien que nous ayons démontré la simplicité de l'ontologie d'explication construite, l'augmentation de la taille du graphe pourrait accroître le temps de l'interrogation et la gestion de l'ontologie.

5.2 Futures pistes de recherches possibles

Le domaine de la présente thèse et les défis associés bien motivants ne font qu'ouvrir de larges perspectives que nous avons pensés, parmi celle-ci nous notons :

5.2.1 Étude des performances de l'apprentissage par renforcement dans la nouvelle ère des données

Nous avons l'intention d'étudier les performances d'autres approches d'apprentissage par renforcement pour améliorer la composition des services de curation de données dans la nouvelle ère des données. En effet, nous pensons que l'apprentissage par renforcement reste une solution pratique pour le problème de la composition des services de curation puisque nous avons recours à l'apprentissage non supervisé. Néanmoins, l'algorithme de Q-learning peut devenir moins efficace en raison des exigences de la nouvelle ère des données. C'est pourquoi nous envisageons d'étudier les performances d'autres méthodes d'apprentissage par renforcement, telles que le Deep Q-Learning et le Multi-agent determinantal Q-learning. Ces méthodes peuvent être moins efficaces dans notre contexte actuel car gourmandes en données de formation (c'est-à-dire en services dans notre cas), mais il serait intéressant d'explorer ses éventuelles capacités pour améliorer la composition de services en termes d'adaptativité à l'avenir.

5.2.2 Prévision des futures crises

Le chapitre 4 a abordé les limites liées à la prédiction des crises, en particulier dans le domaine des soins. Nous avons tiré profit des techniques des systèmes multi-agents pour constituer un système de prédiction qui analyse les signaux faibles pouvant conduire à l'occurrence d'une crise, faisant lien entre la curation de données et la recommandation de mesures de santé. Cependant, nous pensons que la prédiction de crise notamment en temps réel peut être une piste de recherche intéressante. Par conséquent, nous avons l'intention d'examiner plus en profondeur la prédiction des crises en analysant les approches, les techniques et les méthodologies existantes pour construite un système entier avec un continuum de l'ingestion des données en temps réel jusqu'à la recommandation expliquée.

5.2.3 Étudier l'évolution de l'ontologie de l'explication

L'approche ainsi que l'ontologie d'explication que nous avons proposée constituent une contribution originale et générique dans le domaine. Néanmoins, nous avons l'intention d'étudier l'application des approches d'évolution des ontologies afin d'optimiser l'ontologie d'explication construite via notre approche d'explication. À cette fin, nous pensons examiner les techniques existantes pour faire évoluer le schéma et la sémantique de l'ontologie, telles que le word embedding et l'apprentissage automatique. En outre, nous prévoyons d'étudier les métriques permettant de mesurer l'évolution de l'ontologie en termes d'ajouts/changements structurels, relatifs et de sous-classes. À cette fin, en sus de l'examen de la littérature couvrant ces aspects, nous visons à identifier les exigences de l'aspect évolution de l'ontologie afin de les prendre en compte et incorporer dans notre approche d'explication et d'améliorer son évolution.

Bibliographie

[1] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, A. Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion*, 58 :82–115, 2020.

[2] Reginald Bell. Managing the prodromal crisis situation : Two techniques to avoid turning a surge into a mega-tsunami. *Supervision*, 72 :3–6., 01 2011.

[3] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, and Patricia C. Arocena. Data lake management : Challenges and opportunities. *Proceedings of the VLDB Endowment*, 12(12) :1986–1989, 2018.

[4] F. Conesa and F. Destin. Comment utiliser et valoriser les données dans un contexte big data – réflexion méthodologique sur les défis que soulève l'analyse statistique de données hétérogènes massives. *Revue d'Épidémiologie et de Santé Publique*, 63 :S52, 2015. EPI-CLIN 2015.

[5] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review : Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1) :91–99, 2022. International Conference on Intelligent Engineering Approach(ICIEA-2022).

[6] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. A survey on data preprocessing for data stream mining : Current status and future directions. *Neurocomputing*, 239 :39–57, 2017.

[7] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Benítez, and Francisco Herrera. Big data preprocessing : methods and prospects. *Big Data Analytics*, 1, 11 2016.

[8] Firas Zouari, Kabachi Nadia, Khouloud Boukadi, and Chirine Ghedira. Data management in the data lake : A systematic mapping. pages 280–284, 07 2021.

[9] Firas Zouari, Chirine Ghedira, Nadia Kabachi, and Khouloud Boukadi. Towards an adaptive curation services composition based on machine learning. *IEEE International Conference on Web Services (ICWS)*, pages 73–78, 2021.

[10] Firas Zouari, Chirine Ghedira, Nadia Kabachi, and Khouloud Boukadi. A service-based framework for adaptive data curation in data lakehouses. *IEEE International Conference on Web Services (ICWS)*, 2022.

[11] Matei Zaharia, Ali Ghodsi 0002, Reynold Xin, and Michael Armbrust. Lakehouse : A new generation of open platforms that unify data warehousing and advanced analytics. In *11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021, Online Proceedings*. www.cidrdb.org, 2021.

[12] Kathryn Laskey and Kenneth Laskey. Service oriented architecture. *Wiley Interdisciplinary Reviews : Computational Statistics*, 1 :101 – 105, 07 2009.

[13] Angel Lagares Lemos, Florian Daniel, and Boualem Benatallah. Web service composition. *ACM Computing Surveys*, 48 :1–41, 12 2015.

[14] Mahboobeh Moghaddam and Joseph G. Davis. *Service Selection in Web Service Composition : A Comparative Review of Existing Approaches*, pages 321–346. Springer New York, New York, NY, 2014.

[15] Yoav Shoham. Agent-oriented programming. *Artificial Intelligence*, 60(1) :51–92, 1993.

[16] Sabina Leonelli. Classificatory theory in data-intensive science : The case of open biomedical ontologies. *Int. Studies in the Philosophy of Science*, 26(1) :47–65, 2012.

[17] Paul Beckman, Tyler J. Skluzacek, Kyle Chard, and Ian Foster. Skluma : A statistical learning pipeline for taming unkempt data repositories. pages 1–4, 06 2017.

[18] Amin Beheshti, Kushal Vaghani, Boualem Benatallah, and Alireza Tabebordbar. Crowdcorrect : A curation pipeline for social data cleansing and curation. *Information Systems in the Big Data Era*, pages 24–38, 2018.

[19] Amin Beheshti, Boualem Benatallah, Reza Nouri, and Alireza Tabebordbar. Corekg : a knowledge lake service. *Proceedings of the VLDB Endowment*, 11 :1942–1945, 08 2018.

[20] Antonio Maccioni and Riccardo Torlone. Kayak : A framework for just-in-time data preparation in a data lake. *Advanced Information Systems Engineering*, pages 474–489, 2018.

[21] André Pomp, Vadim Kraus, Lucian Poth, and Tobias Meisen. Semantic concept recommendation for continuously evolving knowledge graphs. pages 361–385, 02 2020.

[22] Hassan Alrehamy and Coral Walker. (personal data lake) semlinker : automating big data integration for casual users. *Journal of Big Data*, 5, 03 2018.

[23] Amin Beheshti, Boualem Benatallah, Alireza Tabebordbar, Hamid R. Motahari Nezhad, Moshe Barukh, and Reza Nouri. Datasynapse : A social data curation foundry. *Distributed and Parallel Databases*, 37, 09 2019.

[24] Nikolaos Konstantinou, Edward Abel, Luigi Bellomarini, Alex Bogatu, Cristina Civili, Endri Irfanie, Martin Koehler, Lacramioara Mazilu, Emanuel Sallinger, Alvaro A.A. Fernandes, Georg Gottlob, John A. Keane, and Norman W. Paton. VADA : an architecture for end user informed data preparation. *Journal of Big Data*, 6(1) :1–32, 2019.

[25] Yihan Gao, Silu Huang, and Aditya Parameswaran. Navigating the data lake with datamaran : Automatically extracting structure from log datasets. page 943–958, 2018.

[26] Giovanni Simonini, Luca Gagliardelli, Sonia Bergamaschi, and H.V. Jagadish. Scaling entity resolution : A loosely schema-aware approach. *Information Systems*, 83, 03 2019.

[27] Yihan Gao, Silu Huang, and Aditya Parameswaran. Navigating the data lake with datamaran : Automatically extracting structure from log datasets. page 943–958, 2018.

[28] Rihan Hai, Christoph Quix, and Dan Wang. Relaxed functional dependency discovery in heterogeneous data lakes. pages 225–239, 10 2019.

[29] Seyed Mehdi Reza Beheshti, Alireza Tabebordbar, Boualem Benatallah, and Reza Nouri. On automating basic data curation tasks. *26th International World Wide Web Conference 2017, WWW 2017 Companion*, pages 165–169, 2019.

[30] Gui Santana. Crisis management and tourism. *Journal of Travel & Tourism Marketing*, 15 :299–321, 01 2004.

[31] Fei Hao, Qu Xiao, and Kaye Chon. Covid-19 and china's hotel industry : Impacts, a disaster management framework, and post-pandemic agenda. *International Journal of Hospitality Management*, 90 :102636, 08 2020.

[32] Pradeep Racherla and Clark Hu. A framework for knowledge-based crisis management in the hospitality and tourism industry. *Cornell Hospitality Quarterly - CORNELL HOSP Q*, 50 :561–577, 11 2009.

[33] Daekyo Jung, Vu Tuan, Dai Tran, Minsoo Park, and Seunghee Park. Conceptual framework of an intelligent decision support system for smart city disaster management. *Applied Sciences*, 10 :666, 01 2020.

[34] Mohamed Abdel-Basset and Rehab Mohamed. A novel plithogenic topsis- critic model for sustainable supply chain risk management. *Journal of Cleaner Production*, 247 :119586, 2020.

[35] Senbeto Dagnachew Leta and Irene Cheng Chu Chan. Learn from the past and prepare for the future : A critical assessment of crisis management research in hospitality. *International Journal of Hospitality Management*, 95 :102915, 2021.

[36] Jukrin Moon, Farzan Sasangohar, Changwon Son, and S. Peres. Cognition in crisis management teams : An integrative analysis of definitions. *Ergonomics*, 63 :1–23, 06 2020.

[37] Zhaotong Li, Xueqin Wang, Xue Li, and Kum Fai Yuen. Post covid-19 : Health crisis management for the cruise industry. *International Journal of Disaster Risk Reduction*, 71 :102792, 2022.

[38] Fei Hao, Qu Xiao, and Kaye Chon. Covid-19 and china's hotel industry : Impacts, a disaster management framework, and post-pandemic agenda. *International Journal of Hospitality Management*, 90 :102636, 2020.

[39] Mahmood Mir, Sanjay Jamwal, Abolfazl Mehbodniya, Tanya Garg, Ummer Iqbal, and Issah Samori. Iot-enabled framework for early detection and prediction of covid-19 suspects by leveraging machine learning in cloud. *Journal of Healthcare Engineering*, 2022 :1–16, 04 2022.

[40] Fardin Abdali-Mohammadi, Maytham N. Meqdad, and Seifedine Kadry. Development of an iot-based and cloud-based disease prediction and diagnosis system for healthcare using machine learning algorithms. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 9, 12 2020.

[41] Forsad Al Hossain, Andrew Lover, George Corey, Nicholas Reich, and Tauhidur Rahman. Flusense : A contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4 :1–28, 03 2020.

[42] Xiongwei Zhang, Hager Saleh, Eman Younis, Radhya Sahal, and Abdelmgeid Ali. Predicting coronavirus pandemic in real-time using machine learning and big data streaming system. *Complexity*, 2020 :1–10, 12 2020.

[43] Behnam Nikparvar, Md. Mokhlesur Rahman, Faizeh Hatami, and Jean-Claude Thill. Spatio-temporal prediction of the covid-19 pandemic in us counties : modeling with a deep lstm neural network. *Scientific Reports*, 11, 11 2021.

[44] Alejandro Barredo Arrieta, Natalia Diaz Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado González, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, V. Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 12 2019.

[45] Milad Moradi and Matthias Samwald. Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, 165 :113941, 03 2021.

[46] Roberto Confalonieri and Tarek R. Besold. Trepan reloaded : A knowledge-driven approach to explaining black-box models. In *ECAI*, 2020.

[47] Cecilia Panigutti, Dino Pedreschi, and Alan Perotti. Doctor xai : an ontology-based approach to black-box sequential data classification explanations. 01 2020.

[48] Md. Kamruzzaman Sarker, Ning Xie, Derek Doran, Michael Raymer, and Pascal Hitzler. Explaining trained neural networks with semantic web technologies : First steps. *ArXiv*, abs/1710.04324, 2017.

[49] Montserrat Batet, Aida Valls, Karina Gibert, and David Sánchez. Semantic clustering using multiple ontologies. volume 220, pages 207–216, 01 2010.

[50] Sousa Manuel and DeJoao Ribeiro Leite. Aligning artificial neural networks and ontologies towards explainable ai. *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI'21*, AAAI Press, 2021.

[51] Shruthi Chari, Oshani Seneviratne, Daniel Gruen, Morgan Foreman, Amar Das, and Deborah Mcguinness. *Explanation Ontology : A Model of Explanations for User-Centered AI*, pages 228–243. 11 2020.

[52] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system : A survey and new perspectives. 07 2017.

[53] Jim Weatherall, Faisal M. Khan, Mishal Patel, Richard Dearden, Khader Shameer, Glynn Dennis, Gabriela Feldberg, Thomas White, and Sajan Khosla. Clinical trials, real-world evidence, and digital medicine. In *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, pages 191–215. Academic Press, 2021.

[54] Jacky Akoka, Isabelle Comyn-Wattiau, and Nabil Laoufi. Research on Big Data – A systematic mapping study. *Computer Standards and Interfaces*, 54 :105–115, 2017.

[55] Niccolò Tempini. Data curation-research : Practices of data standardization and exploration in a precision medicine database. *New Genetics and Society*, 40, 12 2020.

[56] Baby Gobin. An agile methodology for developing ontology modules which can be used to develop modular ontologies. 11 2013.

[57] Jeremy Debattista, Christoph Lange, and Sören Auer. daq, an ontology for dataset quality information. *CEUR Workshop Proceedings*, 1184, 04 2014.

[58] Timothy Lebo, Satya Sahoo, Deborah Mcguinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. *PROV-O : The PROV Ontology*. 04 2013.

[59] Zhaoheng Liu, Zhuoming Xu, and Xiutao Xia. Towards systematic analysis and summary of duv-based dataset usage information. pages 169–172, 09 2016.

[60] Doh SHIN, Sang LEE, Junghyun KANG, and Eun PARK. Data catalogue standards based on dcat for transportation data : Dcat-trans. *Journal of Korean Society of Transportation*, 37 :430–444, 10 2019.

[61] Armin Haller, Krzysztof Janowicz, Simon Cox, Danh Phuoc, Kerry Taylor, and Maxime Lefrançois. *Semantic Sensor Network Ontology*. 10 2017.

[62] Riccardo Albertoni and Antoine Isaac. Introducing the data quality vocabulary (dqv). *Semantic Web*, 12, 04 2020.

[63] Vasileios C. Pezoulas, Konstantina D. Kourou, Fanis Kalatzis, Themis P. Exarchos, Aliki Venetsanopoulou, Evi Zampeli, Saviana Gandolfo, Fotini Skopouli, Salvatore De Vita, Athanasios G. Tzioufas, and Dimitrios I. Fotiadis. Medical data quality assessment : On the development of an automated framework for medical data curation. *Computers in Biology and Medicine*, 107 :270–283, 2019.

[64] Dejing Dou, Hao Wang, and Haishan Liu. Semantic data mining : A survey of ontology-based approaches. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, IEEE ICSC 2015*, pages 244–251, 2015.

[65] Csaba Szepesvári. *Algorithms for reinforcement learning*, volume 9. 2010.

[66] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning : An Introduction*. The MIT Press, second edition, 2018.

[67] José Parejo, Sergio Segura, Pablo Fernandez, and Antonio Ruiz-Cortés. Qos-aware web services composition using grasp with path relinking. *Expert Systems with Applications*, 41 :4211–4223, 07 2014.

[68] Honghao Gao, Wanqiu Huang, and Yucong Duan. The cloud-edge-based dynamic reconfiguration to service workflow for mobile ecommerce environments : A qos prediction perspective. *ACM Transactions on Internet Technology*, 21 :1–23, 01 2021.

[69] Wei Zhang, Carl K. Chang, Taiming Feng, and Hsin-yi Jiang. Qos-based dynamic web service composition with ant colony optimization. pages 493–502, 2010.

[70] T. F. Michael Raj, P. Sivapragasam, R. Balakrishnan, G. Lalithambal, and S. Ragasubha. Qos based classification using k-nearest neighbor algorithm for effective web service selection. *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–4, 2015.

[71] Gerardo Canfora, Massimiliano Di Penta, Raffaele Esposito, and Maria Luisa Villani. An approach for qos-aware service composition based on genetic algorithms. *GECCO 2005 - Genetic and Evolutionary Computation Conference*, 3387, 06 2005.

[72] Adam Wade, Anna Petherick, Beatriz Kira, Emily Cameron-Blake, Helen Tatlow, Jodie Elms, Kaitlyn Green, Laura Hallas, Martina Di Folco, Thomas Hale, Toby Phillips, Yuxi Zhang, Jessica Anania, Bernardo Andretti De Mello, Noam Angrist, Roy Barnes, Thomas Boby, Alice Cavalieri, Benjamin Edwards, Samuel Webster, Lucy Ellen, Rodrigo Furst, Rafael Goldszmidt, Maria Luciano, Saptarshi Majumdar, Radhika Nagesh, Annalena Pott, Andrew Wood, Adam Wade, and Hao Zha. Aligning artificial neural networks and ontologies towards explainable ai. *BLAVATNIK SCHOOL WORKING PAPER*, 2020.

[73] Muhammad Ilham Gunawan and Yunieta Anny Nainggolan. Stringency index and stock market return amidst covid19 pandemic : Evidence from emerging stock market countries. *Proceedings of the International Conference on Industrial Engineering and Operations Management Rome*, pages 2513–2521, 2021.

[74] Feng Liu, Meichang Wang, and Meina Zheng. Effects of covid-19 lockdown on global air quality and health. *Science of The Total Environment*, 755 :142533, 2021.

[75] Zander S. Venter, Kristin Aunan, Sourangsu Chowdhury, and Jos Lelieveld. Covid-19 lockdowns cause global air pollution declines. *Proceedings of the National Academy of Sciences*, 117(32) :18984–18990, 2020.

[76] Massimo Pulejo and Pablo Querubín. Electoral concerns reduce restrictive measures during the covid-19 pandemic. *Journal of Public Economics*, 198 :104387, 2021.

[77] Lynn Schriml, Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, Linda Jeng, Cynthia Bearer, Richard Lichenstein, Katharine Bisordi, Nicole Campion, Brooke Hyman, David Kurland, Connor Oates, Siobhan Kibbey, Poorna Sreekumar, Chris Le, Michelle Giglio, and Carol Greene. Human disease ontology 2018 update : classification, content and workflow expansion. *Nucleic acids research*, 47, 11 2018.

[78] Shamik Sural, Gang Qian, and Susnata Pramanik. A histogram with perceptually smooth color transition for image retrieval. pages 664–667, 01 2002.

[79] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 12 2017.

[80] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Matrix factorization and neighbor based algorithms for the netflix prize problem. pages 267–274, 01 2008.

[81] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. ieee, computer journal, 42(8), 30-37. *Computer*, 42 :30 – 37, 09 2009.