Service-Based Approach and Intelligent Agents for Recommendation and Crisis Management : Application to the Analysis and Management of Emerging Diseases

Firas Zouari

N° d'ordre NNT : 2023LYO30023

THÈSE DE DOCTORAT DE L'UNIVERSITÉ JEAN MOULIN LYON 3

Membre de l'université de Lyon

École doctorale n° 512 - InfoMaths, Informatique, mathématiques

Discipline : **Informatique**

Soutenue publiquement le 27/06/2023, par

Firas ZOUARI

Service-Based Approach and Intelligent Agents for Recommendation and Crisis Management:

Application to the Analysis and Management of Emerging Diseases

Laboratoire de recherche :
LIRIS, Laboratoire d'informatique en image et systèmes d'information

Codirectrices de thèse : **Mme Chirine GHEDIRA GUEGAN** et **Mme Nadia KABACHI**

Devant le jury composé de :

Mme Chirine GHEDIRA GUEGAN
Professeure des universités, université Jean Moulin Lyon 3. Codirectrice de thèse

Mme Nadia KABACHI
Maîtresse de conférences,
université Claude Bernard Lyon 1, Villeurbanne. Codirectrice de thèse

M. Richard CHBEIR
Professeur des universités,
université de Pau et des Pays de l'Adour, Anglet. Président du jury

Mme Nada MATTA
Professeure des universités, université de technologie de Troyes. Rapporteure

M. Mahdi ZARGAYOUNA
Chargé de recherche HDR, université Gustave Eiffel, Marne-la-Vallée. Rapporteur

M. Elhadj BENKHELIFA
Professeur, Staffordshire University, Stafford (Royaume-Uni). Examinateur

M. Mohamed Hedi KARRAY
Professeur des universités, école nationale d'ingénieurs de Tarbes. Examinateur

---

Invités hors jury :

M. Claude DUSSART
Professeur des universités - praticien hospitalier,
université Claude Bernard Lyon 1, Villeurbanne

M. Khouloud BOUKADI
Professeur associé, université de Sfax

Acknowledgement

This thesis is a culmination of years of hard work, dedication, and passion, and it would not have been possible without the support and guidance of many people.

I would like to take this opportunity to express my gratitude to my supervisor, Mrs. Chirine GHEDIRA GUEGAN, for her invaluable mentorship, expertise, and encouragement, but also for trusting me during this entire thesis. I would like to acknowledge the beneficial contributions and advice of Mrs. Nadia KABACHI. I would also like to thank Mrs. Khouloud BOUKADI for her availability, contribution and advice that helped me achieve this work.

I would like to thank the members of the jury who evaluated my work: Mrs. Nada MATTA and Mr. Mahdi ZARGAYOUNA for reviewing my dissertation, as well as Mr. Elhadj BENKHALIFA, Mr. Richard CHBEIR, Mr. Claude DUSSART, and Mr. Hedi KARRAY for being part of my jury and their interest in my work.

I would like to extend a special thanks to my parents, my wife, my sister, and my family-in-law for their unwavering support, love, and encouragement throughout my academic journey. I would like to acknowledge the contributions of all my friends, who have supported me in countless ways

I would like to thank all the researchers, teachers, colleagues, friends and staff of the of the LIRIS laboratory for the good moments we had together.

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.

Contents

Glossary

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

List of Figures

List of Tables

INTRODUCTION

CHAPTER 1

Introduction

> All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason. There is nothing higher than reason.
>
> Immanuel Kant

Context

For decades, we have been witnessing positive growth in economic exchanges and migration flows accelerated by globalization. Meanwhile, human migrations and natural disasters have been the cause of many crises, whether economic, political, societal, or public health, due to the emergence of new diseases and the spread of infections in different geographical areas.

Thus, efforts have been focusing on innovative and efficient solutions to face this kind of situation in some serious cases by performing crisis management and detection of abnormal phenomena, especially in the medical field following emerging diseases and epidemics, as was the case with the COVID-19 pandemic. A crisis management system is a structured approach to management and response to emergencies. It encompasses a set of policies, procedures, and protocols to ensure that an organization can effectively respond to any type of crisis, such as natural disasters, cyberattacks, or public health emergencies. Such a system collects data from different sources, including social networks, the Internet of Things (IoT), and institutional databases, which play an important role in generating data related to a domain, including health, in this thesis. These data are often very dynamic and have a very high growth factor, which makes existing infrastructures unable to

handle the huge data volume. Thus, different data repositories, such as Data Lakes and Data Lakehouses, have been proposed to better exploit the masses of data that can be collected in batch and streaming. These data repositories greatly facilitate data integration, whether the data is structured, semi-structured, or unstructured. On the other hand, current crisis management systems still lack real autonomy and intelligence since they often have to rely on predictable and pre-programmed situations. Considering the presented context, our objective is to propose a generic approach to help manage crises based on multi-source data.

To better illustrate the challenge, we present the following two scenarios: Suppose an expert in public health strategy would like to take action following propaganda launched on Twitter about a new virus. Thus, he needs to use a system that allows data collection through different medical data sources and social networks, such as Twitter, Facebook, etc. This system must check the relevance of the new information published on social networks to recommend the most convenient preventive actions. Similarly, a health professional wants to consult medical recommendations about an unknown disease characterized by flu-like symptoms. Considering several countries are in a severe health crisis, he/she wants to consult the latest recommendations to be taken into account to face this disease. In this case, a system that analyzes institutional databases, sensor data, and patient records would help make medical suggestions for handling the situation. In addition to generating recommendations, such a system must provide explanations tailored to each user's role. For example, the strategic expert would want explanations in the form of statistics and arguments (like the situation in neighboring countries and concrete examples) to understand the recommendations. On the other hand, the health professional might be interested in explanations in the form of medical information to better understand the situation. Based on these two scenarios, several questions arise, among them: (1) How can a system adapt and tailor actions to the purposes of several users and the global context? (2) How will this system evaluate the relevance of the data?

The remainder of this chapter explores the thesis from multiple dimensions, including the research questions, the contributions, and the dissertation organization.

Research Problem statement

In this context, the research problem that we address in this thesis is:

> How to generate recommendations intelligently and autonomously on multi-sources, heterogeneous, uncertain, and complex data collected in a data lake/lakehouse without prior knowledge?

We decompose this research problem into three sub-problems.

RP1 - Identifying the data management steps and evaluating and comparing the existing approaches for each step.

The quantitative explosion of data has forced researchers to find new ways to see and analyze data by discovering new ways to capture, retrieve, share, store, and present data. In this context, the data lakes and data lakehouses were proposed as a low-cost data repository that overcomes the limitations of classical data repositories (See Chapter 2). This emergent technology is denoted by its ability to carry massive data in a very heterogeneous form (i.e., non-structured, semi-structured, and structured) at the same time [3]. This valuable characteristic has increased data lake/lakehouse usage, especially for implementing real-time applications when there is a time constraint preventing from performing a process of unifying data schemes before loading data to a repository. Unfortunately, despite the bright side of the data lake, its adoption faces many difficulties for several reasons. First, ingesting data from multi-sources raises many questions about the quality of the ingested data. Data quality addresses "fitness for use" which consists of assessing the quality of the data according to its usage context. Data quality may be appropriate for one use but may not be of sufficient quality for another use [4]. In addition to ensuring data and source quality, the privacy of data carried in the data lake/lakehouse is raised by professionals working with sensitive data (e.g., healthcare, government, social, etc.). To overcome these challenges, scientists found mechanisms to curate data, ensure its quality, and preserve privacy. This work aims to categorize the existing literature and identify the open issues regarding data management in a data lake. Therefore, we consider the following research questions:

- What are the criteria for evaluating and comparing data management processes related to data lakes/lakehouses?
- What are the limitations of existing approaches related to each stage of data management?

In the following sections, we shed light on the underlying research problems related to data management and the recommendation of measures.

RP2 - Managing multi-structured data collected in batch and streaming

Data are often characterized by heterogeneity, complexity, and uncertainty due to their different structure forms (e.g., unstructured data, semi-structured data, and structured data) and the diversity of their provenance. With the growing amount of data, there is an increasing need to implement a specific data management process for big data to overcome these problems. For instance, the healthcare field is facing an explosion in the volume of data generated by advances in genomics and proteomics, combined with laboratory data, patient history, clinical research data, and the health blogosphere. Also, various tools are trying to interact with each other to form increasingly complex platforms and multiply the

heterogeneity of the data. Accordingly, processing this data alone requires much work that classic data management tools cannot do [5]. For this purpose, the data management process may contain several steps, such as data integration, data analysis, and data curation. Data curation is a crucial step that relies on managing and promoting data use from its point of creation by performing enrichment or updating to keep it fit for a specific purpose. Accordingly, a successful decision-making process requires the success of each data management step, including data curation. However, the latter may require too much time and involve more effort, exceeding 50% of the total effort and processing time [6, 7, 8]. On the other hand, critical decision contexts, such as crises, are generally evolving and may impose restrictions on the execution time and the accuracy of information system outcomes. Thus, it is necessary to adapt the data curation step according to the decision factor changes in order to not negatively impact the overall system performance and to align with users' expectations and their decision contexts. Yet, it is challenging to perform data curation for multiple kinds of data sources simultaneously while considering user needs related to different contexts. Therefore, we consider the following research questions:

- How to perform data curation for multi-kind data collected in batches and streaming?
- How do I consider the needs of different users in different contexts while performing data curation?

Once data are collected and curated, it is then possible to get insights from the data. Thus, we present hereafter the second scientific problem focusing on generating recommendations for multi-users with different needs.

RP3 - Generating recommendations for multi-users with different needs

Artificial intelligence (AI) has become a key topic in our personal and professional conversations, political debates, industry conferences, and digital conferences. For instance, AI is increasingly used in the medical field for different purposes, such as patient data and medical image analysis, tumor detection, and health outbreak detection. Hence, AI has shown promising performances in various tasks that are close to those of experts or even better. Nevertheless, although AI models provide good performance, some models, like deep learning models, act like black boxes and show low interpretability. However, experts in certain sensitive domains, such as healthcare and economics, need explanations of the outputs of a recommendation model to better understand the choices and attribute their confidence. Moreover, users may have different needs and perspectives on explanations. To better illustrate the scientific challenge, we present the example of Bob, an infectious disease specialist, and Alice, a deputy senior defense and security officer at the Ministry of Health, using a system to predict and manage health crises. Following the prediction of a threat, this system recommends health measures to treat the predicted disease. Thus, the system has to generate and adapt explanations for each user role. Bob may be interested in medical information such as symptoms, the nature of examinations, and treatments. At

the same time, Alice may be interested in other explanations, such as statistics and the situation in neighboring countries.
Therefore, we consider the following research question:

- How to recommend measures for crisis management while providing explanations tailored to different user roles in different decision-making contexts?

RP4 - Experimenting with crisis management system

The presented data management steps should be encompassed in a crisis management system that helps professionals make decisions by analyzing different data sources to recommend convenient measures (i.e., action to be taken). Thus, it is essential to analyze the performance of the proposed solution to check its practical effectiveness and cost.

Therefore, we consider the following research questions:

- Does the proposal present an effective solution in practice?
- What performance does the test system provide, in terms of processing time, scalability, and cost?

Contributions summary

This dissertation has four contributions, and each one deals with one of the research subproblems that were set up.

C1 - A state-of-the-art of the data management in the data lake/lakehouse in response to RP1.

To have a global view of the data management process, we introduce a systematic mapping that covers data management steps such as curation, quality evaluation, privacy preservation, and prediction using data stored in the data lake. Indeed, it is necessary to perform a first comprehensive analysis of the mechanisms used to manage data in a data lake/lakehouse and perform research using curated data while ensuring the quality of the data and preserving their providers' privacy. We defined criteria such as domains, the proposal type, and the data management step to provide a classification scheme for the studied articles. After that, we provide an analysis of the reviewed articles. This analysis shows the open issues related to data curation, quality evaluation, privacy preservation, and prediction for data stored in data lakes.

This contribution is the basis for the following publication in the 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 [9]

C2 - An adaptive service-based curation approach in response to RP2.

We propose an adaptive data curation approach implemented as a data curation framework for batch and streaming multi-structured data sources while considering the decision maker's needs. The proposed framework is service-based and encompasses the stages of data curation (i.e., data source collection, data quality evaluation, data source characterization, and data curation). The data curation process encompasses curation service composition and data curation modules that employ a library of curation services, where each service presents a curation task. The curation framework relies on ACUSEC, an approach for adaptive curation service composition that we propose, and composes the curation services adaptively to the decision process features to optimize the data curation process in terms of execution time and alignment to user needs. Subsequently, the framework's data curation module invokes the composing services.
Regarding ACUSEC, our original contribution relies on artificial intelligence techniques, particularly machine learning, to generate data curation service composition schemes adaptively. For this purpose, our approach considers the user's functional requirements, such as data source characteristics; non-functional requirements, like user preferences and constraints; and the decision context. Mainly, our proposed framework takes advantage of reinforcement learning as a practical solution that can learn over time to make increasingly effective decisions in a dynamic environment like the one of the data lakehouse.

This contribution is the basis for the following publications in the Cluster Computing Journal, in the IEEE International Conference on Web Services (ICWS 2021), the International Conference on Web Information Systems Engineering (WISE 2022) and the Clustering Computer Journal:

- 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 [10]
- Firas Zouari, Chirine Ghedira Guegan, Nadia Kabachi, Khouloud Boukadi: Towards an adaptive curation service composition based on machine learning. ICWS 2021: 73-78 [11]
- Firas Zouari, Chirine Ghedira Guegan, Khouloud Boukadi, Nadia Kabachi: A service-based framework for adaptive data curation in data lakehouses WISE 2022: 225-240 [12]

C3 - An explainable measures recommendation approach in response to RP3.

We proposed a deep learning-based model for recommending actions for crisis management. Thus, we have leveraged deep learning techniques, to design and propose a multi-output classification-based recommendation model in which each output predicts the stringency level of each proposed measure. Although the performances obtained by deep learning models seem satisfactory, these models (learning models) act as a black box and are not very interpretable. Therefore, we have focused on improving the interpretability of

our models' results. To this end, our contribution is guided by Explainable Artificial Intelligence (XAI), a paradigm related to the explanation of black box models such as deep learning. More precisely, we propose a semantic approach to explain the black box models associated with our recommendation model.

C4 - An evaluation through the implementation of a test system in response to RP4.

We implement our proposed contributions to constitute a multi-user crisis management system that analyzes heterogeneous data collected from diverse sources (e.g., sensors, health institute databases, etc.). Specifically, this system employs a data lakehouse to store the collected data. Then, it curates and analyzes data to recommend appropriate measures for crisis management. Moreover, the implemented system explains the choice of the recommended health measures in an adapted manner for each user role. Accordingly, we examine the effectiveness of the proposed contributions through the different system modules (i.e., data curation, crisis prediction, recommendation of measures and explanation).

Organization of the dissertation

The remainder of this dissertation is organized into five chapters: Chapter 2 gives an overview of the basic concepts related to our work, such as data lakes, data curation, explainable artificial intelligence, etc. Moreover, it analyzes the state-of-the-art of existing crisis management and data management approaches. Chapters 3 and 4 detail our proposed contributions, and Chapter 5 concludes the present dissertation and overviews future endeavors. Below, we give more details about the aspects related to each chapter (See Figure 1.1):

- **Chapter 2 - Basic concepts, state of the art and positioning:** presents and overviews the basic concepts and works related to our proposed contributions. The chapter sets the concepts and vocabulary that serve as a knowledge background for the state-of-the-art and the proposed contributions. Moreover, it surveys the existing health crisis management approaches and data management steps (i.e., curation, prediction, and recommendation) and outlines their strengths and weaknesses. Then, it analyzes the presented state-of-the-art based on the above analysis criteria.

- **Chapter 3 - Adaptive data curation for batch and streaming data** presents the proposed approach for adaptive multi-structured batch and streaming data curation. In particular, our proposed approach is implemented in a service-based framework, encompassing multiple curation services used to generate a service composition scheme. The proposed framework identifies the main data source characteristics that guide the data curation using DAta chaRacterization Quality evAluation oNtology (DARQAN) that we propose to evaluate the data quality and characterize the data sources. Hence, the Adaptive Curation Service Composition (ACUSEC) approach employs the extracted characteristics to compose curation services according

to the user's functional and non-functional requirements. This chapter also details the experiments elaborated to evaluate the performance of our proposal and prove its effectiveness.

- **Chapter 4 - Recommendation and explanation of measures for crisis management** presents the proposed approach for the recommendation of measures for crisis management. In particular, it presents the proposed model, which recommends measures to manage the predicted crisis for different user roles. It also details the semantic-based approach that adaptively explains the choice of the model's recommendations. This chapter also introduces the experimental protocol that assesses the effectiveness of our contributions.

- **Chapter 5 - Conclusion and future work:** summarizes and concludes this thesis by presenting the work carried out, future endeavors related to continuing this work, and enumerating research perspectives addressing the open issues.

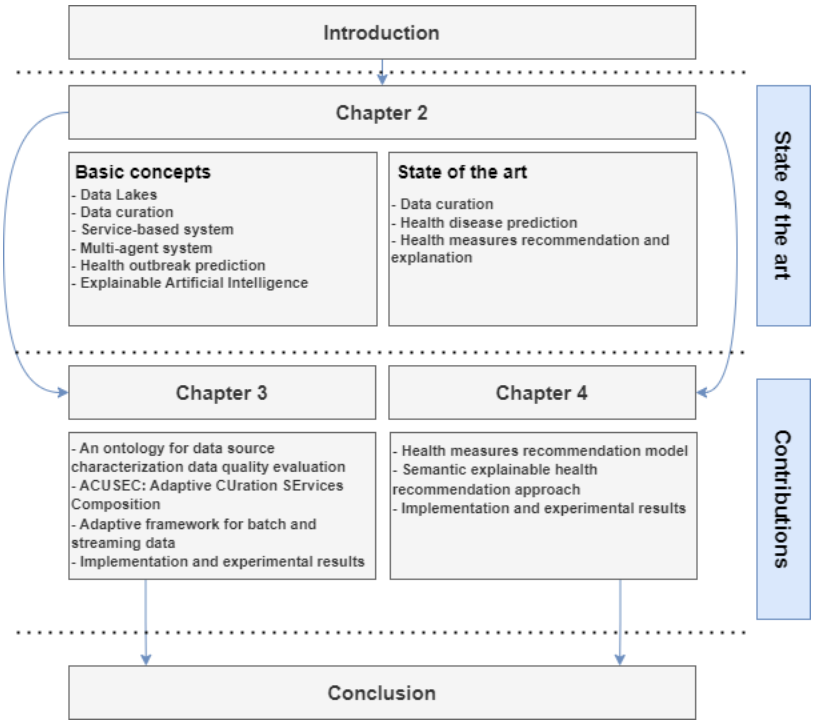


Figure 1.1: Structure of the thesis

CHAPTER 2
Basic concepts, state of the art & positioning

CHAPTER 2

Basic concepts, state of the art & positioning

> If I have seen farther than others, it is because I have stood on the shoulders of giants.
>
> Isaac Newton

Contents

Introduction

Crises could become worse if preventive actions were not taken in time through the intervention of competent human resources and rapidly mobilizable material resources. Thus, the response should be rapid, focused, and coordinated in transboundary crises. Indeed, chemical threats or environmental disasters (e.g., volcanic eruptions) can quickly spread beyond a country's borders or national response capabilities. For instance, the recent COVID-19 pandemic presented a significant health crisis that requires large-scale coordinated action. For this purpose, health crisis management systems need to analyze different data from various sources and providers to predict health crises and mobilize resources to deal with the predicted disease. Hence, these systems may deal with massive, heterogeneous, and complex data. Accordingly, we adopt data lakehouses as a repository to overcome the challenges related to massive heterogeneous data collection and storage.

Data lakehouses combine the key aspects of lakes and warehouses, allowing unifying storage using the single-repository data warehouse model and ensuring data lakes' analytical flexibility. Nevertheless, the crisis prediction and management process still involves unresolved data cleaning, processing, and recommendation challenges that we discuss and tackle throughout this manuscript.

In this chapter, we explain the basic concepts related to our proposals, like data lakehouses, data curation, crisis prediction methodologies, and explainable artificial intelligence. Similarly, we review the work that has been proposed regarding data curation and measures recommendation and explanation, and we discuss their limits. Then, we conclude this chapter by summarizing our study and discussing the position of our contributions.

2.1 Data lakehouses

Nowadays, emerging technologies led to the emergence of a massive amount of data. Big data encompasses too large datasets to be processed by traditional database systems. Hence, they require new processes and methods for storage, integration, processing, and analysis. Due to the heterogeneous aspect of big data, repositories, like classic data warehouses, need to be enhanced to become adequate for big data storage. Indeed, such repositories require the execution of an ETL (Extract - Transformation - Load) process before storing data, which may be costly for big data storage. Accordingly, the ELT (Extract - Load - Transformation) was proposed to overcome the ETL limits. We present an overview of the existing data integration approaches and the data repositories.

2.1.1 Data integration approaches

We identify two approaches proposed to store and integrate data in data repositories. Figure 2.1 depicts an overview of the two data integration approaches. The ETL and the ELT approaches combine three steps that we detail in what follows:

- **Extraction:** This step ingests raw data from infrastructures, software, and applications according to specific rules to be stored in the repository.
- **Transformation:** the transformation step sorts and normalizes the data to be accessed for later operations like data analysis, reporting, visualization, etc.
- **Load:** The load step stores the data in the data repository. Basically, the moment when the transformation and load steps are carried out makes the difference between the two approaches.

The ELT approach is an approach that solves the problems associated with storing large volumes of data. Indeed, the transformation step is considered the most complex step in the traditional ETL process. For this purpose, the transformation step is postponed after data storage in the ELT to optimize the storage costs. Accordingly, the data are stored in the repository in their raw format and could be transformed when needed to generate insights.

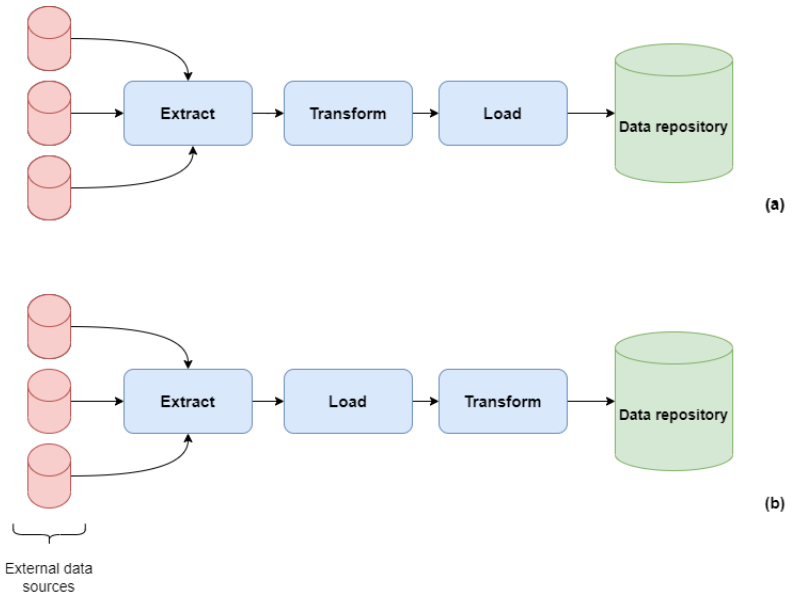


Figure 2.1: Overview of the ETL (a) and ELT (b) approaches

2.1.2 Data lakehouses vs Data lakes vs Data Warehouses

A data warehouse is a central repository of data stored from different sources inside and outside the enterprise. It collects data from various sources, both internal and external and optimizes data retrieval for business purposes. Contrary to classic databases, a data warehouse stores historical, structured, non-volatile, object-oriented data which makes it, basically, designed for data analysis in the context of decision-making. Nevertheless, data collected in databases and data warehouses need to be cleaned and prepared before being stored, which may handicap operations in some decision contexts in which time may be critical. Thus, data lakes were proposed to resolve this problem. As depicted in Figure 2.2, a data lake is a centralized storage location encompassing big data in a raw, granular format from numerous sources. Accordingly, data lakes rely on the ELT process to store data and optimize storage costs. Regardless, data lakes do not afford the same performance in terms of management and optimization as data warehouses. Thus, data lakehouses are proposed as a solution that combines the key strengths of the two data architectures above (i.e., data warehouses and data lakes). Specifically, data lakehouses ensure low-cost storage in an open format accessible by various systems and robust management and optimization features. Unlike data lakes, data lakehouses are directly-accessible storage that provides traditional DBMS features like ACID transactions, data versioning, auditing, indexing, caching, and query optimization [13]. We present a comparison between different data repositories in Table 2.1

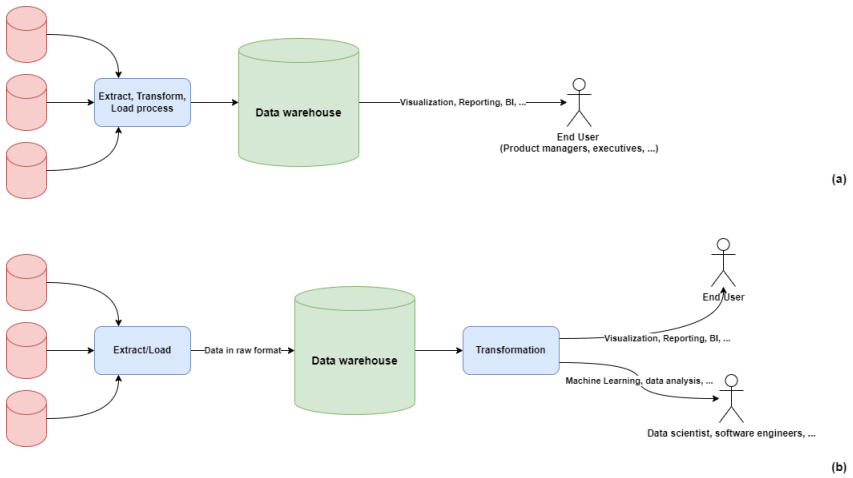


Figure 2.2: Difference between the data warehouse (a) and data lakes (b) architectures

Table 2.1: Comparison of different data repositories

	Data warehouse	**Data Lake**	**Data Lakehouse**
Data format	Processed	Raw	Raw
Data integration	ETL	ELT	ELT
Data quality	Curated	Not guaranteed	Not guaranteed
Schema	Schema-on-write	Schema-on-read	Schema-on-read
Queryable	Yes	No	Yes
ACID Transactions	Yes	No	Yes
Maturity	Mature	Immature	Immature

2.2 Service-based systems

2.2.1 Service Oriented Architecture

Definition

A Service-Oriented Architecture (SOA) is composed of discoverable loosely coupled services. In this architecture, service providers and consumers are independent, and users may compose services into a business process to create new services. [14]

SOA main concepts

We present the main concepts that constitute SOA architecture.

- **Service:** a service represents an independent function or operation that ensures a task from a business process. We distinguish several examples of services such as

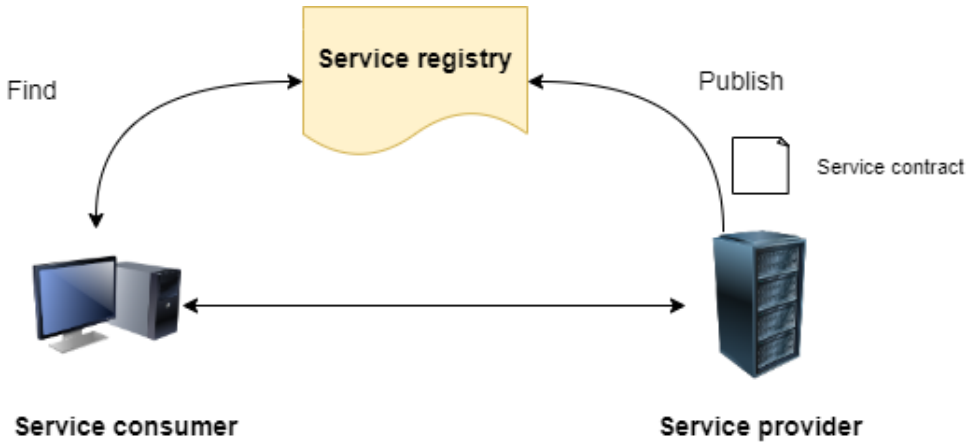


Figure 2.3: Interaction between SOA architecture actors

Web services (e.g., REST, SOAP), Data services (e.g., JDBC), and JAVA Message services.

- **Message exchange:** the messages between services are exchanged according to protocols and patterns and presented in XML or JSON format.
- **Service contract:** service contracts may be considered as metadata in the SOA architecture. Indeed, it is the most basic interaction in this architecture and defines the details of the service offered by a provider to a consumer.

SOA Actors

The service-oriented architecture encompasses two basic actors:

- **Service consumer:** the service consumer seeks the service in a service registry and then invokes a service from one of the service providers.
- **Service registry:** it is a repository or a service broker that plays the role of intermediary between service consumers and providers. It encompasses a set of services offered by providers.
- **Service provider:** the service provider is the service generator. Accordingly, it publishes the service in the service registry after creation.

Figure 2.3 depicts the interaction between the different actors.

SOA principles

The Service Oriented Architected is characterized by a set of design principles and practices that we present through the following points:

- **Standardized:** each service has a service description that provides information about the purpose of the service.

- **Loosely coupled:** the SOA architecture is loosely coupled since it contains services with minimum dependencies on each other. Accordingly, the fault of one service does not impact the general function of the system.
- **Reusable:** this architecture divides logic into services to promote re-use.
- **Composable:** the SOA architecture composes large problems into smaller ones.
- **Autonomic:** the services control the encapsulated logic.
- **Stateless:** the services remain stateless and do not keep data from one state to the other.
- **Abstract:** the services encapsulate the logic and do not provide how they perform its functionality.
- **Discoverable:** the services may be discovered via a service registry that contains information about them.

2.2.2 Service composition

Definition

Service composition is the process of combining atomic services into added-value composite services. It integrates services to achieve a specific task and thus adds more value [15]. Moreover, it relies on service selection that selects the most suitable services from a pool of available services to match a user functional and non-functional requirements and constraints [16].

Composition types

We distinguish two types of service composition:

- **Static service composition:** the control and data flows are defined by the user during design time.
- **Dynamic service composition:** the control and data flows are generated automatically at run-time.

Service orchestration vs choreography

To perform service composition, we identify the two following approaches[16]:

- **Orchestration:** it controls the services and the interactions actively, like the musicians of an orchestra do. Nevertheless, the disadvantage of this approach is that the controller must communicate and wait for each service response which may impact the system's performance.

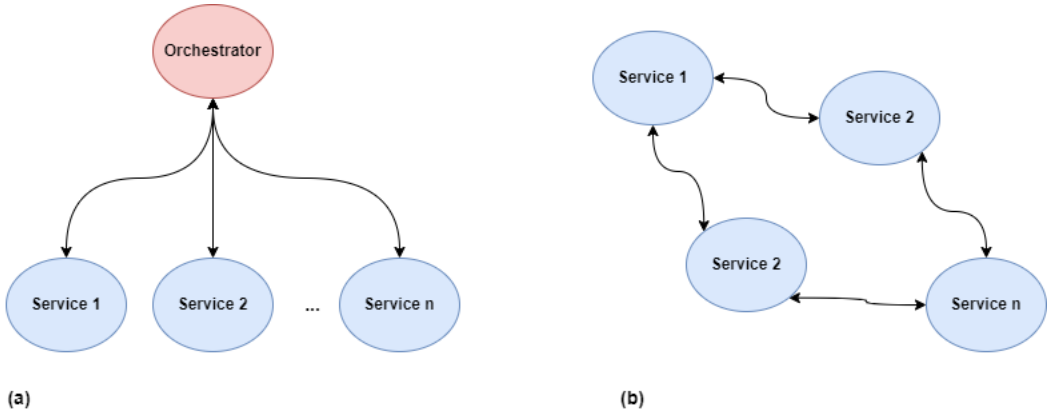


Figure 2.4: Overview of the orchestration (a) and choreography approaches (b)

- **Choreography:** the choreography creates a pattern to be followed by the services without supervision. Hence, it allows for creating faster, more consistent, agile, and efficient systems.

2.3 Multi-agent systems

2.3.1 Definition

We consider an agent as an entity having mental components such as beliefs, capabilities, choices, and commitments [17]. Accordingly, intelligent agents keep performing three actions continuously[18]:

1. Monitoring the environment to detect dynamic conditions
2. Action to affect conditions in the environment
3. Draw inferences and reasoning to interpret perceptions and solve problems

Although both are considered computational units that communicate via messages, agents are different from objects because they have a sense of autonomy and activity. Hence, multi-agent systems are helpful in following all the system dynamics at each moment and in explicitly representing the interactions between the different elements of the system. Thus, they help monitor a system's global behavior and perform simulations, hypotheses, and scenarios. We present in Figure 2.5 an overview of the different components and their interactions in a multi-agent system.

2.3.2 Intelligent agent properties

An intelligent agent is characterized by the following properties [19]:

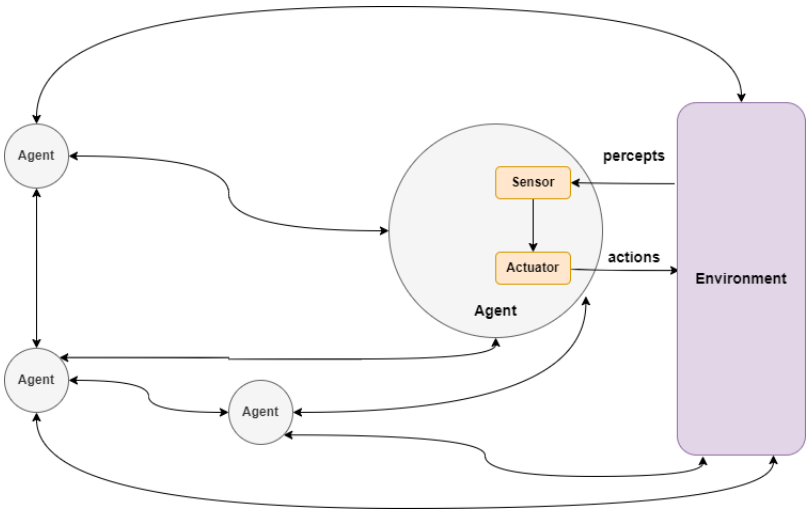


Figure 2.5: Multi-agent system

- **Autonomy:** it concerns the autonomous character of an agent. Specifically, an agent interacts without the direct intervention of humans or other agents.
- **Situatedness:** it states that an agent performs one or some actions that change its environment when receiving input from it.
- **Adaptivity:** it measures the extent to which an agent can react to changes within its environment by learning from its own experiences, environment, and interactions.
- **Social abilities:** it concerns the ability of an agent to interact and exchange with other agents.

2.3.3 Agents typology

We present many typologies proposed to classify agents:

- **Collaborative/Coordinative agents:** these agents possess a non-trivial ability for coordination, autonomy, and sociability.
- **Hybrid agents:** these agents combine the deliberative and the reactive aspects.
- **Intelligent agents:** these agents have the intentional ability to perform reasoning and generate inferences.
- **Wrapper agents:** this type of agent is designed mainly to interact with non-agents.

2.3.4 Types of interactions in multi-agent system

Werner et al. [20] identified the following four interaction types that may occur in a multi-agent system:

- **No communication:** the agents do not communicate; they either interact through the perception of the environment or reach their goal without external help.
- **Sending signals:** the agents synchronize themselves by sending coded messages.
- **Sending plans:** the transfer of information concerns the tasks and beliefs of the agents.
- **Sending messages:** this mode of communication allows agents to exchange their intentions and needs.

We also identify another taxonomy of agent interactions: direct and indirect communications.

Direct interaction

Direct interaction is the classical approach to addressed communication, in which a sender sends a message to a receiver located by its address. It may have different forms, such as peer-to-peer communication and global or restricted diffusion. Direct communication is characterized by its simplicity. Figure 2.6 depicts direct communication in a multi-agent system.

Indirect interaction

Indirect interaction is done by changing the environment's states or adding a blackboard or shared memory. The blackboard is a shared search space where the results obtained by the agents are registered. The blackboard can be read and/or written.

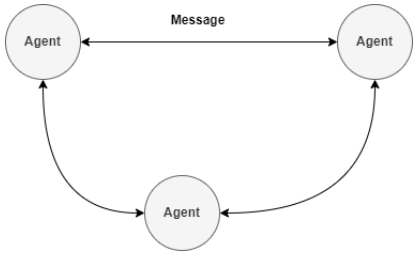


Figure 2.6: Direct communication in a Multi-agent system

2.4 Data curation

Data lakehouses can be an effective solution in today's urgent contexts and situations, such as crises, that require real-time data collection and analysis. Nevertheless, data heterogeneity and complexity remain critical challenges for data lakehouses. Thus, data wrangling, part of data curation, is necessary to enhance data quality before data analysis or visualization. For this purpose, data curation ensures managing and promoting data use from

its point of creation by enriching or updating it to keep it fit for a specific purpose [21]. Accordingly, data curation provides more information about the provenance of the data, the original context of measurement and use, and the object of observation to facilitate the re-use of the data [22]. In the following sections, we present the basic concepts and analyze existing works related to curation.

2.4.1 Data structures and curation

In this section, we present the different data structures which may have specific requirements in designing the data curation process.

Unstructured data

Unstructured data are characterized by the absence of any identifiable structure for data representation. Accordingly, they cannot be stored in rows and columns like relational databases for example [23]. For instance, unstructured data sources may be presented in different forms, such as videos, images, customer interactions, and plain texts. This form of data structure may have the advantage of reducing the effort of its classification. Nevertheless, unstructured data are difficult to analyze since there is no data model to parse, which makes data navigation hard. As there is no respected pattern to present data, data related to entities may have different representations (e.g., IDs of two persons may be represented via different patterns like plain text - XXXXX or between brackets - (XXXXX)). Hence, unstructured data remains a challenge for conventional software to collect and analyze this form of data, especially in the big data era, in which 90% of the collected and generated data is unstructured or semi-structured, which we present in the next section.

Semi-structured data

Semi-structured data are schemaless data that do not respect a specific data model but have some structure. In other words, there is no rigid schema to be respected. Same as unstructured data, semi-structured data cannot be stored in rows and columns. However, they respect some hierarchy and are presented as a graph or tree-like structure[24]. Accordingly, data are represented as entities linked together using edges and characterized by attributes or properties. Hence, this characteristic offers data flexibility since "the schema" of the data can be easily changed. Nevertheless, the lack of a rigid schema makes the data difficult to store due to the changing schemas, which may evolve over several versions. Moreover, semi-structured data sources are not easily queried compared to structured data sets. As examples of semi-structured data, we cite XML, JSON files, e-mails, streaming textual data that is usually presented as semi-structured data, and web pages. For instance, XML files group records presented as a hierarchy of entities encompassing different attributes and values.

Structured data

Structured data sources conform to a predefined schema[1]. This form of data may be stored in rows and columns, which makes it highly organized. The most common form to represent structured data sources is the relational databases managed by a management system known as a Relational Database Management System (RDBMS), such as MS SQL Server, MySQL, PostgreSQL, and Oracle Database. Accordingly, this structure is highly managed and easy to understand, navigate, and query via a standardized language known as SQL. Contrary to unstructured data sources, structured data are easier to access by tools [25].
Nevertheless, structured data sources are limited since most data often exists as unstructured and semi-structured data. Moreover, structured data sources are characterized by low flexibility since they respect a rigid schema, which makes their evolution harder. Accordingly, changes in data requirements require updating the data structures, which is expensive in terms of time and resources. Considering structured data, it is decomposed following normalizing principles to ensure referential integrity and reduce data anomalies (e.g., duplications, errors, violations). Accordingly, data is represented through different entities presented as tables linked via relations.

2.4.2 Curation scopes

In this section, we present the different scopes concerned by curation.

Data

Data curation encompasses curation tasks that are dedicated to data cleaning and enrichment. Indeed, data curation covers data wrangling tasks that encompass data repair and deduplication, which prepare data to be processed in later operations. As the quality of input data reflects the quality of the outcomes of a system, data curation encompasses enrichment tasks that ensure the organization and the maintenance of data sets to be accessed and used by users looking for information. As aforementioned, data lakes/lakehouses hold data collected from different sources that may be complex, uncertain, and heterogeneous. Thus, semantic enrichment, such as linking with external knowledge bases, is needed to promote the quality of outcomes.
On the other hand, contextualization via data profiling, for example, is required to organize the data lake/lakehouse and, therefore, facilitate its indexing and cataloging. Accordingly, data curation is a primordial step that prevents the data lake/lakehouse from turning into a data swamp. As for data swamp, it consists of an unmanaged data lake, making it inaccessible to users or providing little value.

Metadata

Metadata curation is one of the most needed operations to maintain the organization of a data lake. Indeed, metadata management creates a catalog to index the datasets grouped

[1] https://www.ibm.com/cloud/blog/structured-vs-unstructured-data

into the data lake. Accordingly, the system can identify the required datasets for analysis when a data analysis process is executed. Contrary to data warehouses in which the metadata management is well held, data lakes lack metadata management and turn quickly into a data swamp if the metadata management is not well defined. For this purpose, curation tasks such as metadata modeling and extraction are required to ensure the organization of a data lake. Indeed, metadata modeling focuses on the definition of models for dataset representation and identification. For instance, different organization zones (e.g., raw data, curated data zones, etc.) may be created into a data lake to arrange datasets according to the end user needs. On the other hand, metadata extraction extracts the desired metadata from datasets according to defined rules to identify datasets' characteristics (e.g., format, author, etc.). Nevertheless, metadata management is included in data lakehouses, contrary to data lakes.

Schema

As data lakes and lakehouses ingest very heterogeneous data sources from several sources, these data may be presented in different formats, like unstructured data sources. The latter need to be transformed to be used in later analysis operations. However, the unstructured data sources do not have a defined schema and may not respect a specific data representation format. For this purpose, data curation encompasses tasks dedicated to schema curation. Schema extraction tries to identify a data representation to extract a schema for the dataset. While schema mapping and extraction seek links between a dataset and the related datasets grouped in the data repository.
Moreover, as these data repositories keep ingesting data all the time, the datasets' contents and schema may evolve and change over time. Hence, schema evolution tasks keep the datasets up to date by evolving the schema of the datasets each time. This feature is a built-in data lakehouse.

2.4.3 Data curation tasks

By analyzing the literature, we identified several tasks dedicated to data, metadata, and schema curation. Indeed, these tasks need to be re-arranged in a specific order to create a curation pipeline. For instance, a curation pipeline may be constituted from tasks arranged in the following order, "POS Tagging $\rightarrow$ Stem identification $\rightarrow$ Linking with classes of ontologies". POS tagging is the association of words in a text with the corresponding grammatical information, such as gender. Then, this pipeline identifies the stems of the nouns. Subsequently, these latter are linked with classes from ontology to perform semantic enrichment. Thus, we classified the curation tasks into three categories according to the curation purpose.
The first category is data curation, consisting of three sub-categories: contextualization, data repair, and semantic linking. Contextualization tasks cover NLP tasks aiming to contextualize or perform data profiling. We distinguish tasks like POS tagging, stem identification, named entity identification, etc. Data repair are tasks dedicated to data wrangling, such as missing data repair and identification of erroneous data. Semantic linking encom-

passes curation tasks that aim to perform semantic enrichment, like mapping and matching with external knowledge bases.

As for metadata curation, it encompasses metadata extraction and modeling. Indeed, we identified contributions that analyze data sources to identify and construct metadata, while other works have proposed models to represent metadata.

Considering schema curation, we examined several contributions that extract, match, and map data sources schema. Schema curation tasks are needed to handle schema evolution. These tasks are applied mainly to semi-structured and unstructured data to identify the schema of data sources. Considering schema matching and mapping, the former links an attribute from data source A with another one from data source B, while the latter identifies the common attributes having different representations (e.g., name, type, etc.) in two data sources. For instance, we consider three data sets having the following attributes Person(Name, Age, Profession), Population(NameP, AgeP, Job), and Doctors(Name, Degree, Specialty). Accordingly, we can perform schema mapping to identify the correspondence between the attributes of Person and Population datasets (e.g., the name attribute corresponds to NameP, and Profession is represented by Job in the Population dataset). On the other hand, schema matching links Person and Doctors data sets using the Name attribute to create a fourth dataset representing information about doctors (i.e., it works similarly to the JOINT instruction in SQL).

As data lakes and lakehouses collect batch and streaming data, reliable data curation pipelines must be defined to enhance data quality before being employed in data analysis processes. Indeed, batch and streaming data may require specific data curation tasks according to the data format, which makes them need different data curation pipelines. In fact, data streams may be collected from IoTs that group different sensors. Hence, dedicated data stream curation tasks such as data normalization and standardization are required to ensure streaming data curation. For instance, we may have two sensors that capture temperature in celsius and Fahrenheit. Thus, data curation scales the data to create a unified data representation.

By analyzing the existing curation works, we proposed the taxonomy of the main batch and streaming data curation task categories, depicted in figure 2.7. These curation tasks ensure the necessary operations for data curation. We point out that concept drift detection is devoted to streaming data curation. Concept drift tasks detect the deviation of captured streaming data due to a sensor failure. Nevertheless, the curation tasks within the other categories could be employed for batch and streaming data curation.

We present in the following section the examined works related to data, metadata, and schema curation.

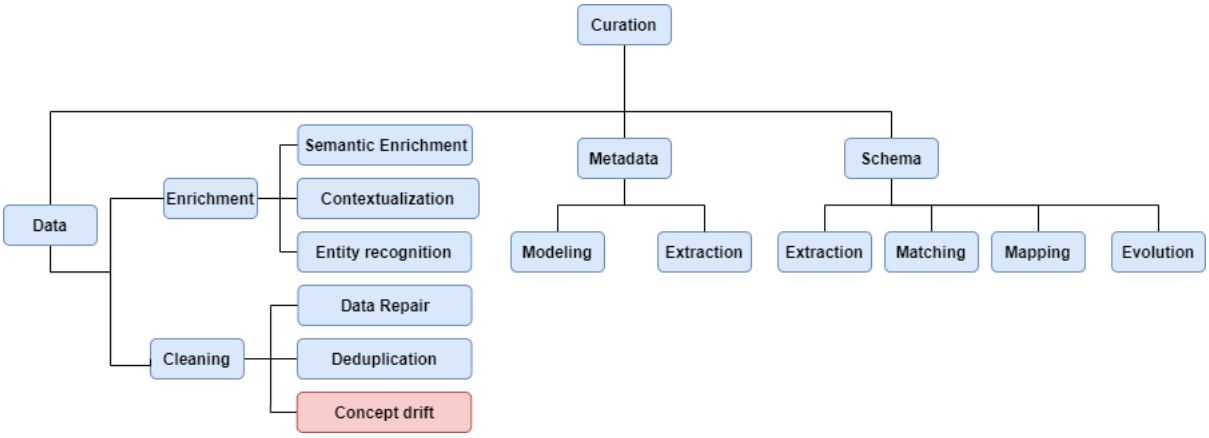


Figure 2.7: Taxonomy of data curation tasks

2.4.4 Existing data curation works review

Data management, generally, and data curation, specifically, have attracted the attention of many researchers. We identified several works in the literature addressing the data curation process. In what follows, we present and examine relevant works addressing the data curation problem. For this purpose, we rely on different criteria that consider the used techniques, the type of data source to be curated, the automation of the approach (i.e., fully automatic, semi-automatic, non-automatic), the adaptation regarding the decision process features, and the nature of the treated data (i.e., batch/streaming data sources). Table 2.2 depicts the examined data curation works

Table 2.2: Overview of the examined data curation works
(Cx: Contextualization, L: Linking, Rp: Data Repair, ML: Machine Learning, S: Semantic Techniques, G: Graph-based techniques, C: Crowdsourcing, R: Rule-based, U: Unstructured data, SS : Semi-structured data, S: Structured data)

Paper	**Architecture**	**Curation tasks**					**Used techniques**					**Data**		
		Data			**Metadata**	**Schema**								
		Cx	L	Rp			**ML**	**S**	**G**	**C**	**R**	**U**	**SS**	**S**
Skluma, 2017[26]	Pipeline	X			X		X					X	X	
Crowdcorrect, 2018[27]	Pipeline	X					X			X			X	
CoreKG, 2018[28]	Service-based	X					X			X			X	
KAYAK, 2018[29]	Framework		X		X				X				X	
Pomp et al., 2020[30]	Framework		X		X	X			X		X			X
Semlinker, 2020[31]	Framework					X		X					X	
Data synapse, 2018[32]	Service-based		X		X	X		X					X	
Hai et al., 2019	Method				X	X					X		X	
VADA, 2019[33]	Framework	X	X	X		X	X				X			X
Lenses, 2015[34]	Method			X		X					X			X
Simonini, 2019[35]	Method					X			X				X	X
DATAMARAN, 2018[36]	Method					X			X				X	

Beheshti et al. [27] presented a data curation pipeline (i.e., a sequence of curation

tasks) that allows analysts to perform social data cleansing and curating. Their approach aims to prepare data for reliable business data analytics. The idea consists of enhancing data curation at each step of the curation pipeline by applying first automatic curation (i.e., data curation using services) and semi-automatic curation (i.e., crowdsourcing, experts annotation) later. The proposed idea was employed in several other works to prove its effectiveness in various use cases, such as [32] and [28]. For instance, Datasynapse [32] is a curation pipeline that curates data for contextualization. Accordingly, the contextualized data are stored in a knowledge-based data lake using CoreKG, a dedicated service proposed by the authors that relies on the curation pipeline to constitute this data lake [28].

Another work proposed by Kanstaninou et al. [33] presented an architecture that contains loosely coupled data preparation components. The latter are constituted from rule-based and machine learning techniques and could be orchestrated dynamically to perform data preparation tasks like matching, data profiling, mapping generation, format transformation, and data repair. Despite the flexibility of this approach, it is devoted only to structured data source curation.

In the context of dynamicity and flexibility, Maccioni et al. presented a framework named KAYAK, which encapsulates graph-based techniques to model pipelines for data preparation [29]. The proposed framework lies between users/applications and the file system layers (i.e., data storage location), and it exposes a set of primitives and tasks for data preparation. Each data preparation task combines a set of primitives representing a straightforward preparation step (e.g., insert dataset, compute joinability). As in [27], the preparation tasks must be combined to accomplish the data preparation task. Based on these tasks, the users can design their data preparation pipeline (i.e., organization of data preparation tasks) represented as a DAG (Direct Acyclic Graph). KAYAK also holds a knowledge base that encompasses metadata about data sources (e.g., attribute names, relation names, etc.), the target schema, and the workflow (e.g., the current state of data preparation, the provenance of information, etc.). The proposed framework also performs schema matching and mapping to maintain schema versioning.

We also examined Skluma [26], an automated system for vast amounts of data processing and deeply embedded metadata, latent topics, relationships between data, and contextual metadata extraction from related documents. To do so, Skluma relies on machine learning techniques to extract metadata from files based on their content and their general file level (e.g., filename, path, size, checksum, etc.).

Pomp et al. proposed ESKAPE[30], an approach for semantic data integration via learning and evolving data representation. Stakeholders may use the proposed approach for semantic enrichment and data repair. For instance, some value ranges (e.g., temperature) may vary according to the context. Hence, ESKAPE may be used to unify data representation and fetch additional information for semantic enrichment. On the other hand, the work proposed by Pomp et al. describes additional required and interesting meta-information with semantic models. As for schema curation, this work unifies data representation via

linking semantic concepts. Thus, it incorporates Autoencoders for an automatic recommendation of data classes by constructing a knowledge graph that combines the different representations for a data instance.

Lenses [34] is a dynamic model for designing data curation tasks based on rule-based probabilistic components called lenses. The authors represent a lens as a component for data processing which may be part of a typical ETL pipeline. In addition, Lenses relies on schema matching to match the data source's schema to a user-defined target schema. For instance, schema matching could be applied to create relations between JSON objects or web tables that may need well-defined schemas.

Moreover, SemLinker[31] handles schema evolution by incorporating semantic technology. Indeed, it constructs a global ontology encompassing the different versions of data schemas. Accordingly, the data representations of each data schema version are linked together semantically.
We also present the approach proposed by Hai et al.[37] for data integration which detects functional dependencies for data stored in heterogeneous sources and, thus, unifies different data representations.

As for the work proposed by Simonini et al., [35], it presents an approach to extract loose schema information, represented as a graph, from a dataset using an attribute-match induction technique. Regarding DATAMARAN [36], it constitutes a tool that automatically extracts structure from semi-structured datasets. To do so, it extracts an extensive collection of structure templates from potential records, presents them as a graph, and then prunes out most of the candidates. Then, it applies two structure refinement techniques to update the structure templates.

2.4.5 Analysis and positioning

Following the above review, we noticed that the literature encompasses diverse proposals that address curation at three levels: data, metadata, and schema. As we rely on data lakehouses to design our solution, we focus on data curation since the metadata and the schema curation are implicitly handled in data lakehouses. Even though some works combined different curation levels, we shed light on works covering metadata and schema curation to have a complete vision of contributions related to all curation levels, in general, and data curation, in particular.

As for data curation, we noticed that most proposed approaches could not be generalized to treat all data source structures by analyzing the works listed above. Hence, they are designed to curate a specific data source format (i.e., unstructured, semi-structured, or structured). Yet, data lakehouses hold various data source formats, which require sophisticated tools to curate data sources. Regarding curation process automation, we identified several approaches that are not fully automatic. In [27], the authors proposed various

services for data curation. Subsequently, automated curation services are combined with crowdsourcing and expert annotation to improve the process of curation. However, their approach can be applied to social data curation, which is generally presented in a semi-structured form. As in [27], the work presented by Kanstaninou et al. [33] aims to prepare structured data sources. Skluma [26], in turn, aims to contextualize unstructured and semi-structured data sources, such as documentation, README files, CSV, and plain text files. Considering ESKAPE[30], it suggests semantic concepts based on structured data source values. Regarding Lenses [34], this framework is dedicated to curating structured data sources.
As for the data curation process automation, our study reveals that the work proposed in [33], [30], and in [26] are fully automatic. In contrast, the other studied contributions are manual, like [29], or semi-automatic, such as [27], and [34]. Nevertheless, as we presented above, the intervention of the human actor may be error-prone and time-consuming.

Our literature analysis identified several techniques used to propose curation contributions, such as machine learning, semantic technologies, graph-based techniques, crowdsourcing, and rule-based techniques. We noticed that the adoption of machine learning techniques is still limited, especially for data cleaning and schema mapping. Indeed, several curation tasks such as deduplication, spotting errors, and violation and repairing data are hard to automate. Accordingly, most of the studied curation approaches rely on rule-based techniques, semantic techniques, or the incorporation of machine learning with one of the above techniques. Consequently, we identified rule-based contributions for curation tasks like detecting violations such as [33]. Otherwise, machine learning techniques are combined with other methods, such as crowdsourcing, to perform the curation task like the work presented in [27]. The latter proposes a semi-automatic approach that relies on automatic curation via curation services and manual annotations via crowdsourcing and experts' annotations. Each service of the proposed curation services [38] represents a curation step (e.g., Linking dataset with knowledge base) that may be employed to constitute a curation pipeline. These services rely on several techniques like rules definition, linking with dictionaries and ontologies, and machine learning. Yet, the proposed curation pipeline also requires human intervention via crowdsourcing, which may sometimes be error-prone and time-consuming like presented above. Nevertheless, the incorporation of machine learning with other techniques can be explained by the subjective aspect of some curation tasks, such as detecting errors, which cannot be identified using a series of rules and require human intervention.

We also investigated the flexibility of the examined approaches. Indeed, we emphasize that all (1) the studied approaches are static regarding the decision process features. Morever, (2) the works presented in [33] and [34] ensure a low level of adaptation by considering end-user needs. In contrast, [29] keeps performing the same curation tasks orchestration regardless of the user's requirements. In addition, to the best of our knowledge, (3) all the examined approaches consider only batch data sources.
Moreover, our study reveals that full autonomy in performing curation and using machine learning and rules generalization are still open curation issues.

Based on the limits of the approaches presented, a solution must be proposed to overcome them through adaptive data curation. Hence, we detail in the next chapter our proposed solution for data curation for batch and streaming data simultaneously, adaptively to decision context, user profile, constraints and preferences, and the type of data source to be treated.

In what follows, we present work related to the design of a recommendation system of measures for crisis management. Indeed, we aim at analyzing curated data to predict changes in situations that may cause a crisis and recommend useful measures to help manage it.

2.5 Crisis management approaches

In this section, we present our findings and limitations after investigating works related to crisis management, which encouraged us to tackle the presented scientific challenge. While elaborating on the state of the art and studying the existing works, we identified several definitions of the concepts of crisis and crisis management, their types, and their scopes. Hence, we identified several definitions like the ones proposed in [39] and [40]. Some authors considered the crisis an event that may have unknown causes, while others saw it as the result of a succession of previous events. However, most proposed works share the common fact that any crisis has severe consequences. Moreover, some authors use specific terms like disaster and catastrophe [41] that may be more related to specific contexts. Following our analysis, we identified that natural disasters and other types of crises, like health crises and economic recessions, are considered macro-level crises and represent the main focus of most existing works. Nevertheless, there are also micro-level crises that include service failures in an organization, for example. Economic crises are the most investigated in the literature, followed by health crises (e.g., epidemics and pandemics). Several works proposed definitions for crisis management planning in the different stages (pre-crisis, mid-crisis, and post-crisis) involved in that process. At each level, crisis management may be changing in terms of needs and requirements. For example, crisis management in the early stages focuses on identifying strategy schedules and their influencing factors, as well as predicting a future crisis. During a crisis, crisis management covers tasks related to identifying and comparing response strategies and factors involved in this process. Also, in post-crisis management, the proposed work analyzes the long-term impact of the crisis and proposes and implements response and recovery strategies. Nevertheless, despite the variety of works proposed for each level, they still lack proper consideration of the different levels of analysis (e.g., national and international) and the incorporation of the requirements of the different actors involved in such a process. For example, the work in [42] highlights the need to take into consideration the expectations, preferences, and satisfactions of stakeholders that may change during the different stages of crisis management (i.e., pre, mid, and post-crisis). In addition, it highlights the need to address the characteristics and differences associated with multi-national crisis management. This work and others like [43] have highlighted the need for building a system's

capacity to be cognitive and adaptable to different needs and changes in crisis management.

2.6 Health outbreak prediction

Crisis prediction is helpful for governments and organizations to prepare their response plan to potential health crises effectively. For example, in the healthcare field, early health disease detection and response preparedness can help save lives, prevent the spread of diseases, and minimize the damage caused by them. In this context and by analyzing the works proposed in the literature, we identified two main approaches for outbreak prediction. The first approach monitors data from different sources like sensors, social networks, and health institutes to detect abnormal changes.

On the other hand, scientists identified a relationship between human and animal health [2]. Thus, the second approach monitors the regions characterized by the wildlife [44]. Indeed, zoonotic diseases or zoonoses may be caused by germs that spread between animals and people. For instance, Salmonella is a disease that may be transmitted from animals like reptiles, amphibians, chicks, and ducklings. Hence, scientists took advantage of artificial intelligence for disease prediction. Indeed, many methodologies are proposed for this purpose, such as risk mapping, regression models, machine learning, and incidence modeling. Accordingly, the proposed works analyze cases, patient health, and meteorological data to predict future cases, outbreak risk factors, outbreak risk, and epidemic dynamics [45].

We present in the following section basic concepts of disease prediction and review some existing relevant works.

2.6.1 Health diseases and outbreak prediction and management stages

The emergence of new infectious diseases such as HIV/AIDS, Severe Acute Respiratory Syndrome (SARS), or pandemic influenza often seems unpredictable. More than 60% of the approximately 400 emerging infectious diseases identified since 1940 are zoonotic, and these pathogens are of particular public health interest. Similarly, specific geographic regions or interfaces between people, wildlife, livestock, and the environment have been identified as the source of recently emerging infectious diseases and are, therefore, targets for intense surveillance. These advances, coupled with a better understanding of the dynamics of transmission, ecology, and evolution of pathogens as they emerge and spread, hold the promise of predicting pandemics. For this purpose, there are three main stages in managing epidemics/pandemics, as depicted in Figure 2.8 [1].

Stage 1 is a pre-emergence, in which naturally occurring pathogens are transmitted between their animal host. Indeed, disruptions in the ecology of the animals (due, for example, to changes in land use) alter the dynamics of microbial transmission and may result in

[2] https://www.cdc.gov/onehealth/basics/history/index.html

an increased risk of pathogen spread to other non-human wildlife or livestock (but not to humans).

Stage 2 is the localized emergence through self-limiting spread events (i.e., green peaks and troughs, representing increases and decreases in the number of infected persons over time) or large-scale spread (i.e., red peaks, representing peaks in the number of infected persons over time), which results in person-to-person transmission for a few generations of pathogens.

In stage 3, certain spillover events can lead to indefinitely sustained person-to-person epidemics, international or global spread, and the emergence of a full-blown pandemic. By dissecting this process and analyzing the interactions between the underlying factors and the spread risk, a more structured approach to pandemic prevention can be developed. Un-

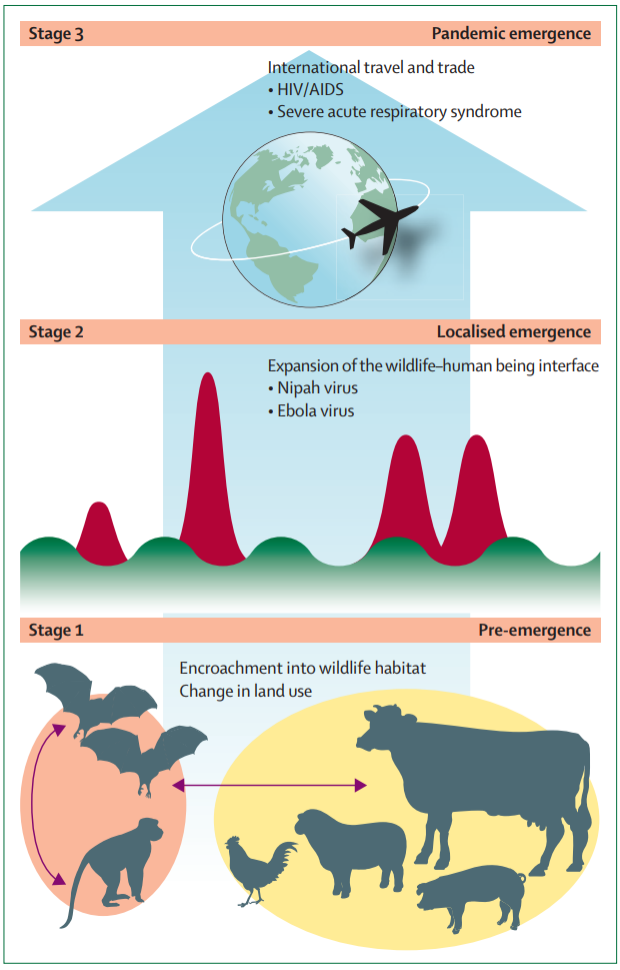


Figure 2.8: Emergence of pandemic zoonotic disease [1]

fortunately, it is tough to predict a pandemic at stage 1. Hence, our objectives focus on preventing pandemics in stage 2 and managing risks in stage 3. The prediction is made

through the analysis and study of "weak signals" a weak signal can be defined as one of the first indicators of a change or an emerging problem, which could take large proportions in the future. Collecting and analyzing these signals and combining them with a prediction system would provide a comprehensive picture of the future impact of a given infectious disease. Several prediction systems, which will be presented in the next section, are part of this prevention effort and use new technologies for data collection, processing, analysis, and diagnosis. Hence, we analyze and categorize the proposed contributions to identify and compare the different prediction systems and to identify the strengths and weaknesses that we present in the next section.

2.6.2 Review of existing disease prediction approaches

Today, the speed and extent of human exchanges and immigration has turned the struggle over infectious diseases into a global issue, as no state can think of hiding behind its borders. Indeed, health crises and emergencies considerably impact the economic, social, and geopolitical levels. Thus, disease prediction and health emergency and crisis management play a significant role in strategic planning. Due to the complexity of crisis management, a crisis management system ensures the organization of health monitoring and safety. With the emergence of the COVID-19 outbreak, several methods, platforms, and approaches were proposed to predict and manage this disease. Table 2.3 depicts the examined works for health disease prediction. We identified two categories of contributions, namely sensor-based and data-based approaches. Considering sensor-based approaches, many of them have employed sensors and IoTs to measure different parameters. Indeed, Mir et al. [46] proposed a framework that mines health parameters (e.g., fever, shortness of breath, cough, fatigue, travel history, oxygen, etc.) collected from sensors and IoTs in real-time and computes the presence of the COVID-19 virus. Following identifying a suspected case, the proposed framework shares data with healthcare centers and professionals and sends patients for tests and consultations. For this purpose, the framework encompasses four main components: user system or data collection, data analysis, diagnostic, and cloud system.
Mohammadi et al. [47] proposed a disease diagnosis system that employs IoTs to collect a patient's courtesy signals. Subsequently, a cloud or local processor stores and processes the collected data. Hence, the system diagnosis to make rational decisions about personal health. If the diagnosis identifies an emergency, a warning will be issued to the nearby hospital for medical emergencies.
Al Hossain et al. [48] proposed FluSense, a syndromic surveillance platform that monitors bio-clinical signals from hospital waiting areas to predict influenza-like illness. To do so, FluSense employs a microphone and a thermal camera to capture patients' behavior (e.g., cough, speech activities, etc.) in waiting rooms.
On the other hand, data-based approaches employ data collected from hospitals, patients, social media, etc. Indeed, Zhang et al. [49] for coronavirus sentiment analysis prediction using two components: an offline sentiment analysis and an online prediction pipeline. The offline sentiment analysis model trains and tests the machine learning models (i.e., Decision Tree, Support Vector Machine, Random Forest, Logistic Regression, and K-Nearest

Neighbors) to find the optimal model used in the online sentiment prediction pipeline. The latter collects streaming tweets and feeds them into the ML model to predict the sentiment analysis of the tweets in real-time.
Nikparvar [50] proposed a multi-variate Long Short-term Memory (LSTM) model that is trained using the mobility on multiple time series (e.g., data collected from Google mobility reports). The proposed model predicts future COVID-19 infection cases, deaths, and daily foot traffic patterns.

Table 2.3: **Overview of the examined works for health diseases prediction**

Paper	Category	Technology	Data	Outcome
Mir et al., 2022 [46]	Sensor-based	IoT-based sensors, linear regression, multilayer perceptron, autoregression	IoT-based data (temperature, audio, heart rate, oxygen, etc.)	COVID-19 Forecasting/ Send patient to health center
Mohammadi et al., 2020 [47]	Sensor-based	Wearable and implementable sensors, Neural networks	IoT-based data	Decision about patient health
Zhang et al., 2020 [49]	Data-based	Decision-Tree, Support Vector Machine, Random Forest, Logistic regression, K-Nearest neighbor	Tweet data about coronavirus	Sentiment analysis
Al Hossain et al., 2020 [48]	Sensor-based	IoT-based sensors, Neural networks	Patients data (coughs, thermal images, etc.)	Influenza
Nikparvar et al., 2021 [50]	Data-based	LSTM (Recurrent Neural Networks)	Mobility reports (e.g., data foot traffic, etc.), health data (e.g., COVID-19 confirmed cases, deaths, etc.), population data (e.g., population density, etc.)	Future confirmed cases, deaths, and daily foot traffic

Following our analysis, we noticed the numerous disease prediction systems devoted to dealing with the COVID-19 outbreak due to the recent emergence of the SARS-CoV-2 virus. Yet, we identified approaches devoted to other diseases, such as influenza, that may cause health outbreaks. The examined approaches generate different outcomes, such as sentiment analysis, future confirmed cases, deaths, patient health decisions, etc. Yet, most studied proposals generate only one type of recommendation for a single user role. Moreover, they lack diversity in developing recommendations and treat all the risks similarly. Yet, some characteristics may influence the recommendation process. For instance, the Health strategies depend on the country's characteristics, which make them differ from one country to another. For example, since the discovery of the first case of COVID-19 infection, China adopted strict health measures to cope with the virus, such as the lockdown of Wuhan city on January 23, 2020, and the "Four Earlys" measures (i.e., early detection, early reporting, early isolation, and early treatment) on February 2, 2020. Alternatively, South Korea detected the first COVID-19 case on January 27, 2020. Accordingly, South Korea adopted less strict measures such as border control, screening, and testing before imposing strict blockades in Daegu city and North Gyeongsang province on February 25, 2020. Unlike China and South Korea, which implemented a containment strategy, Japan adopted a mitigation strategy through different stages to reduce the spread of virus transmission. However, both health strategies have shown different levels of effectiveness. Indeed, health strategy effectiveness may depend on the country's characteristics (i.e., population, Human Development Index), as well as several factors, including the situation in the country, and the outbreak severity [51]. Moreover, health strategies need to evolve by considering the changeable factors and the gained experience during the contest with the virus. As there is no size fits all recommendation, a crisis management system needs to adapt to these changeable factors and propose health measures convenient for each situation. Following our analysis, we also noticed that the examined works do not propose explanations for their outcomes. Yet, as presented above, domain experts may seek reasons to attribute their confidence and trustworthiness to the model and understand the reasoning behind the recommendation [52]. Hence, we overcome the highlighted approaches' limits via our contribution, which is a semantic explainable health recommendation model. Our aim is to consider the country's characteristics and users' preferences to propose different health measures adapted for other countries and multi-user roles. Accordingly, we present in the next sections the basic concepts related to explainable artificial intelligence and recommendation systems, and then, the examined works for explaining AI models using eXplainable Artificial Intelligence (XAI) techniques.

2.7 Recommendation systems

Recommendation systems have been widely studied in various domains, such as the web, e-commerce, and many others. We cite very effective and well-known systems such as Netflix for movie recommendations or Amazon for books recommendation. Recommendation systems generally rely on three main approaches that we detail in what follows.

2.7.1 Content-based filtering

Content-based recommendation determines which items in the catalog best match user preferences [53]. Such an approach does not require a large user community or history of system usage. Indeed, each element of the catalog is described explicitly via a list of characteristics (i.e., attributes). As well, each user possesses a profile encompassing their interests. Accordingly, a similarity distance is measured between the item and the user's interests to find a match. Concretely, we distinguish several techniques to match user interests and items to recommend, such as similarity metrics (e.g., Dice[54], Cosine [55], Jaccard[56], etc.). Content-based recommendation systems are a practical solution since they consider individual user preferences and propose customized recommendations for each user. Indeed, this approach does not need the interests of other users to perform recommendations.
Moreover, it allows for the recommendation of new items or even items that are not popular (i.e., having less interest from most users). Yet, various elements (e.g., images and videos) could not be described by keywords, which makes content-based recommendations less effective in this case. In addition, the content-based recommendation could be confused with elements described with the same keywords. The content-based recommendation also does not support user interest evolution. This recommendation approach requires that all user interests and element descriptions be declared in advance, which is only applicable in some contexts.

2.7.2 Collaborative filtering

Somewhat of analyzing the relation between users' preferences and items for the recommendation, collaborative filtering-based systems measure the similarity between users' preferences to propose similar items. Indeed, this approach captures the users' preferences and clusters them into subgroups according to their preferences. Similarity measuring in collaborative filtering relies on two major approaches: Item-to-Item and User-to-User recommendation. The former measures the similarity between the items to recommend, making it suitable when dealing with many users and items.
On the contrary, the User-to-User approach measures the similarity between users of the recommendation system. Item-to-Item approaches employ several techniques, such as the cosine measure, to measure the similarity between items. In contrast, User-to-User approaches rely on other methods like the Pearson correlation coefficient. We identified other techniques used to measure similarities, such as matrix factorization, which is widely used. The idea of using similar other users' scores to measure the interest of the concerned user in an item has shown its efficiency. Yet, the cold start problem overwhelms the collaborative filtering approach. We distinguish the cases of the addition of a new user whose preferences are not known and the addition of a new item. Moreover, this approach may be costly regarding resources when dealing with numerous items and users.

2.7.3 Hybrid filtering

As depicted in Figure 2.9, a hybrid recommendation system employs different information, such as external knowledge and item characteristics, to recommend items for users. Specifically, a hybrid recommendation combines different recommendation approaches (e.g., content-based filtering, collaborative filtering, etc.) by taking advantage of the approaches while limiting their disadvantages. For this purpose, several combination approaches were proposed, such as monolithic hybridization design, parallelized hybridization design, and pipelined hybridization design [53]. Monotholic hybridization uses additional input data from other recommendation algorithms. For instance, a content-based recommendation system that also leverages community data to determine similarities between items. On the other hand, parallelized hybridization design combines the results generated by recommendation systems functioning in a parallel way, while pipelined hybridization combines them sequentially.

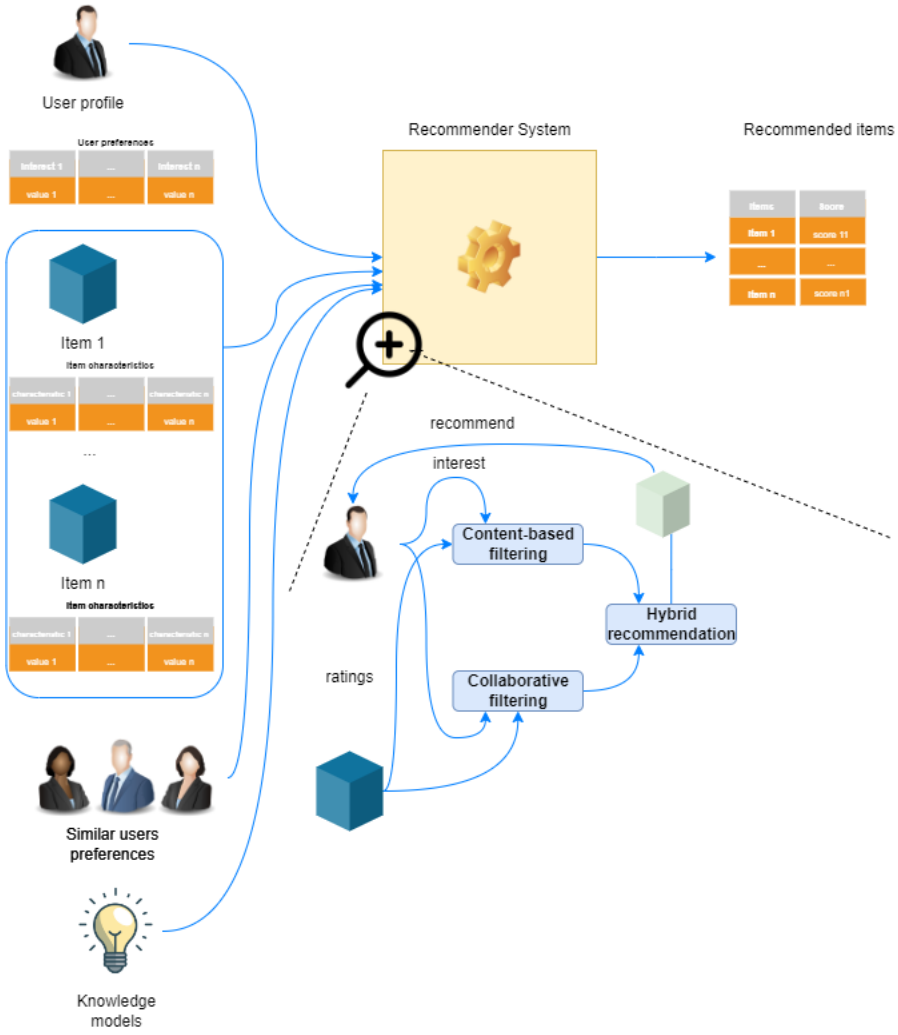


Figure 2.9: Hybrid filtering

2.7.4 Explanation of recommendations

The explanation of recommendations is one of the rights that the social community keeps insisting on. The latter needs to understand explanations for ethical reasons, the establishment of responsibilities, the understanding of the model semantics, and economic reasons [57]. Moreover, domain experts may seek explanations to attribute their confidence and trustworthiness to the model and understand the reasoning behind the recommendation since data in a chaotic situation, such as a crisis, may be poorly managed and biased [52] [58]. To the best of our knowledge, although some existing recommendation models provide explanations for stakeholders, they do not consider different user roles. Hence, we deem it challenging to adaptively explain recommendations for different user roles using different explanation types such as examples, counter-examples, feature importance, etc., as we detail in the following sections.

2.8 Explainable Artificial Intelligence approaches

Explainable artificial intelligence (XAI) has recently received much interest from researchers. Early artificial intelligence systems, such as expert and rule-based systems, were easy to interpret. With the advent of machine learning, AI systems are becoming more and more opaque. For instance, traditional classification rules provide a higher level of interpretability than decision trees. Figure 2.10 depicts the interpretability and accuracy of common artificial intelligence models. As shown in this Figure, the better the AI model performs, the less it becomes interpretable. Despite the performance and accuracy of deep learning models, these models are complicated to interpret.
For this reason, deep learning models are qualified as black box models. This is why the efforts of researchers have been very much turned to the interpretation and explanation of the results generated by these models. In this section, we present an overview of the explanation of artificial intelligence paradigm from different points, such as the terminology, the scope, and the principles of XAI.

2.8.1 Explanation approaches

We identified two approaches that interpret and explain the black-box models: the first consists of explaining the inner functioning of the model's layers and interpreting the intermediary processes. However, given the complexity of these models, this approach is limited and requires a trade-off between the interpretability and accuracy of the model. The second approach, "Post-hoc explanation" does not impact the performance of the deep model and ensures the explanation of the results generated by analyzing inputs and outputs without affecting the model performance. Post-hoc explainability targets models that are not easily interpretable by design. For this purpose, it encompasses various means to improve black-box models interpretability, such as textual explanations, visual explanations, local explanations, explanations by examples, explanations by simplification, and feature relevance explanation techniques. Each of these techniques covers one of the most common ways humans explain systems and processes.

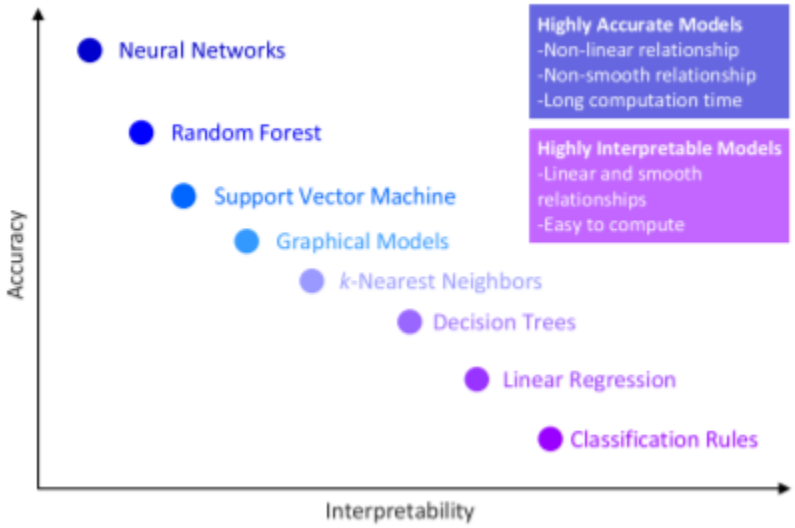


Figure 2.10: Accuracy and interpretability of AI models. [2]

2.8.2 Terminology

Considering our work, we adopt the terminology proposed in [2]

- Understandability or intelligibility is the characteristic of a model to make a human understand its function without explaining its internal function.
- Comprehensibility denotes the ability of a learning algorithm to represent its learned knowledge in a human-understandable fashion.
- Explainability refers to the ability to provide explanation or meaning in understandable terms to a human.
- Transparency refers to the degree of transparency of the model by itself to be understandable.

2.8.3 Explanation scopes

We present the scope of explanations, one of the widely used criteria to classify XAI systems. Indeed, we identify two main scopes of explanations: local and global.

- **Global explanations** illustrate the systems' operations or reasoning adopted to generate the outputs. For instance, a set of rules that imitate the internal functioning of an AI system may be considered a global explanation.
- **Local explanations** try to explain individual predictions generated by the system rather than the whole AI model functioning. As an example of local explanation, we present illustrating examples of outcomes generated in similar cases as the proposed outcome.

2.8.4 The four principles of XAI

We present below the four fundamental principles of XAI defined by NIST[59] that need to be considered when designing an explainable system.

- **Explanation:** the following principle defines that an AI system needs to supply evidence, support, and reasoning for each decision.
- **Meaningful:** this principle states that explanations provided by the AI system need to be understandable and meaningful to a user or a group of users.
- **Accuracy:** the following principle insists upon the accuracy of explanations provided by the AI system.
- **Knowledge Limits:** this principle reflects that an AI system must have knowledge limits regarding explanations.

2.9 Semantic-based AI explanation types and techniques

We rely on semantic technologies that have long proven their efficiency and performance in illustrating knowledge in several fields. However, their adoption is still limited in designing explainable artificial intelligence systems. Hence, we present hereafter the explanation types and techniques related to XAI and review existing semantic-based explanation works.

2.9.1 Explanation types

In the present work, we adopt the following classification of explanation types proposed in [60]:

- **Case-Based**: provides solutions based on actual past cases that can be presented to the user to convincingly support the system's outcome and can involve analogical reasoning based on similarities between the characteristics of the case and the current situation. *"To what other situations have this recommendation been applied?"*
- **Contextual:** refers to information about elements other than explicit inputs and outputs, such as information about the user, situation, and general environment that affected the calculation. *"What broader information about the current situation prompted the suggestion of this recommendation?"*
- **Contrastive:** answers the question *"Why this output rather than that one?"* by contrasting the given output with the facts that led to it.
- **Conterfactual:** addresses the question of what solutions would have been obtained with different data than the one used. *"What if input A was over 1000?"*

- **Everyday:** uses real-world stories that appeal to the user, given their understanding and general knowledge. "Why does option A make sense?"
- **Scientfic:** references the results of rigorous scientific methods, observations, and measurements. *"What studies have backed this recommendation?"*
- **Simulation-Based:** uses an imagined or implemented imitation of a system or process and the resulting results of similar inputs. *"What would happen if this recommendation is followed?"*
- **Statistical:** represents a report of the result based on data about the occurrence of events under specified (e.g., experimental) conditions. *"What percentage of people with this condition have recovered?"*
- **Trace-Based:** provides the underlying sequence of steps used by the system to arrive at a specific result. *"What steps were taken by the system to generate this recommendation?"*

2.9.2 Explainable Artificial Intelligence techniques

Throughout our analysis, we identified different techniques used for post-hoc explanation. Among them, we present the following ones:

- **Text explanations:** are generated to explain the outcomes of an AI model. These explanations may rely on means of semantic mapping from model to symbol.
- **Visual explanations:** employ visualization techniques to explain a model's outcomes for the user. For instance, heat maps that highlight a part of an image that leads to the generation of the outcome could be considered visual explanations.
- **Explanation by examples/counter examples:** proposes extracting data examples concerning the generated outcomes. Accordingly, it presents similar and different situations to illustrate the model's behavior. Humans basically use this type of explanation to explain a given process.
- **Explanation by simplification:** (i.e., surrogate model) reduces an AI model complexity by generating a more transparent model that imitates the behavior of the target model. For instance, a decision tree-based model could be extracted from a deep learning-based model since the former is more transparent than the latter. Hence, exploring the decision tree hierarchies explains the reasoning behind the deep learning model behavior.
- **Semantic mapping/matching:** is the least used technique compared to the abovementioned one. It relies on linking with a semantic component (i.e., knowledge graph, ontologies, etc.) through mapping or matching to explain the outcomes of a model semantically. Semantic explanations are still less used compared to visual and surrogate model explanations.

2.10 Measures recommendation and explanation

As our goal is to propose a system for recommending appropriate actions for crisis management, we have studied the proposed works for recommendation and the underlying works, such as eXplainable Artificial Intelligence (XAI), to overcome the poor interpretability of AI models. Thus, we have selected the works of XAI that take advantage of semantic technologies (such as ontologies) to add semantics to AI models. In what follows, we review and detail the underlying work.

2.10.1 Semantic-based explanation approaches: state of the art

In recent years, eXplainable Artificial Intelligence (XAI) has attracted the attention of many researchers. Indeed, there is a trade-off between the performance of an AI model and its degree of interpretability. For instance, neural networks outperform decision trees. However, they are less interpretable. In this context, several approaches and techniques were proposed to promote AI model interpretability, particularly deep learning models. Following a literature analysis, we identified two categories of XAI contributions. The first category encompasses works that aim at explaining the inner function of an AI model (e.g., deep learning models) by adding semantics for each layer, while the second category includes *post-hoc explanation* methods that explain AI models by extracting relationships between the inputs and the outputs of a model to learn the relation between them and generate explanations [61]. Despite the diversity of approaches proposed for post hoc explanation, there are few semantic approaches for explaining deep learning models. Nevertheless, semantic technologies such as ontologies and knowledge graphs provide reasoning with concepts and relationships in a way that is close to how humans perceive related concepts. Semantic technologies are more consistent and ensure easy navigation, scalability, flexibility, and interoperability. In this section, we present and examine some relevant semantic-based XAI works by examining the techniques used, the explainability scope, and the explanation types.

Table 2.4 depicts the examined works for semantic-based explainable artificial intelligence approaches.

The authors in [62] extend TREPAN, an existing algorithm for artificial neural network simplification to decision trees, by including ontologies for domain knowledge modeling in generating explanations. The Doctor XAI approach [63] proposes an explainability technique that deals with multi-labeled and ontology-linked data. The authors in this paper focus on Doctor AI explanation, which is a model for the next visit prediction using the patient clinical history.

The work [64] proposed an explanation approach based on DL Learner, a framework for learning concepts in Description Logic from user-provided examples. The ontologies are used as an intermediary to explain the input-output behavior of trained artificial neural networks.

We also examined [65], which is an approach proposed to cluster AI model features semantically using an ontology to assess the meaning of the values during the clustering

Table 2.4: Comparison of the examined semantic-based works for eXplainable Artificial Intelligence

Title	**Used technique**	**Scope**	**Explanation type**	**Target user**	**Aim**
TREPAN, 2020 [62]	Surrogate Model	Global	Contrastive/ Contextual	Static	Àn approach that explains TREPAN algorithm
Doctor XAI, 2020 [63]	Rules Extraction and features perturbation	Local	Counterfactual	Static	An approach that explain Doctor AI model
DL Learner, 2017 [64]	Matching	Global	Everyday	Static	An approach that explain annotated images
Semantic Clustering, 2020 [65]	Clustering	Local	Contextual	Static	An approach for features clustering
De sousa et al., 2021 [66]	Matching	Global	Trace-based	Static	Approach that extract explanation rules
Explanation Ontology, 2020 [60]	Matching	Local	Contrastive/ Everyday/ Contextual	Multi-user	Ontology that could be instantiated for the explanation of the AI models.

process.
De Sosa et al. [66] leveraged ontologies and small classifiers within the neural network model to link its internal states with ontology concepts. Thus, the authors aim to establish the mapping to understand the model's internal behavior.
We also examined the Explanation Ontology [60] that provides a structured representation of several explanation types and models the role of explanations, accounting for the system and user attributes in the process.

2.10.2 Analysis & discussion

After analyzing the presented works, we identified various techniques used in the examined works. Accordingly, our analysis reveals that matching with external ontologies is the

most used technique for AI model explanation. [64] performs matching with the framework DL Learner to learn input-output behavior, while [66] relies on matching with external ontology to extract explanation rules. As for [60], the explanation ontology could be used for matching to enrich semantically the mechanisms that generate explanations (e.g., surrogate models for simplification). On the other hand, we identified works that had used ontologies to derive a surrogate model, such as [62]. TREPAN Reloaded [62] extends the TREPAN algorithm using ontologies by creating more understandable decision trees. Specifically, ontology helps to determine which features are more understandable for a user, and assign priority in the tree generation process. As for Doctor XAI [63], this work employs an ontology to extract rules and perturbation of features semantically. As for explanation scope, the examined works cover different scopes, such as local [63], [65], [60], and global [62], [64], and [66]. Local explanations aim to explain a model instance, while global explanations illustrate how the AI model works and try to imitate it. Works of the both scopes had proved their effectiveness. As for the explanation types, most of the examined approaches propose a single explanation type. Indeed, explanation types range from counterfactual [63], everyday [64], contextual [65], [66].

In contrast, only TREPAN Reloaded [62] and explanation ontology [60] propose different explanation types. The former proposes contrastive and contextual explanations, while the latter combines contrastive, every day, and contextual explanations. Our analysis also covers the target users and reveals that most examined works provide explanations for a single user. We noticed that only explanation ontology considers several user roles for the explanation. However, the examined explanation ontology does not provide dynamic explanations since it predefines links between users and explanations in advance. Considering the limits of the presented approaches, it is essential to propose a solution that overcomes them by considering several users roles and providing different explanations dynamically and adaptively according to different user roles and needs.

Considering our context, we aim to recommend health measures by considering several contextual information (e.g., country, virus reproduction rate, etc.). Moreover, we consider user preferences in terms of explanations to propose a convenient explanation for him. Our contribution relies on two models for the recommendation. Indeed, we rely on deep learning techniques to design a content-based model for measures recommendation since it outperforms traditional recommendation techniques (e.g., vector spacing models) [67]. Moreover, we aim to design a semantic model that explains the recommendation model's outcome. As for the explanation approach, it relies on semantic-based techniques and collaborative filtering (i.e., Matrix Factorization). We adopt collaborative filtering since our contribution is dedicated to several user roles that may have similar interests in terms of explanations. Hence, identifying and exploring similar users' interests helps identify new user preferences. As our models generate several outcomes (i.e., measures and explanations) for the user, we combine both recommendation approaches to constitute a hybrid recommendation approach. Indeed, the results generated by the measures recommendation model are used as input for the semantic explanation model.
Accordingly, we overcome the weaknesses of the content-based and collaborative filtering

approaches by proposing such a hybrid contribution. Considering the sequential recommendation process, we adopt the pipelined hybridization to design our contribution for measure recommendation for crisis management and a semantic explanation approach.

Conclusion

This chapter presented the concepts and technologies underlying our research by presenting an overview of measures recommendation for crisis management and its relation to the different artificial intelligence technologies. This chapter also covers technical concepts involved in our contributions, like explainable artificial intelligence, multi-agent systems, and data lakehouses, which we adopt as a data repository.

After presenting and overviewing the basic concepts, we presented a detailed exploration and analysis of the state of the art around the sub-problems of our research work, such as curation, disease prediction, and measures recommendation. First, we illustrated a study about curation from different perspectives, namely data, metadata, and schema curation. We identified several tasks for data curation, such as contextualization, linking with external knowledge bases, data repair, schema, metadata extraction, etc. The examined contribution relied on several techniques, such as machine learning, semantic technologies (i.e., knowledge graphs, ontologies, etc.), graph-based techniques, crowdsourcing, etc. Moreover, they curate different data structures, namely structured, semi-structured, and unstructured data. Despite the diversity of the techniques and purposes of the examined approaches, we noticed that most of them consider only batch data and do not treat streaming data. They also show static behavior while performing data curation. Indeed, they need more flexibility to adapt their curation processes according to the decision context and user requirements. Hence, they do not provide dynamic organization and rearrangement of the curation process. We also identified that the generalization of curation tasks presents an issue, especially for the rules-based and machine-learning-based approaches. Our study also shows that some curation works lack full autonomy and require human intervention, which may be time-consuming and error-prone.

In the second step, the chapter studied and highlighted the different works related to measure recommendations. Nevertheless, most examined works for prediction do not propose monitoring and managing the situation (e.g., a health outbreak) via continuous recommendations of measures. Moreover, the examined approaches do not consider the different user preferences or roles to adapt and align their outcomes according to their needs. We also stated that these contributions do not propose explanations for the end user. However, they have high importance in critical fields such as healthcare. For this purpose, we examined the existing works on explainable artificial intelligence. Accordingly, we identified two main approaches: AI model interpretability and post-hoc explanation methods. The first approach imposes a trade-off between the AI model's accuracy and interpretability, which may affect its performance. On the other side, the post-hoc method does not affect it. Following the analysis of the existing post-hoc methods, we noticed that only some works

rely on semantic technologies to generate explanations. Nevertheless, the latter encompasses several techniques that may enhance the semantics of the explanations provided to the user. We also stated that the examined approaches do not provide more than one type of explanation, which may not be adapted for multi-user needs.

Our work aims to propose a system that perform explainable recommendation of measures to manage crises using curated data collected from different sources while considering different user preferences in various decision contexts. The above contributions are introduced and discussed in the following chapters.

CHAPTER 3
Adaptive data curation for batch and streaming data

CHAPTER 3

Adaptive data curation for batch and streaming data

> Errors using inadequate data are much less than those using no data at all.
>
> Charles Babbage

Contents

3.1 Introduction

This chapter presents our approach to adaptive batch and streaming data curation. Data quality, heterogeneity, and complexity remain critical challenges for big data. Indeed, data cleaning is necessary to enhance the quality of the data before analysis or visualization. For this purpose, data curation ensures managing and promoting data use from its point of creation by enriching or updating it to keep it fit for a specific purpose [21]. It provides more information about the provenance of the data, the original context of measurement and use, and the object of observation to facilitate the re-use of the data [22].

Nevertheless, the existing data curation approaches are no longer sufficient to curate multi-structured big data collected from multiple sources using different ingestion modes (i.e., batch and streaming data) [68]. Besides, factors such as the characteristics of the data source and the decision context may affect the data curation process. Indeed, critical decision contexts, such as crises, are generally evolving and imposing restrictions on the execution time and accuracy of information system outcomes. Therefore, it is paramount to consider the data characteristics and usage context to identify and perform the convenient data curation process [69]. Besides, the value of data is never settled, as its semantics are continuously changing, which forces the data curation process to be rearranged and changed over time [70]. Hence, the data curation needs to be aware of the changeable decision process features to optimize the quality of the decision process outcomes and its execution time and to align with user expectations. Hence, it is challenging to identify the convenient data curation tasks and rearrange them regarding the data source characteristics, the decision context, and user expectations.

As presented in the previous chapter, most existing data curation approaches are static and do not consider the abovementioned decision context features. The latter are handicapped decision-makers (i.e., those dealing with critical situations) who want to make decisions promptly and effectively. Besides, some existing approaches require human intervention, which can be time-consuming and error-prone. Nevertheless, the static aspect of the data curation step may handicap other steps in the decision-making process, such as data integration and analysis. For instance, stakeholders' needs may differ according to the decision context. We assume that Alice, a deputy senior defense and security officer at the Ministry of Health, and Bob, an infectious disease specialist, use a system to get recommendations for managing health crises. This system collects multi-structured data (i.e.,

structured, semi-structured, and unstructured) with different qualities that are collected in batch and streaming modes. Alice uses the system in a crisis context, while Bob uses it in an ordinary situation. Hence, they may have different outcome accuracy and system response time needs. Indeed, response time may be significant for Alice, while outcome accuracy in Bob's case is less critical. Thus, the data curation in Alice's case differs from that in Bob's case regarding decision context and user needs. Considering that such a system employs a data management process, the characteristics of the data source may directly impact the requirements regarding data curation. For instance, structured batch data sources may require a data curation pipeline different from streaming semi-structured data sources. Accordingly, data curation needs to consider the decision context and the user's functional and non-functional requirements that may impact the quality of the outcomes, like accuracy, response time, etc. Hence, our objective is to design a solution for adaptive data curation for multi-structured batch and streaming data while considering the abovementioned requirements. To do so, we consider semantic technologies as a practical solution by proposing an ontology to characterize and evaluate data quality. Ontologies can significantly impact data quality by providing a common framework for representing and defining data. Thus, ontologies can help to standardize and validate data inputs by ensuring that the data conforms to the expected format and type, which can also facilitate data reasoning and integration.

The present chapter is organized as follows: First, we overview in Section 3.2 the general idea of our proposed approach for adaptive curation of batch and streaming data. Specifically, we present in Section 3.3 the proposed DAta chaRacterization and Quality evAluation oNtology (DARQAN). Then, we detail in Section 3.4 our proposed approach, ACUSEC, for Adaptive CUration SErvice Composition. Subsequently, we present in Section 3.5 and Section 3.6 the implementation and the experiments elaborated to evaluate our proposal. Finally, we present Section 3.7, which concludes the chapter.

3.2 General Idea

We propose a service-based approach named ACUSEC for adaptive batch and streaming data curation according to the user's functional and non-functional requirements, such as his preferences and requirements, his decision context, and the quality of the curation services. As depicted in Figure 3.1, we perform data curation after evaluating data and source quality. For this purpose, we propose a modular ontology DARQAN that plays a double role in assessing the quality of data and data sources through different dimensions and extracting the data characteristics that guide the data curation process. Indeed, the characteristics of the data source may impact the selection of curation services, which, consequently, influences the composition of the overall services. Thus, we employ ACUSEC and DARQAN ontology to constitute a new data curation framework.

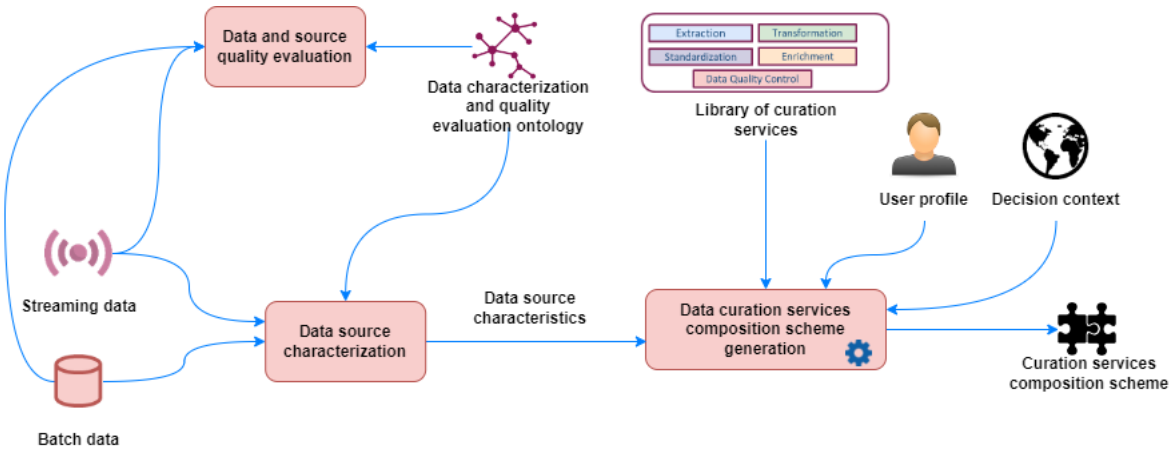


Figure 3.1: Overview on the adaptive curation service composition

3.3 DARQAN: DAta source chaRacterization and Quality evAluation oNtology

As presented previously, we aim to propose an ontology that plays a double role. First, it evaluates the quality of the data source to judge the need for data curation. Second, it guides the proposed ACUSEC approach in the curation services composition process by identifying the main data source characteristics that impact the composition. For this purpose, we adopt the AOM methodology [71] to propose, DARQAN, a modular ontology that describes data sources from different perspectives. As depicted in Figure 3.2, our ontology encompasses four modules: data source description, data quality, provenance, and platform modules. The following sections describe each module and how we adopted it in our context.

3.3.1 Data description module

The data source description module provides several kinds of information, such as information on the data itself, such as the period and location of observations, the linguistic system, the different forms of data it contains, and information about the provider. Figure 3.17 depicts the core classes of the data description module. This module also encompasses technical information in the data format in which the dataset is provided (Distribution) (e.g., an XML dataset, a plain text file, an SQL database, etc.). The data format may be combined with data properties like the URL to access a MySQL database, for example, the username and the password. Moreover, we have described the way of usage of the dataset, the tool that exploits it, the right statements, and the license to use it. As for data curation, it relies mainly on the information presented as data properties of the classes of the data source description module. We present the following features that characterize a data source:

- *- The data source format (structured (S), semi-structured (SS), or unstructured*

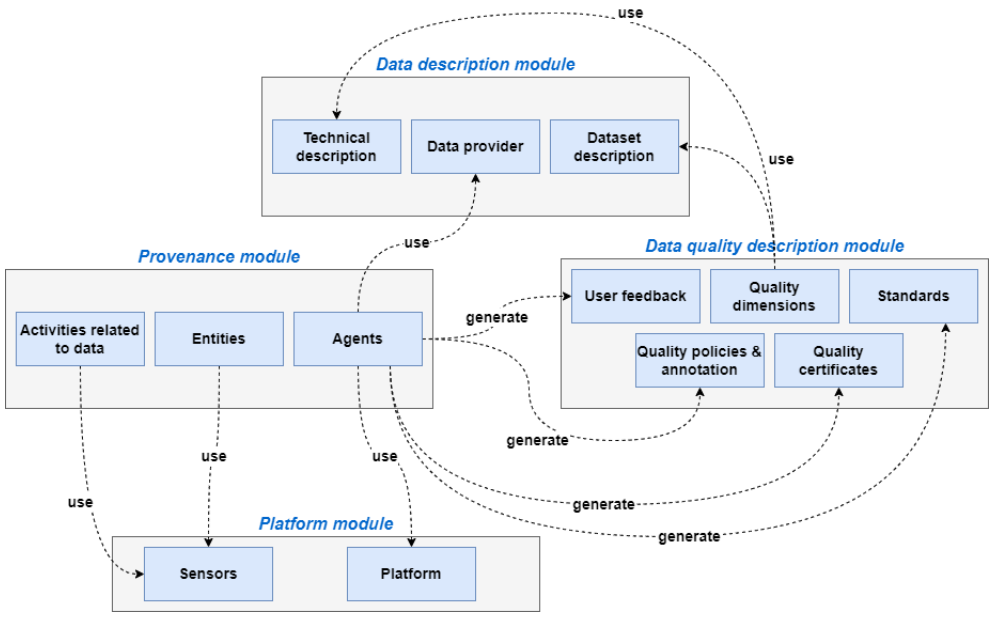


Figure 3.2: The data source characterization and data quality evaluation modules

(US)*: this feature guides the service composition to select suitable curation services according to the type of source.*

- ***Does the data source include an URL in its data values?*** *This feature helps determine whether an URL extraction service that extracts additional enrichment information from the URL's website should be invoked.*
- ***Does the data source need to be converted to another format?*** *Some data sources require format conversion before being curated. For example, a plain text file that presents an unstructured data source could be converted into a semi-structured data source (e.g., an XML file) to enhance the curation process. Using this feature, we can distinguish whether the data source needs conversion via the "Converter Service" invocation.*
- ***Does the data source need to undergo a PoS Tagging process?*** *Some data sources contain paragraphs that need to be annotated via POS Tagging to enrich them semantically. Indeed, POS Tagging is the association of words in a text with the corresponding grammatical information, such as part of speech, gender, number, etc. Hence, this feature allows for identifying whether the data source contains paragraphs that need to be annotated via the POS Tagging process.*
- ***Is it streaming data?*** *This feature allows distinguishing between batch and streaming data to invoke the convenient curation services for each data type.*

Following data characterization, we provide the formal description of a data source as follows:

$$\textbf{DS=<}DN, DAtt, Do, MAtt, DCh\textbf{>}$$

where:

- **DN** is the data source name
- **DAtt** represents the data attributes
- **Do** represents the data records
- **MAtt** is the set of attributes taken from a Metadata MD
- **DCh** represents the characteristics needed for adaptive data curation that were extracted from the data source via the data description module.

Metadata are defined as:

$$\textbf{MD=<}Mn, MAtt, MVal\textbf{ >}$$

where:

- **Mn** is the metadata name
- **MAtt** represents the metadata attributes
- **MVal** represents the data objects

In addition to the above data characteristics, we track the different actors who can perform the data generation to trace the origins of the data. Indeed, we trail the relationship between actors, data-related activities, and entities via the provenance module. As for entities, we design the platform module in our proposed modular ontology that explicitly describes sensors' generated data, like real-time data. In the following sections, we describe the ontology modules involved in the data characterization (i.e., platform and provenance modules) and quality evaluation steps (i.e., data quality module).

3.3.2 Data quality module

Data quality evaluation module evaluates the quality of data and data sources via reasoning based on several quality factors, such as quality dimensions, standards, certificates, quality policies, and user quality feedback. Indeed, we adopted and reused standards proposed by W3C, such as [72, 73, 74, 75, 76], to design a data quality evaluation module. By investigating the standards presented in [77] and considering the context of the present work, we relied on several data quality dimensions to evaluate the quality of the source and the data from different perspectives. Based on what we have identified in the literature, we mainly focus on evaluating data quality from the following perspectives:

- Data and Source accuracy: it is primordial to check the data's precision and source relevance.
- Time-related accuracy dimensions: In our context, the time aspect is crucial for checking the temporal validity of the data, which affects prediction reliability. As data lakehouses keep ingesting raw data sources, the data sources may continue to evolve, and some data may become obsolete. Thus, it is necessary to test whether the data are still temporally valid.

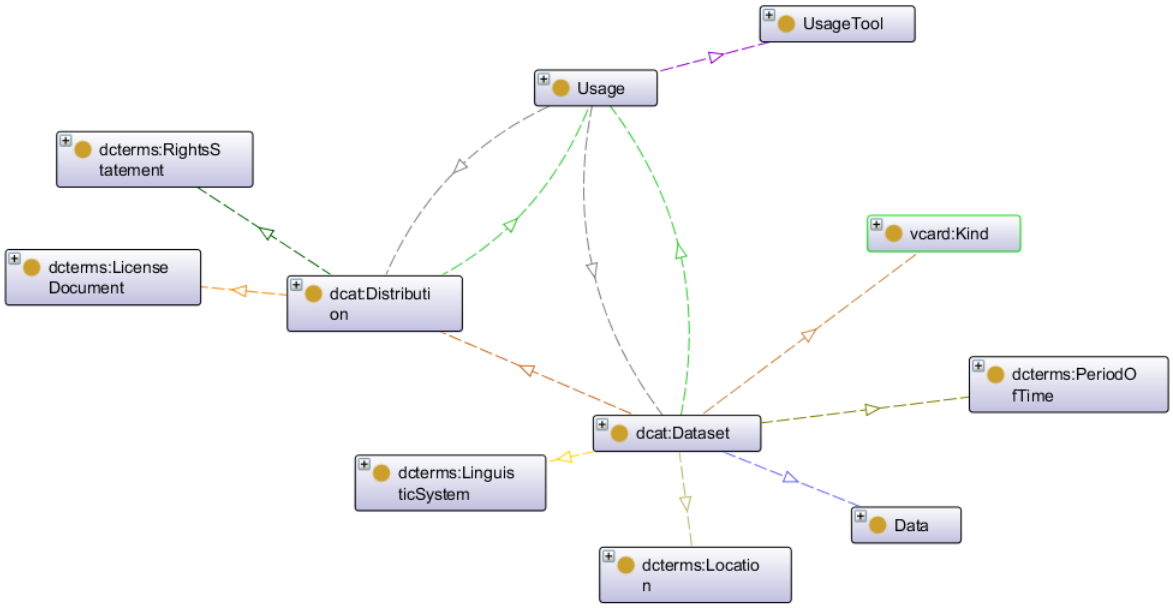


Figure 3.3: The core classes of the data description module

- Trustworthiness: to attribute confidence to a data source, we need to check its reputation as a data source as well as the reputation of its publisher. We can ask questions like "Can we believe its contents?", "Who is the publisher of this data source?" and "What is the users' review?"

Figure 3.3.2 depicts the core classes of the data quality module, and Table 3.1 depicts the data quality dimensions involved in our work.

Table 3.1: **Quality dimensions definitions**

Perspective	Dimension	Definition
Accuracy	Data Accuracy	Accuracy is defined as the closeness between a data value *v* and a data value *v0*, considered as the correct representation of the real-life phenomenon that the data value *v* aims to represent [78].
	Source Accuracy	Source accuracy is a ratio which is calculated between accurate values and the total number of values [78].
Time-related Accuracy Sub-Category	Currency	Currency concerns how promptly data are updated with respect to changes occurring in the real world [78].
	Volatility	Volatility describes the period for which information is valid in the real world [79].

Table 3.1 **– continued from previous page**

Perspective	Dimension	Definition
	Timeliness	Timeliness expresses how current the data are for the task at hand [78].
Trustworthiness	Believability	Believability is the extent to which data are accepted or regarded as authentic, genuine, and credible [78, 80, 81].
	Verifiability	Verifiability is the degree and ease with which the data can be checked for correctness [78, 82].
	Reputation	Reputation is a judgment made by a user to determine the integrity of a source [78].

Based on these data and the source's quality dimensions, the value of each quality dimension is calculated using SWRL rules that we define to perform reasoning over the proposed ontology.

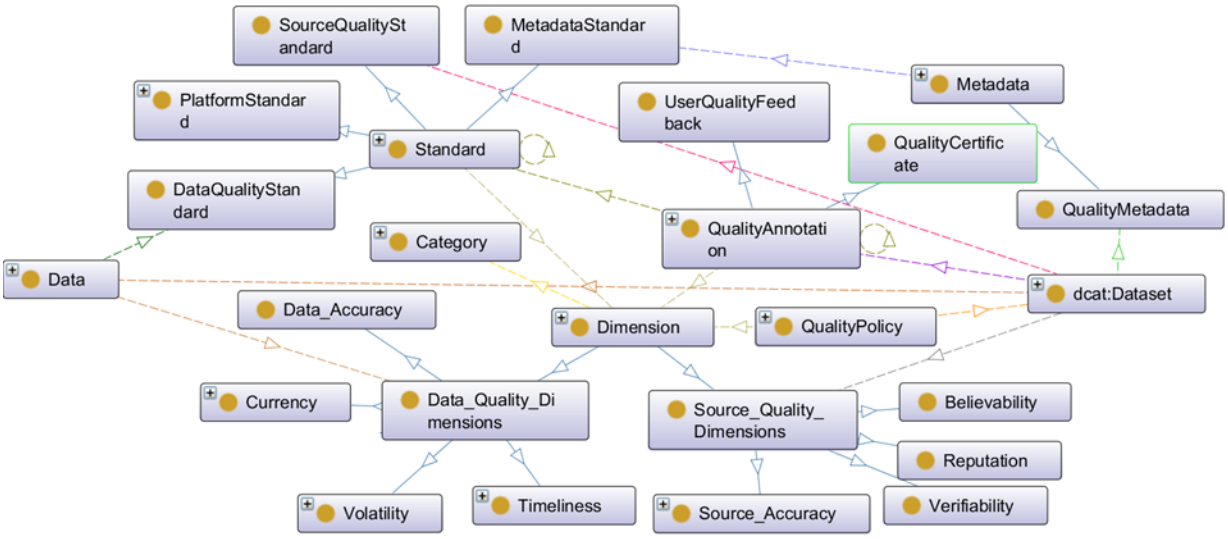


Figure 3.4: The core classes of data quality module

3.3.3 Provenance module

The provenance module is based on the logic proposed in [73] and tracks the origins of a data unit. As depicted in Figure 3.3.3, the three main concepts of this module are Entity, Activity, and Agent. The Agent class represents the agent who manipulates the activities (e.g., SoftwareAgent, Person, and Organization). An agent may act on behalf of another agent and use an entity or ensure an activity. The Activity class is designed to illustrate the activities leading to the generation of the data. These activities, in turn, could be associated with agents and use entities. Each activity may have start and end dates and may transmit information to other activities. As for the class entity, it describes entities dealing with data units (e.g., Sensor). An entity may be derived from another entity or generated from an

activity. For instance, we have represented the Collection class that groups entities (e.g., the Sensor Network collection groups the entities *Sensor*). To illustrate the utility of each class, we assume *agent* Bob is executing the *activity* data analysis, in which it checks and analyzes the data collected from the *entities* sensors.

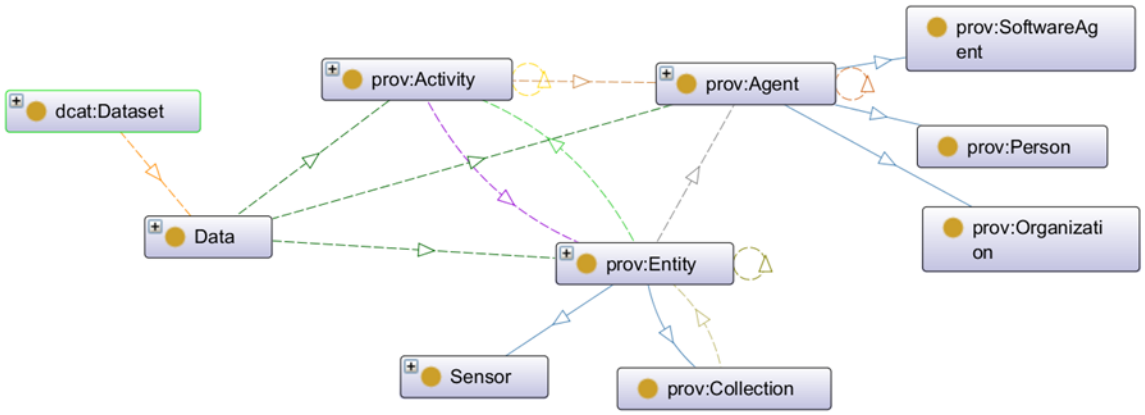


Figure 3.5: The core classes of the provenance module

3.3.4 Platform module

The platform module describes the platform from which the observations (e.g., heart rate) may be collected. Indeed, a platform, such as smartphones and satellites, may host entities like sensors. As depicted in Figure 3.6, every sensor is characterized by an observable property and a feature of interest. A feature of interest describes the thing whose property is being estimated or figured in the observation to get a result. On the other hand, the observable property class depicts an observable quality of a FeatureOfInterest. We give the following example to illustrate the role of each class. Assuming the heart rate of an individual is measured via a smartwatch. Hence, we can present this information via the class *Smartwatch*, which is a subclass of the class *Platform*. The *Smartwatch* class may encompass an individual named "Apple Watch Series 8" that contains a *Sensor* represented by the individual "photoplethysmography", which is the sensor used to measure heart rate in Apple watches. In the following section, we detail the proposed approach for adaptive curation service composition.

3.4 ACUSEC : Adaptive CUration SErvice Compostion

In this section, we present our approach to compose curation services using the identified characteristics. As we presented earlier, context awareness and user requirements should be considered during the entire decision process. Indeed, context awareness allows the involvement of contextual information in the system, which helps it adapt its outcomes according to the user's needs. Furthermore, user preferences play an important role in

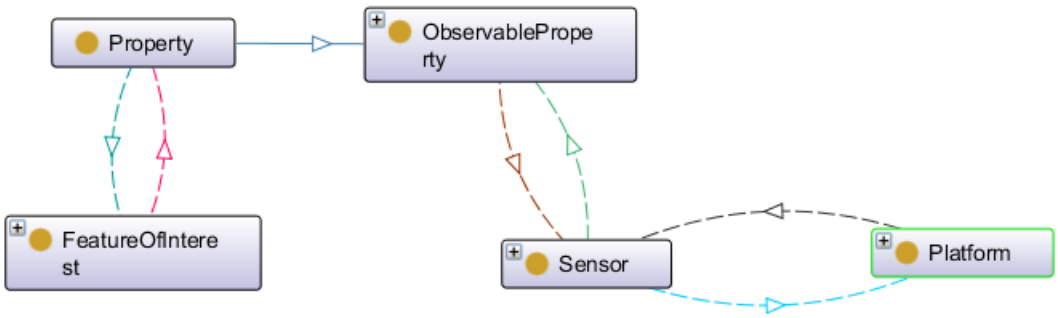


Figure 3.6: The core classes of the platform module

linking users with contexts to help them decide better and show behaviors more consistent with user expectations [83]. On the other hand, data curation encompasses many tasks, such as data enrichment, cleaning, metadata extraction, and schema extraction, to handle data, schema, and metadata management for later data analysis. Thus, data curation tasks need to be arranged in a specific order to perform successful data curation.
Moreover, each data format and ingestion mode implies the application of specific curation tasks. For instance, concept drift detection is applied for streaming data curation. Accordingly, we represent data curation tasks as services and treat the data curation as a service composition problem. Hence, our objective is to compose them according to the users' and the decision context's requirements. Thus, our challenge in this work is identifying convenient data curation services for multi-structured data collected in batch and streaming and their adaptive composition according to the data source characteristics, user preferences, constraints, and the decision context.

For this purpose, we designed a library of existing curation services that have proven their effectiveness. Then, we rely on the reinforcement learning paradigm to perform curation service composition. As depicted in Figure 3.7, our contribution, ACUSEC, is a two-stage approach. First, ACUSEC employs reinforcement learning to learn the possible curation service compositions according to the user's functional and non-functional requirements. Accordingly, the training stage identifies the optimal policy (i.e., set of composition schemes) to perform curation service composition. Following this stage, a composition stage composes curation services using the learned policy. In what follows, we describe the components that constitute our approach, such as the library of curation services, the learning and the composition processes.

3.4.1 Designing a library of curation services

The proposed Adaptive CUration SErvice Composition (ACUSEC) approach ensures adaptive, context-aware, and user-oriented data curation. Thus, it employs a library of curation services from which it selects and composes curation services. The curation service library encompasses existing curation services proposed in [38] and [84]. Figure 3.8 depicts the incorporated curation service library. The proposed library groups four categories of ser-

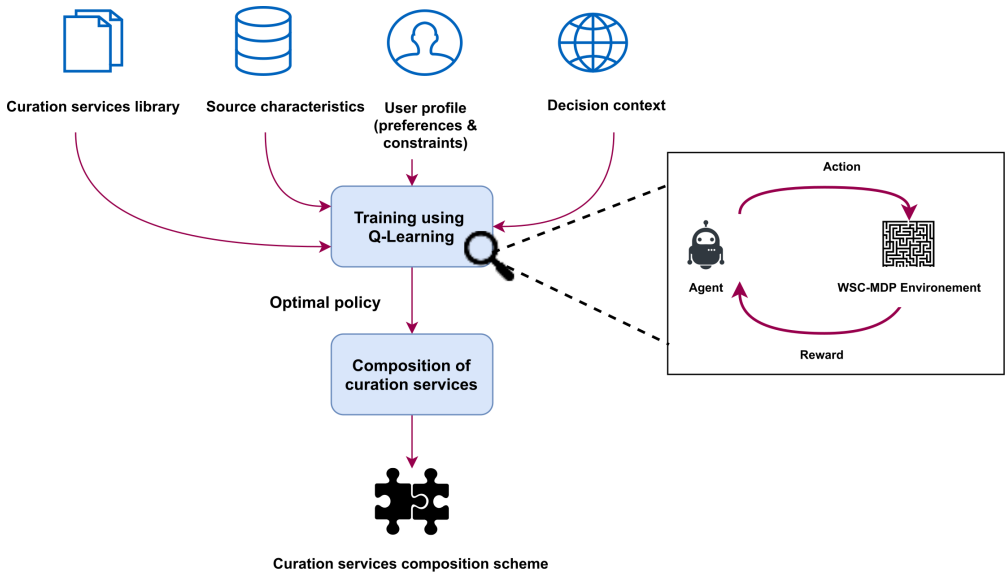


Figure 3.7: Learning process for adaptive service composition

vices: extraction, enrichment, data quality control, and data standardization services.

- **The extraction services category** encompasses services that ensure NLP (Natural Language Processing) tasks like named entity extraction, POS Tagging, and Stem extraction. Incorporating NLP tasks helps extract features that may be embedded in an enrichment process for later data analysis.
- **The enrichment services category** groups services for semantic data enrichment through knowledge-based annotation mechanisms and term similarity extraction. Enriching input data is convenient for data management as it can improve the generated outcomes [85].
- **Data quality control services category** contains two services that ensure missing values and data anomalies detection (e.g., value deviation). The services in this category perform data cleaning to improve data quality.
- **The data standardization services category** encompasses services that unify data using a knowledge base. This latter is used as a reference model, containing parameters describing the variables with their types and ranges. We state that data quality control and standardization services are suitable for curating structured data sources, even though most extraction and enrichment services could be used to curate data sources independently of the treated data source type.

We present hereafter the services grouped in each of the above categories.

Extraction services

The proposed library of curation services contains the following six extraction services:

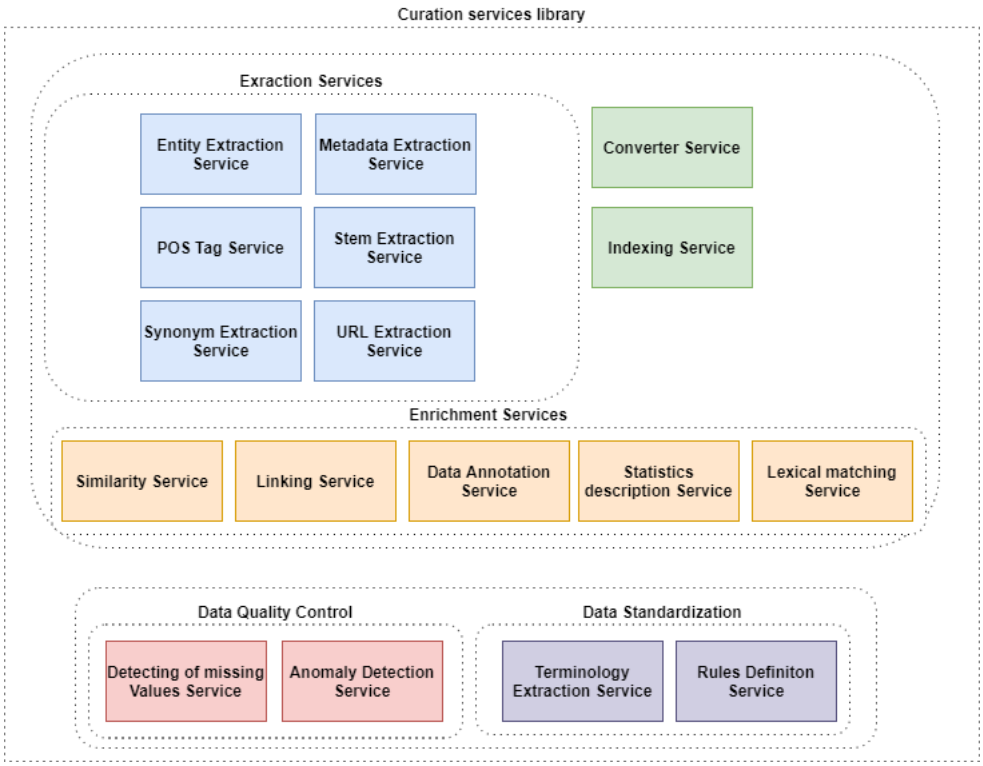


Figure 3.8: Curation service library

- **Entity extraction service:** entity extraction service classifies and locates atomic elements in a text into categories like the names of persons, organizations, and locations. Entity Extraction (EE) or Named Entity Recognition (NER) is a crucial step of data pre-processing that enables later information filtering and linking. For instance, we can identify three entities from the following sentence "Mark Zuckerberg is one of the founders of Facebook, a company from the United States", namely Mark Zuckerberg: Person, Facebook: Company, and the United States: Country.

- **POS Tag service:** a Part-of-Speech identifies the grammatical information of words such as nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, interjections, numerals, articles, and determiners. These annotations help analyze paragraphs and texts by examining the rhetorical relations between words. Figure 3.9 depicts an example of the PoS Tagging process.

- **Synonym extraction service:** synonym extraction identifies the words that have exactly or nearly the same meaning (e.g., end and finish). The semantics of words could change according to the context in which they become not synonymous. For instance, the words "long" and "extended" are synonymous if we talk about time. However, they do not have the same meaning when qualifying the word "work".

- **Metadata extraction service:** the metadata extraction services provide statistics about the dataset that may be needed later for data analysis, like the number of fea-

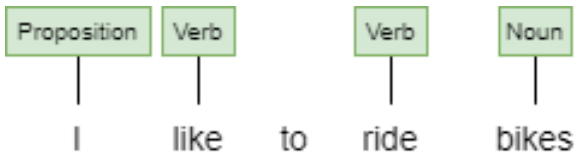


Figure 3.9: An example of PoS Tagging

tures, the size of the dataset, the value ranges, the number of continuous, categorical, and discrete features, and the number of missing values.

- **Stem extraction service:** the stem represents the basic form of the word in which affixes can be attached. For instance, the word "championships" includes the stem "champion" attached to the derivational suffix "-ship", which constitutes the stem "championship". The latter is attached with the inflectional "-s", constituting "championships". This task is crucial to natural language processing (NLP), allowing us to analyze textual data better. For instance, an analyst may seek different forms of the word "health" to identify tweets related to healthcare.
- **URL extraction service:** a Uniform Resource Locator (URL) is identified by an address that references a web resource (e.g., a web page). This curation service aims to locate URLs in texts to fetch additional information from web pages, including web page titles, paragraphs, sentences, keywords, phrases, and named entities.

Enrichment services

The enrichment category includes the following services:

- **Similarity service:** similarity approximates data matching using functions that measure similarity between different elements by assigning a score to a pair of data values. For instance, Jaro [86], Jaccard [56], and Cosine [55] measure the similarity between string data (i.e., character-based or token-based). Similarily, relative [87], and hamming distance [88] are devoted to measuring the similarity between numeric data. For instance, the Hamming distance between these two vectors ([0,1,0,1,1], [1,0,0,1,0]) is 0.4.
- **Linking service:** the linking service extracts information from entities existing in knowledge graphs like Google Knowledge Graph[1] and Wikidata[2]. To do so, we extract keywords from the text to enrich and then search for the classes related to the keyword to enrich by measuring the similarity between terms described in the similarity service. For instance, searching for the keyword "cancer" on wikidata returns various information related to this disease, such as descriptions in different languages, medical information, causes (e.g., smoking), and links to other external knowledge bases.

[1] https://developers.google.com/knowledge-graph
[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

- **Data Annotation service:** the data annotation service categorizes features as continuous or discrete, categorical or numeric, according to their data type and range values. Thus, it classifies data into predefined categories according to their quality. For instance, the authors in [84] proposed to classify data into three categories: good for datasets without missing data, fair for datasets having less than 25% missing data, moderate for datasets with less than 50% missing data, and bad otherwise.
- **Descriptive statistics service:** this service provides statistics about a dataset's features using the mean, median, minimum, maximum, standard deviation, kurtosis, and skewness. These statistics shed light on valuable information used to identify statistical differences between features, which help identify data anomalies.
- **Lexical matching:** lexical matching extracts structural information from the dataset to create a vocabulary containing the frequent terms and their ranges of values. Then, the identified metadata are used to extract additional information from a reference model (e.g., a domain ontology, an external knowledge graph) and link it with the dataset. For this purpose, the lexical matching service employs the Jaro distance to measure the distance between the terms to enrich and the fetched information.

Data quality control

We adopted the following services to constitute the data quality control category:

- **Missing values detection service:** this service is devoted to structured data source curation, which detects rows containing missing values. Other curation services, like the data annotation service, may require this information.
- **Anomaly detection service:** this service detects anomalies present in data by separating the core of regular observations from some polluting ones that present the outliers. Thus, this service measures the distance between a feature's value and the population mean. The latter could be acquired from other curation services, such as descriptive statistics and metadata extraction.

Data standardization

We also adopted the following services to constitute the data standardization category:

- **Terminology extraction service:** the following service extracts feature labels and adds constraints to remove incompatibilities, like parentheses and commas, to standardize the data representation of each feature. For instance, some information has different representations (e.g., dates could be presented in different formats like 13/12/2023, (13-12-2023), 2023-13-12, 23-13-12), which must be unified into a unique representation to better analyze data.
- **Rules definition service:** this service extracts domain knowledge rules from a reference model (e.g., domain ontology) to represent data. Thus, it relies on statistical

information like the ranges of each dataset's features. For instance, the "Gender" attribute could be equivalent to the "Sex" attribute and presented via different values, like 0 to denote male and 1 to denote female. We present other possible representations like "M/F" and "Male/Female". Hence, this service aims to identify such rules to standardize the data representation.

Other services

The proposed library also includes other curation services that do not figure in the above categories but are required for data curation, namely:

- **Converter service:** this service is applied when the conversion of data types is needed. For instance, some data types (e.g., text files) need to be converted to another one (e.g., XML), which is easier to curate.
- **Indexing service:** this service is based on *Apache Lucene* [3], a full-text search library that indexes data to facilitate searching. This service acquires a token or a phrase and returns the sentences that contain the given token. This service may help fetch datasets related to a specific context (e.g., healthcare, natural disaster), especially from unstructured data that contains large texts.

The presented services are involved in a reinforcement learning-based approach that we propose to compose curation services adaptively to the user's functional and non-functional requirements, as described in the following sections.

3.4.2 Reinforcement learning-based curation service composition

Since we are dealing with learning scenarios in dynamic environments, we rely on reinforcement learning, which has proven its efficiency in this kind of problem [89]. Specifically, we rely on the Q-Learning algorithm, one of the most popular algorithms for reinforcement learning, to learn the optimal curation service composition scheme adaptively. The Q-Learning algorithm is a model-free reinforcement algorithm that defines an agent interacting with an environment, usually defined as a Markov Decision Process, to learn the optimal actions to carry out a transition from one state to another. By adopting this logic and learning transition weights during the learning process, the learning agent can then assign them to transition actions. These weights represent rewards accumulated after each transition. Hence, we treat the composition of the curation services as a gain maximization problem that aims to maximize the overall reward. In the following section, we formally describe the environment designed as a Markov Decision process and used to perform reinforcement learning.

[3] https://lucene.apache.org/

The Markov Decision Process

A Markov decision process is a stochastic model where an agent makes decisions and the outcomes of his actions are random. As depicted in Figure 3.10, we represent the curation services in a Markov Decision Process (MDP) in which each transition action presents a curation service. Thus, in the MDP environment, we present all the valid possible compositions of all the curation services for all data source types, regardless of user requirements and environmental factors. During the training stage, the learning agent explores and exploits the environment to identify the optimal curation service composition according to the functional and non-functional requirements. The Markov Decision Process is denoted as[90]:

$$\textbf{WSC-MDP=<}S, s_0, S_r, A(.), P, R\textbf{>}$$

where:

- ***S*** is a finite set of states of the world
- $\mathbf{s_0} \in \mathbf{S}$ represents the initial state
- $\mathbf{S_r} \subset \mathbf{S}$ is the set of terminal states
- ***A(s)*** is the set of services that can be executed in state $s \in S$ to perform a transition from one state to another
- ***P*** represents the probability of transition from one state to another
- ***R*** that computes the reward following the transition from one state to another

The transition actions allow the transition from one state to another. Besides, the actions' weights represent rewards accumulated at each transition.

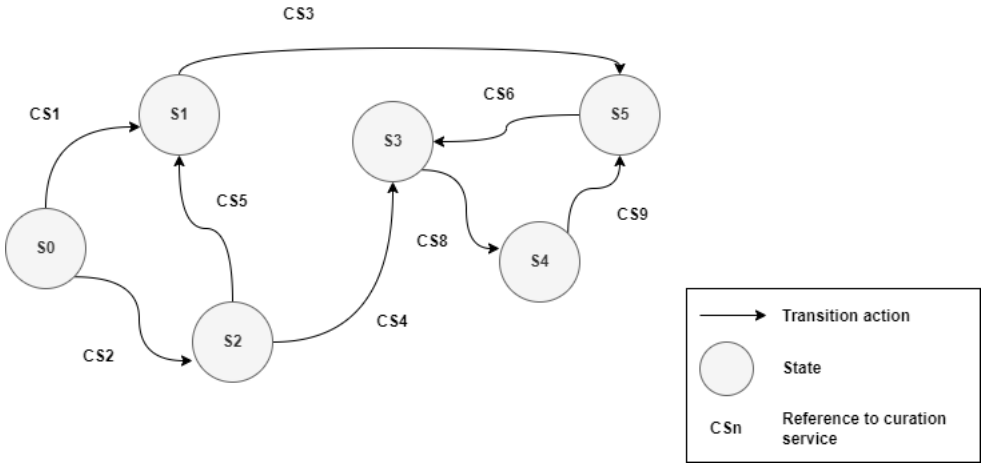


Figure 3.10: An example of a Markov Decision Process in which each action references a curation service

Training process based on reinforcement learning

We propose Equation 3.1 to compute the transition rewards. The proposed equation relies on curation services QoS, user preferences, and constraints (e.g., the QoS response time value >90%). The Quality of Service (QoS) is a measure to assess how well a service serves the end-user [91]. The training algorithm applies Equation 3.1 to compute the transition reward according to several QoS dimensions.

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

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

where:

- **w** represents user preferences regarding a QoS, defined as weight ranging from 0 to 1
- **D** is a normalized value of QoS dimension evaluation ranging from 0 to 1
- **M** represents a minimum threshold set by the user for QoS that needs to be fulfilled to invoke the service. The value of M ranges from 0 to 1.
- ϕ is a normalization value that needs to be strictly higher than 0 and lower than 1

We use user preferences as weights to promote one QoS dimension over another. Moreover, the training algorithm considers the defined constraints over QoS values by setting a minimum threshold M that should be satisfied to invoke a service. Considering the QoS, user preferences, and constraints, the reward function returns a positive value when all users' constraints are satisfied. Otherwise, the function returns a negative value. Since the service composition is a gain maximization problem, the negative rewards prevent the agent from choosing curation services that do not fit the user's constraints. The first part of equation 1 computes the difference between user-imposed constraints and QoS values. It returns either 1 if all user constraints are fulfilled or -1 otherwise. Equation 3.1 relies on equation 3.2 to compute the difference between one QoS dimension and the threshold M defined by the user. Subsequently, the value of equation 3.2 is normalized to -1 or 1 according to the obtained value. The second part allows assigning user preferences to QoS dimensions. Therefore, the preferences are defined as weights to multiply evaluated QoS dimensions values. Afterward, the multiplication of the two parts of the equation returns the reward value according to user preferences, constraints, and QoS values. In what follows, we formally describe the elements used to compute the reward function, such as the quality of services, user preferences, and curation services. Each curation service CS is characterized by an ID, a name, its quality (QoS), and an operation. We define a curation service as:

CS = <Id, CSN, QoS, Op >

where:

- ***Id*** represents the curation service Id
- ***CSN*** is the curation service name
- ***QoS*** is a set of evaluated QoS dimensions. It encompasses QoS dimension $\mathbf{QoS_D}$ and the evaluated QoS value $\mathbf{QoS_V}$ presented as pairs and assigned for each assessed QoS dimension
- ***Op*** is the operation name to be executed following a service invocation.

Illustration. As an example, we present the formalization of the URL extraction service that is identified by the Id 5 and characterized by the following QoS dimensions and values: "Availability: 70%, Accuracy: 80%, Reliability: 90%, Response Time: 70%, Reputation: 90%, Security: 70%". The formal description of this service is the following:
CS=<5,"URL Extraction Service", [("Availability", 0.7), ("Accuracy", 0.8), ("Reliability", 0.9), ("Response_Time", 0.7), ("Reputation", 0.9), ("Security", 0.7)],"Extract_URL()">
As for the user profile, it encompasses user preferences and user group preferences. Indeed, each user can be part of a group of users, and the group preferences may be aggregated from the users' preferences. During the training process, the user can use either his own preferences or his group preferences to promote one QoS dimension over another. Group preferences are useful in sharing information about specified preferences between users, which makes it possible to save effort and learn from other members. Thus, we define the user profile as:

U = <Np, Pru, G >

where:

- **Np** represents the user profile name
- **Pru** represents user preferences regarding a decision context **C**
- **G** represents a group of user profiles. A group is characterized by group name **Ng** and group preferences **Prg** concerning a decision context **C**

We define the decision context that represents the user's situation and surroundings as:

C = <Nc,Tc >

where:

- **Nc** represents the name of the context
- **Tc** is the decision context type (e.g., crisis, ordinary situation, etc.). We rely on the proposal in [92] to design the types and the characteristics of the decision context

Adaptive curation service composition process steps

We rely on the reward function described above during the training stage, which encompasses, in turn, three steps: environment initialization, exploration, and exploitation. As curation services are devoted to curating either semi-structured, unstructured, or structured batch data sources, as well as streaming data, the environment adapts itself by disabling some actions that are not convenient to the treated data source type. For this purpose, the environment initializes the reward returned by the disabled actions to a negative value. Thus, the agent will avoid these actions and select only the transition actions worth a positive reward value. Following environment initialization, the agent performs an exploration and exploitation process to identify the optimal service composition. Indeed, during the exploration and exploitation process, the learning agent uses a Q-Table, which stocks the probability of transition from one state to another. The transition probabilities will be used during the composition stage to identify the optimal curation service composition scheme encompassing the actions worth the maximum transition probability. During the training stage, the Q-Table values are updated progressively using a recursive function (Equation 3.3).

$$Q(s,a) = Q(s,a) + \alpha(r + maxQ(s',a')) \tag{3.3}$$

where:

- **s** represents the actual state
- **a** is the selected action
- α represents the learning rate
- $\mathbf{s}'$ represents the next state to select from actual state s
- $\mathbf{a}'$ is the next action to select to perform the transition from s to s'

This function is applied to calculate the probabilities of transition using the reward value (Equation 3.1), computed according to user preferences, constraints, and the decision context, as described above. At the end of the training process, we obtain the optimal policy $\pi*$ representing the final Q-Table. After the training and composition stages, the learning agent uses the optimal policy (i.e., the learned Q-Table) to retrieve the convenient curation service composition scheme by choosing the combination of actions that maximizes the overall gain. Algorithms 1 and 2 represent the used algorithms for training and generating the curation service composition scheme. After identifying the optimal curation service composition scheme, the data curation is performed by invoking the curation services in the composition scheme.

After the presented formal description, we illustrate, in the next section, the implementation and evaluation of our proposed approach.

Algorithm 1 Training algorithm

Require: User preferences and requirements, Decision context, Data source characteristics

Ensure: Q-Table that represents the optimal strategy learned during the process of training

```
Initialize Q-Table to 0
Initialize Reward Matrix R
Gamma ← Learning rate
Initialize the number of episodes E
for i = 0 to E do
    Current_state ← Choose random state
    Available_act ← Check available actions from current state
    if size(Available_act >0) then
        Act ← Choose random next action reward
    else
        Act ← 0
    end if
    Update(current_state,act,gamma)
end for
```

Algorithm 2 Composition algorithm

Require: Q-Table

Ensure: Curation service composition scheme

```
while there is available states to visit and the goal state is not reached do
    Choose the next state from Q Table which returns the maximum reward
    Append the next state to the list of visited states
    Eliminate the current state of the list of available states
    Current_state ← next_state
end while
```

3.5 ACUSEC : Implementation

As depicted in the following sections, we rely on our approach to constitute the data curation framework.

3.5.1 Adaptive framework for batch and streaming data curation

We rely on the presented service composition approach to design an adaptive data curation framework for batch and streaming data sources. We adopt the service-oriented architecture to design this framework since this architecture is reliable, scalable, and loosely coupled. Thus, we aim to optimize further data analysis steps in terms of execution time and alignment with user needs. As depicted in Figure 3.11, our framework encompasses the following four layers: data collection, data quality control, data treatment, and data curation layers. Indeed, the data collection layer ingests batch and streaming data sources and information about streaming data, data providers, location, and temporal information as metadata. Thus, our framework ensures adaptivity from the moment of data collection up to generating a curation pipeline. Subsequently, the framework evaluates the quality of the collected data via a data quality evaluation module and a data streaming monitoring module. For this purpose, the data quality module evaluates the data quality and the data source's quality. Hence, it considers quality dimensions, including data accuracy, timeliness, believability, verifiability, and reputation. The data curation framework judges whether the data source needs to be curated based on data quality. Data curation is performed when one of the evaluated data quality dimensions is below a threshold β, which the user can define. Following the data evaluation, the data quality dimensions and source values are transmitted to the data characterization module, which we define in the data treatment layer. The data source characterization module extracts the data source characteristics required for data curation, like the data source format, type, and specific data curation tasks.

Based on the extracted features, the user profile, and the decision context, the data curation layer selects the most convenient curation services from a library of curation services to constitute the data curation pipeline. As presented earlier, each curation service ensures a curation task (e.g., removing duplicate records, anomaly detection, etc.). These curation tasks could also be combined in a specific way to curate a data source. Our framework relies on our proposed approach ACUSEC to compose the curation services. As ACUSEC relies on machine learning techniques, specifically reinforcement learning, it aims at resolving the challenges (Cf. Introduction) that are related to data source heterogeneity, decision context instability, restriction in terms of execution time, and accuracy of outcomes. Indeed, machine learning algorithms can automate curation tasks organization and gain increasing experience as they improve accuracy and efficiency to make better decisions. Thus, the composition of the curation service is enhanced as the learning algorithms gain experience each time. In the following section, we prove the effectiveness of machine learning techniques for data curation.

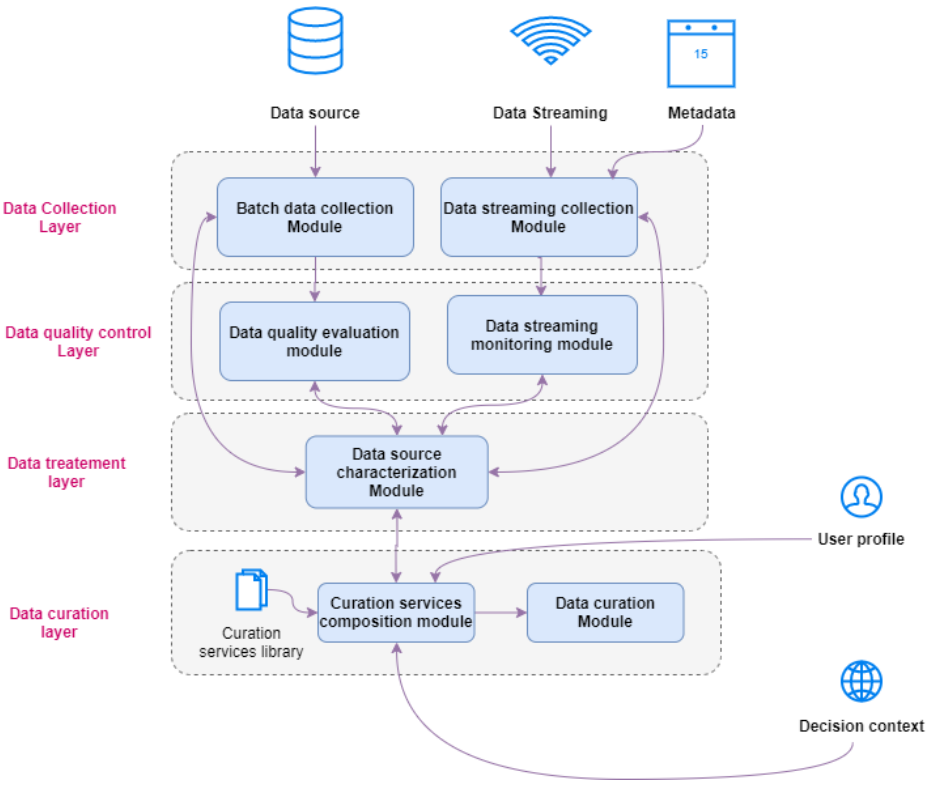


Figure 3.11: Adaptive data curation framework

3.5.2 Demonstration

We present the following example to illustrate the idea behind the proposed ACUSEC approach and the adaptive data curation framework. We assume that the adaptive data curation framework is implemented in a crisis management system that relies on different data management stages, including data curation. We suppose that Alice and Bob, whom we introduced in the introduction, use this system to predict and manage health crises. As presented above, response time may be significant for Alice, while outcome accuracy in Bob's case is less critical than response time. Moreover, we assume they use this system to analyze multi-structured data sources ingested in batch and streaming modes from different providers (e.g., the web, sensors, social networks, etc.). In the following example, we focus on data curation, which may be a component of such a crisis management system. We suppose that Alice wants to decide on a critical health crisis using various data sources, including sensors. Hence, the data curation framework collects data using the data streaming collection module while monitoring the data streams via the data streaming monitoring module. Then, the framework extracts data characteristics using the data source characterization module to identify different data characteristics, among which data are collected from streaming sources in JSON format. Since the curation service composition module deals with streaming semi-structured data, it initializes the MDP environment by disabling the transition actions referencing inconvenient curation services (e.g., the curation services for batch or structured data curation). Subsequently, during the exploration/exploitation

process, the learning agent learns the optimal composition service policy π^* using the equation 3.1. Then, using the learned policy π^*, the curation service composition module composes the curation services that fit Alice's needs by selecting services with a high response time QoS value. Later, the curation service composition scheme is transmitted to the data curation module to invoke the curation services and perform data curation. Considering another decision context, we assume that Alice uses the system in an ordinary situation to get some statistics from it. Hence, she has other preferences regarding the decision context. Accordingly, the curation service composition module adapts itself (i.e., reinitializes, re-explores, and re-exploits the MDP environment) and generates another scheme to meet Alice's needs. As for Bob, we assume that he is using the crisis management system to check the last recommendations to treat a new infectious disease. Thus, the system collects databases from diverse sources, like health institutions, to generate these recommendations. Hence, the curation service composition module initializes the MDP environment differently than in Alice's case by enabling curation services for batch and structured data curation since Bob is more interested in the accuracy of the results. The curation service composition module adjusts equation 3.1 weights during the exploration/exploitation process to promote accuracy over response time. Accordingly, it generates a different curation service composition scheme that meets Bob's needs.

The following section focuses on the evaluation of our proposed approach.

3.6 ACUSEC: Evaluation

We present in this section the experimental protocol that assesses the effectiveness of our proposal. Specifically, the elaborated experiments focus on the scalability of our approach regarding (1) the number of users, (2) the number of services, (3) the adaptivity and alignment with user requirements, and (4) the effectiveness of the data curation process. For this purpose, our experimental protocol relies on three different data sources, namely an unstructured dataset[4], a semi-structured dataset[5], and a structured dataset[6]. We compared the performance of our curation service composition method with the First-visit Monte Carlo and Temporal-difference Learning [93], two well-known reinforcement learning algorithms. Technically, the experiments were performed on an Intel Core i7-6500HQ PC with 16 GB of RAM using Python 3 and NumPy.

3.6.1 Data characterization and quality evaluation ontology

We conducted experiments to assess data quality through our proposed ontology for data characterization and quality evaluation. For this purpose, we rely on the Google Mobil-

[4]https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter: A dataset that contains health news from more than 15 major health news agencies such as the BBC.

[5]https://www.ncbi.nlm.nih.gov/datasets/coronavirus/genomes/: A dataset provided by the National Center for Biotechnology Information (NCBI) that contains data about COVID-19 genomes

[6]https://www.google.com/covid19/mobility/: A dataset that contains Community Mobility Reports providing insights into the changes in response to policies while combating COVID-19

ity dataset, since it contains dates and periods, allowing us to measure the timeliness and currency quality dimensions. Hence, we aim to measure the effectiveness of the defined inference rules. To do so, we propose a tool that translates the requirements set by the user into SQWRL queries to query the ontology. Then, we use the reasoner PELLET [94] to reason over the ontology to infer each quality dimension value. Thus, we noticed that our ontology has successfully computed the value of each data quality dimension using the defined inference rules. For instance, we found that the ontology has assigned 100% for the trustworthiness quality dimension since a trustworthy provider provides the dataset. We also elaborated on further evaluation to measure ontology performance using metrics. Thus, such metrics measures the ontology's structural and knowledge qualities that reflect its functional, analytical, pragmatic, syntactic, cognitive, semantic, social, and practical capabilities. For instance, structural quality may impact later activities such as ontology merging and alignment. On the other hand, knowledge quality may measure the extent of the richness of an ontology[95]. Hence, we relied on schema, knowledge, and graph evaluation metrics to evaluate the presented dimensions. Tables 4.2-4.4 depict the evaluation results. Table 4.2 shows the structural richness of the data characterization and evaluation ontology structure. Indeed, more than one-third of the ontology's structure is represented via inheritance and relation shapes, illustrating the richness of our ontology's structural quality. In addition to its simplicity (i.e., non-complex ontology), the performance metrics show the knowledge richness depicted via the absolute and maximal breadth. The vast knowledge provided by this ontology is guaranteed by the average population and the class richness values. As mentioned above, the structural and knowledge qualities ensure our ontology's ease of merging, alignment, and reuse.

Table 3.2: Schema evaluation metrics

Inheritance richness	0.30
Relationship richness	0.46
Axiom/Class Ratio	6.15
Class/Relation Ratio	1.76

Table 3.3: Knowledge metrics

Average population	1.13
Class richness	0.3

3.6.2 Scalability according to the number of services and users

These experiments aim to assess the scalability of our curation service composition method according to the number of users using it simultaneously and the number of curation services. To do so, we used multithreading to create a simulation environment to simulate the curation service composition, which may be executed by several users simultaneously.

Table 3.4: Graph evaluation metrics

Absolute root cardinality	5
Absolute root node	16
Absolute leaf cardinality	16
Absolute sibling cardinality	16
Maximal depth	2
Absolute breadth	16
Maximal breadth	8

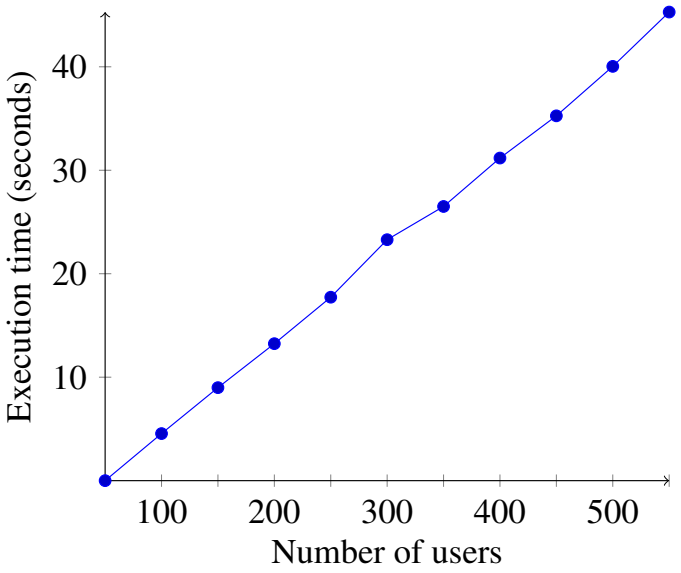


Figure 3.12: Execution time per number of users

Therefore, we developed a set of threads, where each one simulates a curation service composition request by one user. Moreover, to have similar conditions during the experimentation, we defined the same input parameters (i.e., user preferences, constraints, etc.) for all the threads. Then, we progressively executed and increased the number of threads to examine our proposal's response time to these queries. Indeed, we considered the average execution time as a result. Figure 3.12 depicts the overall execution time according to the number of threads. As we performed curation service composition for three different data structures, we tested whether the data type impacted the composition process. Accordingly, we noticed that the data source type does not affect our approach performance.

Regarding the scalability according to the number of services, we simulated the increasing number of services by increasing the size of the Q-Table. Indeed, each Q-Table entry corresponds to one service. We executed the service composition three times at each iteration and took the average execution time. We defined random rewards to generate

Q-Tables of different sizes. Afterward, we examined the curation service composition for each data source. Figure 3.13 depicts the evolution of execution time according to the number of states. Following these experiments, we noticed that the overall process execution time was less than one second using a Q-Table of size less than 4000x4000. Otherwise, the execution time can reach about 7 seconds when the size of the Q-Table is 12000x12000. Accordingly, the experiments showed that the service composition scheme is generated in near real time using 12000 services. However, in our case, the Monte Carlo and Temporal-difference algorithms cannot generate a service composition scheme using more than 200 services. Thus, the results show that our proposed method outperforms the two reinforcement learning algorithms. We also noticed that the composition process lasts almost the same regardless of the data source types. We extended these experiments by evaluating the performance of our proposed curation service composition approach against service composition benchmarks and baselines that have proven their effectiveness in composing services, such as greedy randomized adaptive search procedure (GRASP) [96], random composition [97], ant colony [98], k nearest neighbors (KNN) [99], and genetic algorithm (GA) [100]. Thus, we first employed the library of curation services, consisting of 18 services, and then simulated the increasing number of services to assess the scalability using 50, 100, 200, 300, 1000, 10000, and 12000 services. We intentionally stopped the experiments when the run time exceeded 1 hour, as we assumed it was already immensely time-consuming. Table 3.5 compares the execution time of each service composition method regarding the number of services. As depicted in the Table, the K nearest neighbors, the genetic algorithm, and the GRASP algorithm take too long to generate a curation service composition scheme. These algorithms take more than an hour to create a composition scheme using a library of curation including more than 200 services. This huge execution time may be explained by the long training process required to generate a composition scheme. Nevertheless, considering data lakehouses, they may contain data ingested in real time that needs curation promptly. Hence, it is tedious to launch a composition process that takes too much time for each ingested data source. Moreover, we emphasize that user requirements regarding curation may be unstable and highly changeable. Thus, re-executing a training process that takes too much time to cope with these changes may take time and effort. Accordingly, these composition algorithms may not be convenient for this kind of service composition characterized by dynamicity, uncertainty, and a highly changeable environment. We also investigated the performance of the curation service composition using random and ant colony algorithms. Although the random algorithm needs less time to generate a composition scheme, it generates composition schemes randomly that may sometimes be invalid or contain non-convenient services. Similarly, the ant colony algorithm showed good performance in terms of execution time. However, both algorithms (i.e., random and ant colony) require more than 2 minutes to generate a composition scheme using a library of services grouping more than 10,000 services. The table shows that our curation service composition approach outperforms the examined composition algorithms regarding execution time and scalability since it generates a scheme using more than 10,000 services in less than 7 seconds.

Regarding the complexity of the designed algorithms, we found that the complexity of

Table 3.5: Comparison of the performance of different service composition methods

Number of services	KNN	GA	GRASP	Random	Ant	The proposed method
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

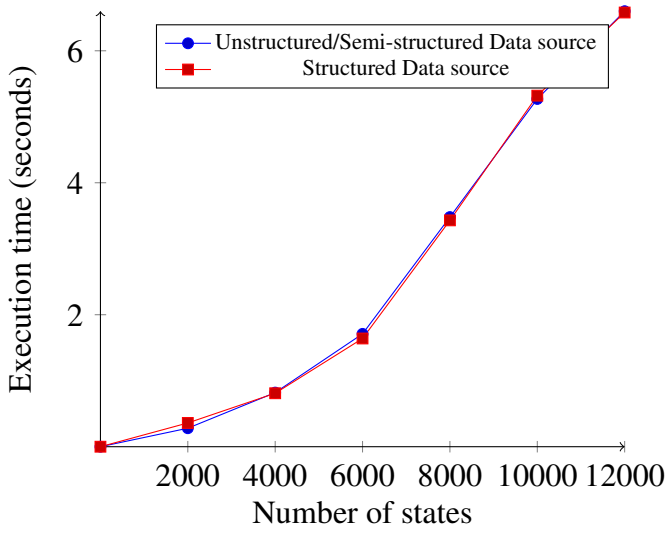
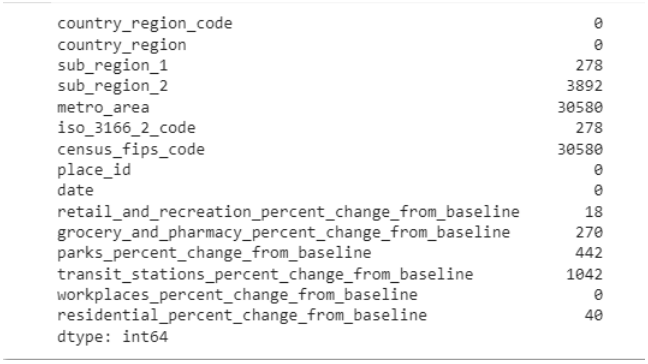


Figure 3.13: Execution time per number of states

the training algorithm is $O(E)$ and the complexity of the composition algorithm is $O(n)$, where E represents the number of episodes and n is the number of environment states. Thus, linear complexity is a promising result. Indeed, linear complexity algorithms perform better in execution time than other algorithms with different complexity classes, such as quadratic complexity. Also, we found that the algorithm converges from 400 episodes for a Q-Table of size 18x18, which makes it converge quickly. Hence, ACUSEC performs well in terms of execution time and scalability.

3.6.3 Effectiveness of the data curation process

Following the evaluation of the generated curation service composition scheme, we conducted further experiments to assess the effectiveness of the data curation process using different schemes. Hence, we invoked the services constituting the composition schemes to examine the curation process for different data structures. In other words, we inves-



```
country_region_code                                    0
country_region                                         0
sub_region_1                                         278
sub_region_2                                        3892
metro_area                                         30580
iso_3166_2_code                                      278
census_fips_code                                   30580
place_id                                               0
date                                                   0
retail_and_recreation_percent_change_from_baseline    18
grocery_and_pharmacy_percent_change_from_baseline    270
parks_percent_change_from_baseline                   442
transit_stations_percent_change_from_baseline       1042
workplaces_percent_change_from_baseline                0
residential_percent_change_from_baseline              40
dtype: int64
```

Figure 3.14: The extracted statistics about missing data

tigated the impact of the selected services for data curation using datasets with different structures. Specifically, we relied on the same "COVID-19 Community Mobility Reports" dataset in its structured form to conduct the first experiment. We also employed an unstructured dataset, containing health tweets, to conduct the second experiment. We relied on the data characterization and evaluation ontology to evaluate data source quality and identify the main characteristics needed for data curation. Then, we evaluated the effectiveness of the data curation process via three-step experiments, namely: (i) the generation and verification of a curation service composition scheme, (ii) validation of the outcomes, and (iii) evaluation of the effectiveness of the data curation.

Experiment 1. The characterization ontology describes the dataset as structured and requires specific curation tasks dedicated to structured data. (i) Considering the COVID-19 Community Mobility Reports dataset, the generated curation service composition scheme is as follows: Metadata extraction service → Descriptive statistics service → Missing values service → Terminology extraction service → Lexical Service → Rules extraction service → Entity extraction service → Linking service → Synonym service.
(ii) This composition scheme contains services adapted for structured data source curation. Specifically, the metadata extraction and the descriptive statistics services provide statistics about the dataset that may be needed later for data analysis, like the number of features, the size of the dataset, the ranges of the values, the number of continuous, categorical, and discrete features, and the number of missing values. Figure 3.14 depicts an extract from the generated metadata and descriptive statistics about missing data.
(iii) Finally, we investigated the effectiveness of the curation services by verifying whether the curation process would repair the identified missing data. After invoking the missing data service, we found that our curation framework had successfully filled in the missing data. Nevertheless, fields like "meteo_area" and "place_id" are kept empty because the number of missing data in this field equals the number of rows in the dataset. As for the "iso_3166_2_code" and the "sub_region_2" attributes, they still contain missing values since some regions do not have a second sub-region and may not have an ISO 3166-2 code. Figure 3.15 depicts the number of missing data by feature after invoking the missing values service. Then, we monitored the invocation of the next curation services present in

```
country_region_code                                    0
country_region                                         0
sub_region_1                                           0
sub_region_2                                         556
metro_area                                         30580
iso_3166_2_code                                      278
census_fips_code                                   30580
place_id                                               0
date                                                   0
retail_and_recreation_percent_change_from_baseline     0
grocery_and_pharmacy_percent_change_from_baseline      0
parks_percent_change_from_baseline                     0
transit_stations_percent_change_from_baseline          0
workplaces_percent_change_from_baseline                0
residential_percent_change_from_baseline               0
dtype: int64
```

Figure 3.15: The missing data statistics after invocation of the missing data curation service

the composition scheme (i.e., Terminology extraction service → Lexical Service → Rules extraction service → Entity extraction service → Linking service → Synonym service) constituting the scheme to examine their impact on data source curation.
We stated that the terminology service extracts the dataset features' names to construct a reference model employed by the rules extraction service. The latter identifies rules related to features such as the maximum and minimum value ranges to detect any semantic violation (i.e., using the extracted rules) and, therefore, check any possible anomaly that may degrade dataset consistency. Then, the entity extraction service extracts the named entities to be linked with external knowledge bases and enriched via the linking and synonym extraction services. As most rows in this dataset contain numeric values, the role of linking and synonym extraction services is not apparent here. Thus, we emphasize their roles in the second experiment since the employed dataset is unstructured and contains several tweets constituting a set of words that need enrichment.

Experiment 2. In this experiment, the data characterization ontology identified data characteristics such as the unstructured form of the dataset and rows containing URLs. (i) Based on these characteristics, the curation layer generates the following curation service composition scheme: (URL extraction → Entity extraction service → Linking service → Synonym service).
(ii) We noticed that our framework selected, in this case, only the services devoted to extracting and enriching data, which is appropriate regarding the tweets' characteristics.
(iii) Finally, we investigated the results generated by this composition scheme to evaluate the effectiveness of the enrichment process. Figure 3.16 illustrates an example of a tweet about cancer enriched with information extracted from the URL (i.e., the URL that figures in the tweet) and the keywords linked with external knowledge bases. The enrichment is performed via the entity extraction service that described the term "cancer" as a cause of death.

Since the tweet contains an URL from the BBC website, the URL extraction service fetches further information, such as "The international team analyzed 77 genes". The linking service, in turn, extracts more information from external ontologies, like that breast cancer is a "cancer that originates in the mammary gland". A domain expert validated the

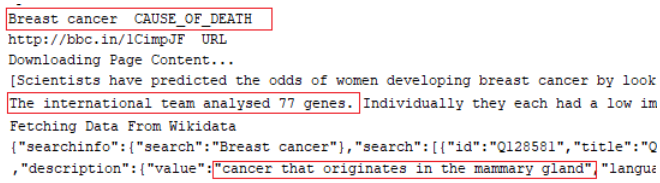


Figure 3.16: An extract from the enrichment information for a tweet concerning breast cancer

obtained enriched tweets, demonstrating the efficiency of the service composition scheme and the appropriate curation. These tweets can thus be used later in data analysis within a prediction model, for example, to forecast cancer cases.
To sum up, the conducted experiments prove our proposal's effectiveness in data repair (e.g., replacing missing data) and enrichment by fetching additional information from trusted sources to avoid data noise.

3.6.4 Adaptivity to changes

We defined a set of experimental scenarios to measure the adaptivity of our approach, ACUSEC according to the user's functional and non-functional requirements. The testing scenarios[7] focus on changes regarding the data source type, user preference and constraints, and QoS changes. In each scenario, we defined different user preferences and constraints for each data source type to check their impact on the curation service composition. We also randomized the QoS values to check their impact on the composition process. We also compared the generated curation service composition scheme via ACUSEC with the static composition of services. This experiment aims to analyze the added value of the dynamic composition process compared to the static one. Technically, we have implemented a prototype that allows the user to define the input parameters, such as the data source type, user preferences regarding a decision context, and user constraints, as depicted in Figure 3.18. The prototype also allows the user to select data source characteristics, like whether the source contains URL values and whether the data source needs to be converted or indexed. This experiment examines the impact of such characteristics on the generation of the curation service composition scheme. Moreover, our prototype allows the end user to choose the execution of specific curation tasks, such as PoS Tagging or enrichment process. On the other hand, it allows him to prevent or promote the presence of a curation service in the composition of services. In the first evaluation scenario, we investigated the curation service composition scheme changes according to the data source type. Indeed, as we presented previously, some curation services may be convenient exclusively for curating structured or unstructured/semi-structured data sources. For this purpose, we generated curation service composition schemes for each data source structure type. Moreover, we have defined equal preferences for all the QoS dimensions and specified that the decision context is ordinary. Following these experiments, we noticed that our composition approach could distinguish curation services convenient for

[7] https://www.youtube.com/channel/UCoxXOUewbVQy6IROo9DnlUA

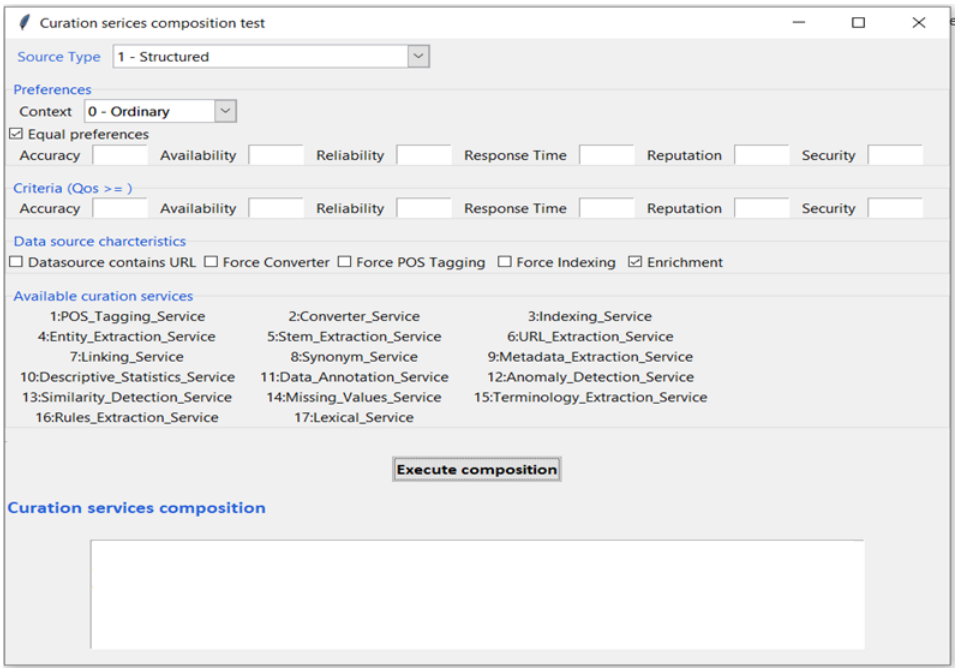


Figure 3.17: An overview of the tool designed to generate compositions based on ACUSEC

each data source type. As common curation services may curate unstructured and semi-structured data sources, we illustrate in Figure 3.18 and Figure 3.19 different curation service compositions for unstructured and structured data sources. By examining the two composition schemes, the results depict that the two compositions are entirely different. For instance, the second service composition scheme includes services dedicated only to structured data sources curation like Metadata Extraction Service, Descriptive Statistics Service, Missing Values Service, Terminology Extraction Service, Lexical Service, and Rules Extraction Service. The second experimental scenario focuses on investigating the impact of user preferences and constraints on the generation of the curation service composition scheme. To do so, we generated two curation service composition schemes, one using equal preferences and the other using the following user preferences: "Accuracy: 10%, Availability: 10%, Reliability: 50%, Response Time: 10%, Reputation: 10%, Security: 10%". The sum of all user preferences should equal 100%. Accordingly, we noticed that our approach generates two different composition schemes according to each user's defined preferences. Although the data sources may have a common structure (i.e., a structured data source in the presented example), the changing user preferences impacted the curation service composition process by replacing some services that were not present in the first composition. In this example, the entity extraction service has been replaced by the stem extraction service because its reliability QoS (90%) is greater than the entity extraction service's reliability (80%). Following this experiment, we tested the impact of user constraints on the curation service composition process. Specifically, we defined the following constraint (Accuracy >= 80%) to generate a composition scheme and specified equal preferences for all QoS dimensions. Thus, the approach generated a different

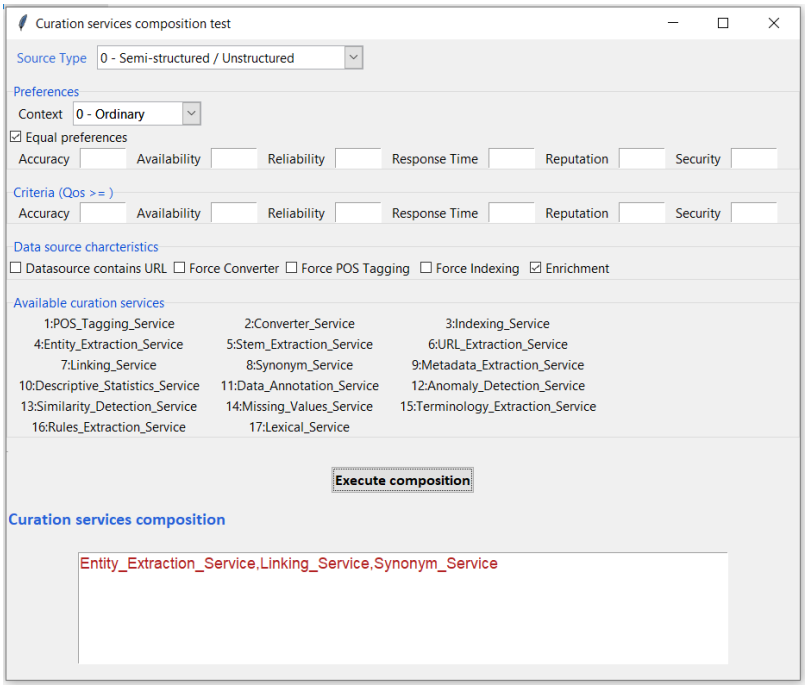


Figure 3.18: Curation service composition generated using the prototype for unstructured data source

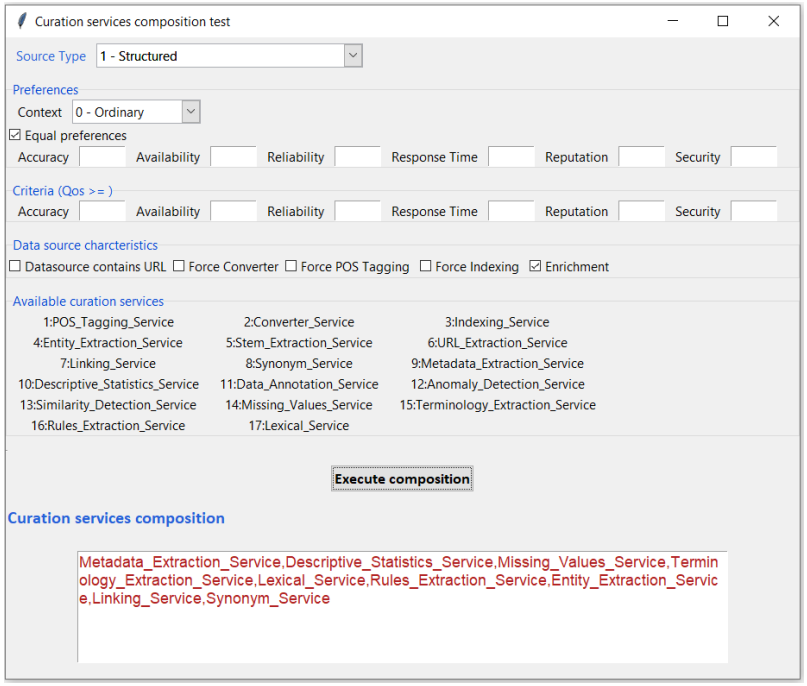


Figure 3.19: Curation service composition generated using the prototype for structured data source

curation service composition scheme from the one without constraints. Indeed, the rules extraction service and the linking service do not appear in the service composition because they do not fulfill the imposed constraints. The accuracy of the rules extraction and the linking services are equal to 76% and 71%, which are lower than 80%.

We also investigated our approach's performance regarding the changes in the QoS values to examine their impact on the curation service composition process. Specifically, in each experimental iteration, we assigned random QoS values to perform several curation service compositions for each data source structure. Thus, our approach can consider changeable QoS values by selecting the services with the highest quality. Our experiments also evaluate the effectiveness of our composition approach in selecting the curation services according to QoS and user preferences. To do so, we defined three curation services with different qualities, each ensuring a specific curation task. Through this experiment, we aim to investigate the cost of the generated composition scheme. For instance, we consider the stem extraction services ST1, ST2, ST3, and the synonym extraction services SY1, SY2, and SY3 candidates to constitute the abovementioned scheme. By investigating the proposed composition scheme, we noticed that our method selected ST2 and SY1, which have the highest Response Time QoS (i.e., ST1: 71%, ST2: 85%, ST3: 50%, SY1: 99%, SY2: 78%, SY3: 84%). Then, we changed user preferences by promoting the accuracy dimension (Accuracy: 50%, Availability: 10%, Reliability: 10%, Response Time: 10%, Reputation: 10%, Security: 10%) to check whether this would have an impact on the curation service composition scheme. Thus, we noticed that our approach adapts well to the changeable preferences and QoS since the generated composition scheme encompasses ST3 and SY3, which are the most accurate services (i.e., ST1: 48%, ST2: 66%, ST3: 77%, SY1: 77%, SY2: 40%, SY3: 98%).

Then, we investigated the added value of the dynamic service composition compared to the static one. This experiment aims to simulate the curation service composition in a critical decision context that may require high accuracy and low response time. Such quality dimensions have high importance in such decision contexts. We assume a semi-structured data source is curated using a static curation pipeline composed of the following services: Stem Extraction Service, Synonym Extraction Service, and Linking Service. As depicted in Table 3.6, each curation service is characterized by different quality dimensions. In this example, we consider the accuracy and response time quality dimensions. We assume that the user imposes the following constraints: (Accuracy >75% and Response Time >70%). However, the static curation pipeline does not consider the user constraints regarding QoS values. Hence, it invokes curation services that may not meet the user's requirements and handicap the curation process.
Contrary to static curation, our ACUSEC approach considers the defined constraints and generates the appropriate curation service composition constituted of "Stem Extraction Service ->Synonym Extraction Service". Indeed, ACUSEC has eliminated the Linking Service from the generated curation service composition even though its response time QoS value was greater than 75% since its accuracy did not meet user requirements.
We also investigated the adaptivity of our proposal according to the characteristics of the

Table 3.6: Extract from curation services QoS values

Service/QoS	Availability	Accuracy	Reliability	Response Time	Reputation	Security
Stem Extraction Service	70%	80%	90%	70%	90%	70%
Linking Service	88%	71%	83%	92%	83%	86%
Synonym Extraction Service	85%	81%	79%	87%	78%	74%

treated data source. For instance, following the characterization of a semi-structured data source containing URLs, the scheme generated by our ACUSEC (Stem Extraction Service $\rightarrow$ URL Extraction Service $\rightarrow$ Entity Extraction Service $\rightarrow$ Synonym Extraction Service) contains a service dedicated to fetching data from URLs. In another scenario, we present a schema generated for streaming data. By investigating the scheme generated for the streaming data (Anomaly detection Service $\rightarrow$ Stem Extraction $\rightarrow$ Entity Extraction $\rightarrow$ Linking Service), we noticed that the generated scheme is different from the other schemes generated for batch data since it contains a service dedicated to streaming data.
Following these experiments, we state that the experimental results are encouraging in terms of execution time and adaptivity to functional and non-functional requirements.

3.6.5 Alignment with user needs

We conducted experiments on the presented datasets to assess our curation service composition method's alignment with user expectations. As we adopted the reinforcement learning paradigm, we relied on the returned cumulative reward to assess the alignment of our approach with user needs in terms of QoS, user preferences, and constraints. Indeed, as the value of the cumulative reward increases, the curation service composition scheme becomes more aligned with user needs. To do so, we used the library of curation services to generate service compositions for the presented datasets using ACUSEC (i.e., which employs Q-Learning), First-Visit Monte Carlo, and Temporal Difference algorithms. We defined similar experimental settings for all the tested algorithms. Specifically, we defined similar user preferences, QoS values, decision context, user constraints, and data source formats. We took the average rewards as a result and presented the performance of each algorithm in Figure 3.20. As depicted in this Figure, our curation service composition aligns better with user needs than the First-visit Monte Carlo and Temporal-difference Learning algorithms since it returns a higher cumulative reward. Indeed, the cumulative reward gained by our service composition method exceeds 9, while the maximum rewards returned by the other reinforcement learning algorithms are less than 6.

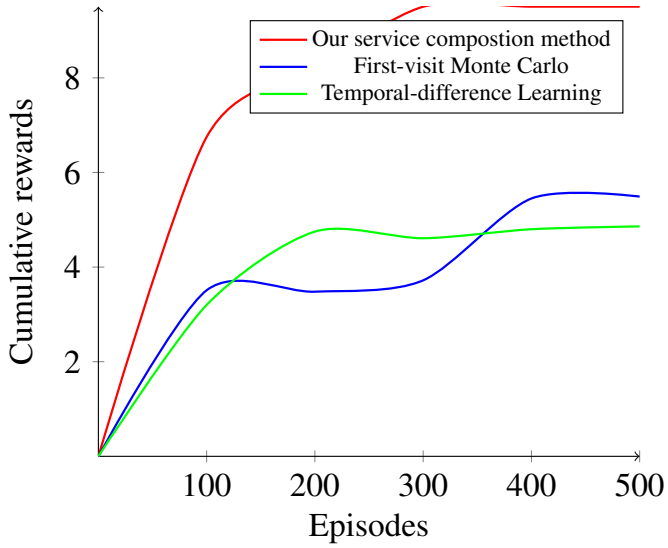


Figure 3.20: Cumulative rewards by each algorithm

3.7 Conclusion

3.7.1 Summary

The presented chapter dealt with the first research question of thi;s thesis, which encompasses the two following sub-questions, namely, "How to perform data curation for multi-kind of data collected in batch and streaming?" and "How to consider the needs of different users in different contexts while performing data curation?". We proposed a new approach for adaptive data curation. Specifically, our proposed approach considers the user's functional and non-functional requirements to generate a curation service composition scheme. To do so, we followed the AOM methodology to design, DARQAN, a modular ontology that extracts the characteristics of data sources and evaluates their quality from different perspectives, such as data provenance, the platform used to collect data, quality dimensions, standards, user feedback, etc. Then, our approach relies on the extracted data sources characteristics as well as the reinforcement learning paradigm to compose curation services from a library of services we design using existing curation services. We have also validated our proposed approach via experiments evaluating scalability, adaptivity to changes, and alignment with user needs.

3.7.2 Limitations & Enhancement ideas

Extension of the library of curation services

We figured out that the library of curation services could be more enriched by including more services that ensure other curation tasks. Indeed, the library of services encompasses

services that perform extraction, enrichment, data standardization, and quality evaluation. Hence, we believe that these categories of services could be extended by adding more services that perform more granular tasks for curation. For instance, an entity resolution service could be added to the library to be included later as curation in the generated service composition scheme. Accordingly, we can investigate later the effectiveness and efficiency of adding such tasks to the curation pipeline.

Consider other adaptivity dimensions

In future work, we aim to consider other adaptivity dimensions for curation service composition. Indeed, our approach generates a composition scheme according to user preferences, constraints, the decision context, the quality of services, and the characteristics of the data source. Thus, we think that considering other dimensions could be one of the exciting research tracks. For instance, we can define more characteristics that detail the decision context of the user, according to the user role. Accordingly, presenting more granular information related to the decision context and user preferences is possible. On the other hand, we think that considering more quality dimensions to assess the quality of services could be one of the possible improvement tracks.

CHAPTER 4
Towards an explainable recommendation approach for crisis management

CHAPTER 4

Towards an explainable recommendation approach for crisis management

> Time and health are two precious assets that we don't recognize and appreciate until they have been depleted
>
> Denis Waitley

Contents

4.1 Introduction

In the previous chapter, we presented our proposed data curation approach, which considers the user's functional and non-functional requirements to adaptively generate a curation service composition scheme. The proposed approach adapts the composition scheme according to the changing situation of the user and its decision context. We also proposed an evaluation protocol that assesses the effectiveness of our proposal in terms of execution time, adaptivity to changes, and alignment with user needs.

Following the data curation step, the data are analyzed to predict the risk of the occurrence of a crisis. Subsequently, selecting suitable actions will be necessary to manage the crisis in the case of an identified risk. Accordingly, stakeholders can adopt these actions to cope with the situation and prevent it from worsening. These actions constitute strategies for monitoring and controlling emergencies. The response strategies depend on the country's characteristics, which make them different from one country to another. For instance, in the healthcare field (see Chapter 2), since the discovery of the first case of COVID-19 infection, China has adopted strict health measures to cope with the virus, such as the lockdown of Wuhan city and the "Four early's" measures (i.e., early detection, early reporting, early isolation, and early treatment). On the other hand, South Korea first adopted less strict measures like border control, screening, and testing. Then South Korea imposed strict blockades in some provinces. Unlike China and South Korea, Japan adopted a mitigation strategy through different stages to reduce the spread of virus transmission. The effectiveness of the different strategies depends on several factors, such as the country's characteristics (i.e., population, Human Development Index), the situation in the country, and the outbreak severity [51]. Accordingly, these factors must be considered while setting up health strategies to evolve health measures and adopt them according to the changing situation. Considering a crisis management system, it needs to learn from its previous and other countries' prevention and management strategies. For this purpose, several recommendation approaches were proposed in the literature. As aforementioned, these approaches rely on diverse techniques like collaborative filtering, content-based recommendation, and deep learning-based recommendation [101]. Despite the superiority of deep learning-based recommendations, deep learning models remain complicated and have low interpretability. Hence, they cannot explain the decisions generated, a common problem haunting the deep learning community [102]. Hence, the challenge is to tackle the adaptive explanation of recommendations for different user roles using various explanation types.

To attempt this challenge, this chapter presents our two-fold contribution:

- A multi-output deep learning-based model for measures recommendation. The proposed model generates the stringency of measures related to crisis management

while considering multi-user roles, needs, and multi-country characteristics.

- A semantic-based approach that explains the recommendation models adaptively, while considering different users' roles, preferences, and decision contexts. The proposed approach dynamically constructs an explanation ontology by mapping with external ontologies. The constructed ontology encompasses several explanations (i.e., neighboring countries, counter-examples, etc.) adapted for different users. Then, based on matrix factorization techniques, we extract the suitable explanation sub-graph according to the user's role and needs.

We highlight that we treat the explainability problem as post-hoc explainability, which focuses on the explanation of the non-interpretable models. These models differ from the IA models interpretable by design (e.g., decision trees) (see Chapter 2). For this purpose, our contribution relies on eXplainable Artificial Intelligence (XAI) techniques.

The remainder of this chapter is organized as follows: Section 4.2 represents a motivating scenario highlighting our proposed contributions' challenges. Section 4.3 depicts the proposed cross-country deep learning-based model for measures recommendation for crisis management that is adopted for multi-user needs. Section 4.4 illustrates the proposed approach that adaptively explains the recommendation model for multi-user roles. Section 4.5 details the implementation and the conducted experiments that assess the effectiveness of our proposals. Finally, Section 4.6 concludes the chapter and provides an overview of the limitations and future perspectives.

4.2 Motivating scenario

As depicted in Figure 4.1, we assume Alice and Bob are using a crisis management system that recommends health measures (i.e., preventive actions) to manage health outbreaks and prevent them from worsening. Such users may need different explanations according to their role to understand the system's choices and attribute their trustworthiness to the system. Bob may be interested in explanations expressed as medical information such as symptoms, nature of tests, treatments, etc. Based on this information, Bob will understand the nature of the pathogen agent that caused the health outbreak (e.g., virus, bacteria, fungi, etc.), which justifies the choice of the recommended treatments. For instance, infections caused by viruses (e.g., colds and flu) must be treated via vaccines, ensuring preemptive protection by training the body's immune system. However, bacterial infections (e.g., tuberculosis) are more complex than viruses and spread through the air. Indeed, a bacterium can live and reproduce almost anywhere (e.g., soil, water, the human body, etc.). Hence, they require treatment with antibiotics. Considering fungi infections (e.g., valley fever), they are more complicated than viruses and bacteria.
Nonetheless, they are slower to mutate, which makes them easier to target with antifungal medications than bacteria are with antibiotics[1]. Similarly, Alice may be interested in other types of explanation, such as statistics or the situation in neighboring countries. This

[1] https://www.cedars-sinai.org/blog/germs-viruses-bacteria-fungi.html

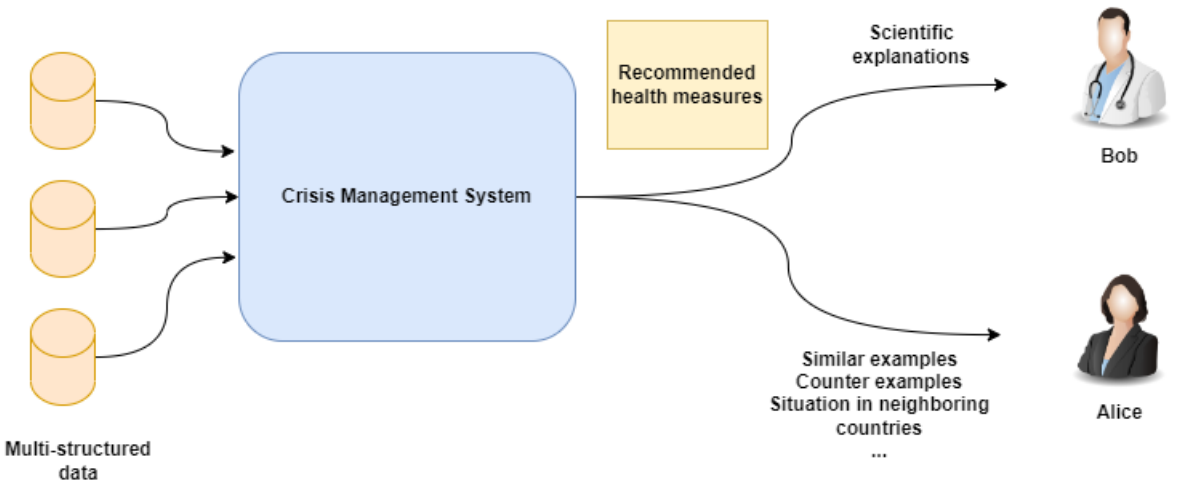


Figure 4.1: Overview of multi-users with different needs using a crisis management system

type of explanation is primordial for a decision-maker to better understand the situation and take suitable strategic actions to prevent it from spreading in the country [103]. For instance, when an infectious virus spreads in a neighboring country, it will be safer to limit travel there or close borders in some cases. In other cases, similar examples (e.g., the situation in countries that adopted the health measures) or counter-examples may help understand the recommended health measures and ensure trustworthiness between the user and the system. Accordingly, a crisis management system should provide different explanations adapted to each user's role.
Hence, our challenge in the present work focuses on generating recommendations for multiple users with different needs and requirements. In particular, we want to consider the needs and requirements of the users in terms of recommendations and explain the choice of recommendations appropriately for each user role. We detail in the following sections the proposed solution to overcome the presented issues.

4.3 Model for crisis management measures recommendation

Crisis management requires effective preparedness, planning, response, and continuous evaluation and improvement. Hence, recommendation models could be a practical solution that offers personalized and efficient recommendations to manage crises while continuously improving them based on past experiences. Therefore, we aim to attempt this objective by proposing a multi-output deep learning-based recommendation model. For this purpose, after predicting the occurrence and cause of a crisis (e.g., the pathogen causing the risk), the recommendation step needs to identify, firstly, the recommendation model convenient to treat the predicted crisis. Then, the model should recommend measures addressed to several user roles incorporated into the crisis management process, such as health and economic experts and strategic analysts in the health field, in our case. Strategic

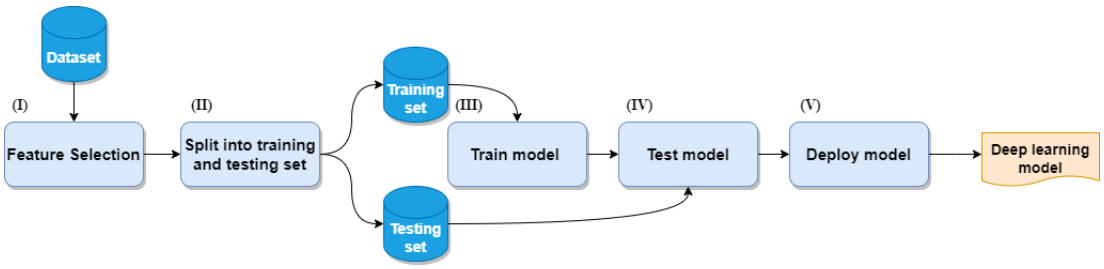


Figure 4.2: Overview of the recommendation model construction steps

analysts may be interested in measures impacting daily life and people's movements, like school and border closures, restrictions on gatherings, etc. However, economic experts may focus on other financial measures, such as income support and debt or contract relief for households.

Thus, we designed a multi-output deep learning model that takes advantage of previous experience in dealing with crises to recommend future measures by predicting each measure's future stringency scale. Moreover, our recommendation model considers the country's characteristics to recommend suitable measures, making it the first cross-country measures recommendation model to the best of our knowledge. Figure 4.2 depicts the process adopted to constitute the recommendation model, constituted of five steps: (I) feature selection, (II) splitting of the dataset into training and test sets, (III) training, (IV) test, and (V) deployment of the model. First, we applied (I) a feature selection process to select the best feature combination that maximizes the model's performance. To do so, we adopted the measures proposed in the Oxford Covid-19 Government Response Tracker (OxCGRT) [104], which leverages prior countries' experience in terms of crisis management, particularly disease outbreaks. OxCGRT provides standardized measures and metrics that assess the efficiency of the government responses during the entire period of the disease's spread. Moreover, this framework is updated regularly following the evolution of the pandemic and governments' responses. Several works have employed OxCGRT data and metrics to measure the effectiveness of the adopted health policies, such as [105], [106], [107], and [108]. Although the OxCGRT measures (i.e., sometimes referred to as indicators) are employed to measure the stringency of the government's health policies, we identified proposals that used these measures in other fields and contexts [109], such as in the environmental [110, 111] and political [112] fields.

The OxCGRT framework combines 21 indicators (i.e., considering the version released in August 2022), combining different measures. This framework's ordinal measures (i.e., different from non-ordinal ones that do not have a stringency scale) have a scale of stringency of application. For instance, the indicator "school closing" contains the following measures: (1) recommendation of school closures or schedule changes; (2) obligation to close certain school levels (e.g., primary schools); and (3) obligation to close all school levels. Hence, each measure may represent the stringency level of application of a health indicator. The OxCGRT indicators are grouped into five groups, namely containment and

closure policies (C), economic policies (E), health system policies (H), vaccination policies (VC), and miscellaneous policies (M). Containment and closure policies encompass measures related to school closures and restrictions on movement. On the other hand, economic policies include measures related to income support or the provision of foreign aid. As for health system policies, the latter concerns measures such as testing policies, facial covering, and contact tracing. Regarding vaccination policies, they record aspects related to vaccination, such as the eligible groups, the cost of vaccinations, etc. The last category, miscellaneous policies, is defined to allow the user to add other information as plain text. Table 4.1 depicts the indicators proposed in the OxCGRT framework. We emphasize that the non-ordinal indicators are described via text, numeric and categorical information. Moreover, we stress that the indicators E3, E4, and H4 are removed from the latest OxCGRT version.

Table 4.1: Measures included in the OxCGRT framework
(NO : Non-ordinal indicator (i.e., do not have a scale of stringency))

Category	ID	Measure	Scale
Containment and closure policies	C1	School closures	3
	C2	Workspace closing	3
	C3	Cancel public events	2
	C4	Restrictions on gathering	4
	C5	Public transportation	2
	C6	Stay at home order	3
	C7	Restrictions on internal movement	2
	C8	International travel controls	2
Economic policies	E1	Income support	2
	E2	Debt/contract relief for households	2
Health system policies	H1	Public information campaigns	2
	H2	Testing policy	3
	H3	Contact tracing	2
	H5	Investment in Covid-19 vaccines	NO
	H6	Facial covering	4
	H7	Vaccination policy	5

Table 4.1 **– continued from previous page**

Category	ID	Measure	Scale
	H8	Protection of elderly people	3
Vaccination policies	V1	Vaccine prioritisation	NO
	V2	Vaccine eligibility/availability	NO
	V3	Vaccine financial support	NO
	V4	Mandatory vaccination	NO
Miscellaneous	M1	Other responses	NO

Following a feature selection process, we select the most consistent, non-redundant, and relevant features to construct the deep learning model. We have chosen nine out of fifteen features as input features, including country, actual indexes (i.e., stringency, government response, containment health, and economic support), reproduction rate, positive rate, population rate, median age, and life expectancy, and Human Development Index. As for outputs, we selected the categorical features related to the measures related to closure (i.e., C1-C8), economic (i.e., E1, E2), and health (i.e., H1, H2, H3, H6, H7, H8) policies. It is important to note that we selected features with a stringency scale that could be predicted via our proposed deep-learning model. Accordingly, we rely on the measures rather than the non-selected measures from the presented categories (e.g., E4) and the vaccination measures since they do not have a stringency scale. Then, we rely on Multi-output deep neural networks to constitute a model including four hidden layers and sixteen output layers (i.e., according to the number of the selected measures), in which each output layer predicts the stringency level of each measure. For instance, the measure "School Closures" is characterized by a three-level severity scale ranging from recommendation to obligation for school closures. Hence, the output layer devoted to C1 prediction predicts four classes (i.e., no measure applied, recommending school closures or schedule changes, etc.). We formally describe the architecture of the proposed model through the following equations:

$$f(x) = W_o * h_4 + b_o \tag{4.1}$$

$$h_4 = W_{h4} * h_3 + b_{h_4} \tag{4.2}$$

$$h_3 = W_{h3} * h_2 + b_{h_3} \tag{4.3}$$

$$h_2 = W_{h2} * h_1 + b_{h_2} \tag{4.4}$$

$$h_1 = W_{h1} * x + b_{h_1} \tag{4.5}$$

where $h1$, $h2$, $h3$, $h4$ are the outputs of the first, second, third, and fourth hidden layers, respectively. W_{h1}, W_{h2}, W_{h3}, W_{h4}, b_{h1}, b_{h2}, b_{h3}, b_{h4} are the weight matrices and biases for the different hidden layers. Then, (II) we split the OxCGRT dataset, which contains the previous experiences of several countries in crisis management, into training and testing sets. This step could be performed thanks to two popular methods, namely 80/20 split or

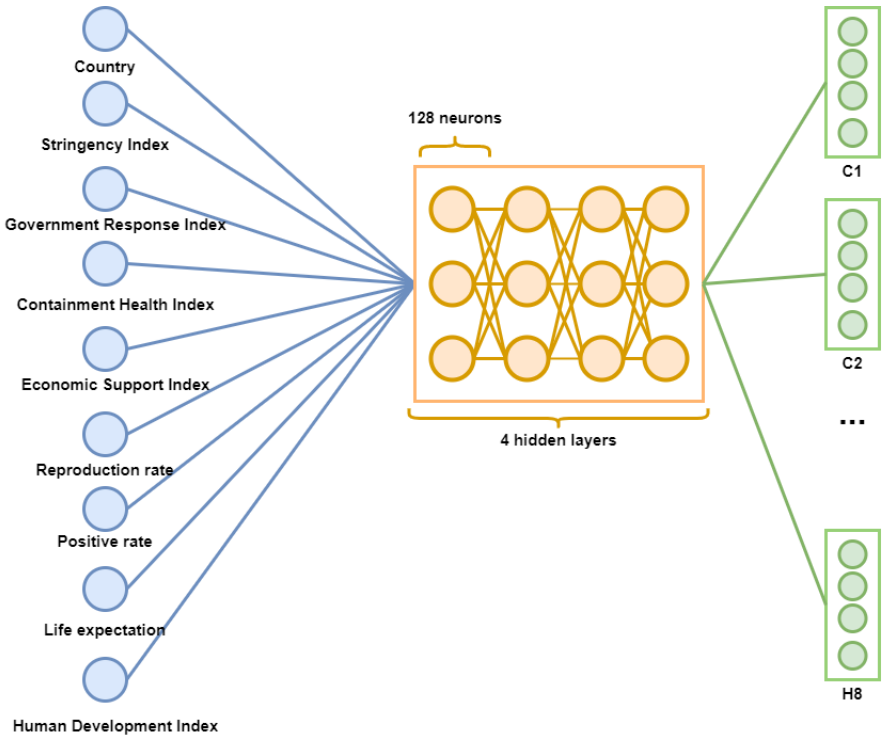


Figure 4.3: Overview of the deep learning-based health measure recommendation model

k-fold validation. For the third step, (III) we trained the recommendation model using the training set and the Sparse Categorical Cross-Entropy as a loss function and regularized using the ADAM stochastic optimization algorithm [113], which is an algorithm used for first-order gradient-based optimization of stochastic objective functions. This algorithm is based on adaptive estimates of lower-order moments. Thus, the optimizer modifies hyperparameters and minimizes the loss function. Regularization is helpful in calibrating machine learning models to adjust the loss function. After the training process, (IV) we evaluate the performance of the proposed model using the testing set, as we detail in Section 4.5. Finally, we (V) deployed the recommendation model when it became stable. The following section illustrates the semantic-based approach that we propose to explain the choice of this recommendation model.

4.4 Semantic approach for the explanation of recommendations

No one can deny the high performance ensured by deep learning models. Although effective, they act like "black boxes" since they are complicated and less interpretable than other AI models, like glass-box models (e.g., rule-based systems). Hence, we proposed a semantic-based approach to explain the outcomes of the recommendation model and promote model explainability. Figure 4.4 depicts an overview of our explanation approach.

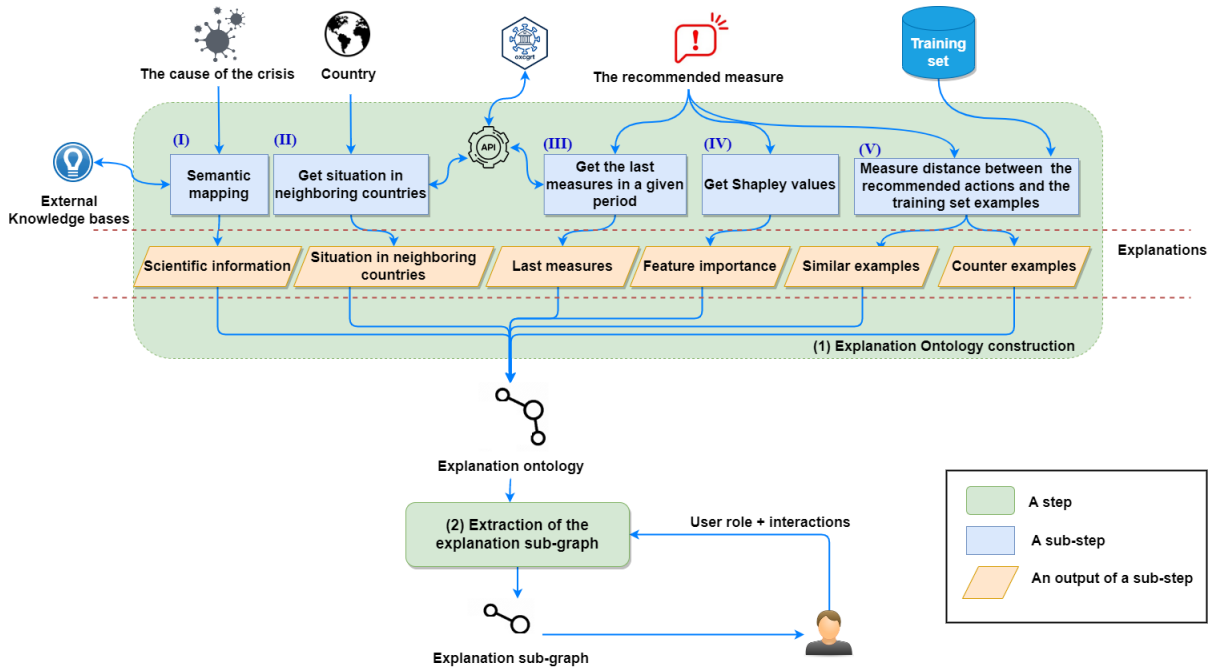


Figure 4.4: Overview of the explanation approach

We propose a two-step approach, namely (1) Ontology construction and (2) Extraction of the suitable explanation sub-graph. Specifically, the proposed approach uses semantic technologies to dynamically construct an explanation ontology using mapping with external semantic graphs, data sources, and frameworks. Then, our approach extracts an explanation subgraph from the constructed explanation ontology according to the user's role and preferences. To do so, we keep the explanation process transparent for the user by implicitly inferring his preferences via analyzing his interactions with the model and, thus, adapting the proposed explanations according to his needs. In what follows, we detail the adopted reasoning to design each step.

4.4.1 Explanation ontology construction step

Semantic technologies have long proven their efficiency and performance in illustrating knowledge in several fields. Therefore, a diverse range of ontologies was proposed, which may help to provide various possible explanations for the end-users [114].

Thus, we rely on semantic technologies to design our approach to explain the recommendation model. We consider the needs and roles of multiple users and their decision contexts to offer them adapted explanations. Thus, we propose different explanation types such as explanation by scientific information and statistics (I), neighboring countries (II), last measures adopted during a given period (III), feature importance (IV), similar countries, and counter-examples (V), as depicted in Figure 4.4. For instance, in a crisis, a scientist may be interested in scientific information, while a strategic expert may want to

know statistics and the situation in neighboring countries. In addition, both users may be interested in similar and counter-examples that help understand the reasoning behind the recommendation model and investigate the recommended measures' effectiveness.

In particular, (I) the constructed ontology groups scientific information related to the concerned domain. Algorithm 3 depicts the steps proposed to construct the explanation ontology, while Figure 4.5 presents the classes that constitute the core of our explanation ontology, which is built dynamically through mapping and importation from external knowledge graphs, data sources, and frameworks (see Algorithms 4-9).

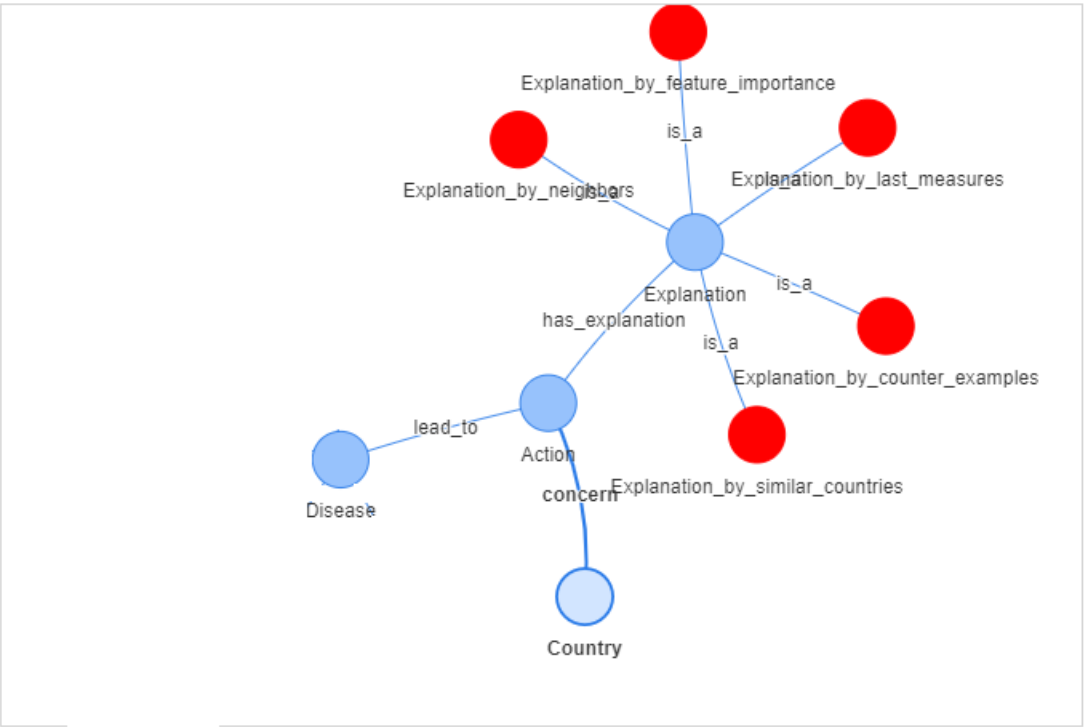


Figure 4.5: Core classes of the constructed explanation ontology

We represent each explanation type with a red-colored class. These classes will be linked and instantiated to constitute explanations. Hence, we seek in the external domain ontology those concepts that contain the name of the cause of the crisis. Also, we import from the domain ontology more information about synonyms and links to external scientific and government libraries. Moreover, our explanation approach analyzes the situation in neighboring countries (II) and proposes it as an explanation for users. To do so, we use a database that contains each country in the world with its borders. The explanation approach then gets information about the situation in neighboring countries. On the other hand, our explanation approach employs internal elements used during model training, like the training set and model features, to constitute other explanations. For instance, it involves the training set (i.e., the dataset used to train the recommendation model) as a dataset containing different information that could constitute examples for

Algorithm 3 Algorithm for explanation ontology dynamic construction

Require: Training set, The recommended measures, country
Ensure: Explanation Ontology
{/* Creation of the core classes */}
Explanation Ontology = New empty Ontology
Explanation Ontology.addClass(Action)
Explanation Ontology.addClass(Disease)
Explanation Ontology.addClass(Explanation)
Explanation Ontology.addClass(Country)
Explanation Ontology.addIndividual(Action, The recommended measures)
Explanation Ontology.addIndividual(Country, country)
{/* Creation of each explanation */}
Create Explanation by scientific Information(The cause of the crisis)
Create Explanation by counter examples(Training set, The recommended measures, country)
Create Explanation by similar examples(Training set, The recommended measures, country)
Create Explanation by feature importance(Training set)
Create Explanation by situation in neighboring countries(The recommended measures, country)
Create Explanation by the last measures(period, country)
Link classes via objectProperties

Algorithm 4 Algorithm for scientific information construction

Require: The cause of the crisis
Ensure: Scientific explanations
Search for classes that are related to the cause of the crisis in external ontologies/-knowledge graphs
Import the identified classes
Import description, synonyms and links to external scientific and governmental libraries as instances of the created class

Algorithm 5 Algorithm for explanation by counter-examples

Require: Training set, The recommended measures, country
Ensure: Counter-examples explanations

for all Training Set records **do**
    Get the records that adopted different measures than the recommended ones
    Extract the records referencing countries that are facing a close crisis severity by measuring the cosine distance using the reproduction rate and positivity rate
    Filter the records by identifying the closest countries by measuring the cosine distance using the Human Development Index, median age and Life expectancy
end for
Create "Explanation by counter examples" class as subclass of Explanation
for all Extracted records **do**
    Create a new class that represents an example as subclass of "Explanation by counter examples"
end for

Algorithm 6 Algorithm for explanation by similar examples

Require: Training set, The recommended measures, country
Ensure: Similar examples explanations

for all Training Set records **do**
    Get the records which adopted the same measures as the recommended ones
    Extract the records referencing countries that are facing a close crisis severity by measuring the cosine distance using the reproduction rate and positivity rate
    Filter the records by identifying the closest countries by measuring the cosine distance using the Human Development Index, median age, and Life expectancy
end for
Create "Explanation by similar examples" class as subclass of Explanation
for all Extracted records **do**
    Create a new class that represents an example as subclass of "Explanation by similar examples"
end for

Algorithm 7 Algorithm for explanation by situation in neighboring countries

Require: The recommended measures, country
Ensure: Measures adopted in neighboring countries

Get the neighboring countries of the concerned country
for all Neighboring country **do**
    Get the actual measures
    for all Neighboring country **do**
        Create a new class as subclass of explanation by neighboring countries
        Create an instance of the created subclass containing the set of measures
    end for
end for

Algorithm 8 Algorithm for explanation by last measures

Require: Period, country
Ensure: Measures adopted in a given period of time
1: **for all** Date in the period **do**
2: Get the measure adopted in that date
3: Create a new class for the measure as a subclass of "Explanation by last measures"
4: **end for**

Algorithm 9 Algorithm for explanation by feature importance

Require: Training set
Ensure: Feature importance
1: **for all** Features of the recommendation models **do**
2: Measure its contribution in the recommendation of measures using SHAP method
3: **end for**
4: Create a subclass of "Explanation by feature importance"
5: Create an instance of the created class, containing the shapely values of each feature

explanations. The training set contains the history of each country's measures, as well as (III) practical contextual information. Indeed, we measure the distance between the recommended measures and the training set records using cosine similarity to constitute explanations by similar examples and counter-examples. Thus, we extract the data records related to countries that applied the same recommended measures (V). We also measure the cosine distance using contextual information (e.g., the Human Development Index) to filter the dataset records. The filtering process aims to extract the most similar countries with the concerned country regarding contextual information (i.e., having similar characteristics). Then, the records taken out will be proposed as similar examples (V), while the others (i.e., that were not taken out) will be proposed as counter-examples (V), and we present them as subclasses of the "Explanation_by_similar_countries" and "Explanation_by_counter_examples". In addition to the explanations presented above, our explanation approach provides the feature's importance as an explanation to investigate each feature's impact on decision generation. Accordingly, a user could check whether the recommendation model is biased and may understand the reasoning behind the recommendation. To do so, we employ SHAP (SHapley Additive exPlanations), a game theory approach that aims to explain the output of machine learning models via local explanations of Shapely values [115]. Then, the Shapley values of each feature are proposed as an explanation for the end user. As we intend to consider the healthcare field as the application domain, we illustrate the constitution of explanations via our approach in this field. For instance, we apply our approach to constitute explanation ontology that includes (I) scientific explanations, particularly medical information related to the predicted or closest disease (e.g., in the case of identification of an unknown disease like COVID-19 in 2020), such as symptoms, treatments, synonyms, and links to external medical and government libraries. To do so, it performs mapping with external knowledge graphs, such as Human Disease Ontology in healthcare. The latter is an online open-access semantic

database with specific formal semantic rules to express meaningful disease models and multiple-inferred mechanistic disease classifications [116]. Then, we present the imported classes as subclasses of the "Disease" class, for example, in the explanation ontology. Figure 4.6 depicts and example of explanation via an ontology constructed after mappings with Human Disease Ontology (HDO) to explain COVID-19 diseases. We also present in Figure 4.7 an example of an explanation for a severe COVID-19 infection. Follow-

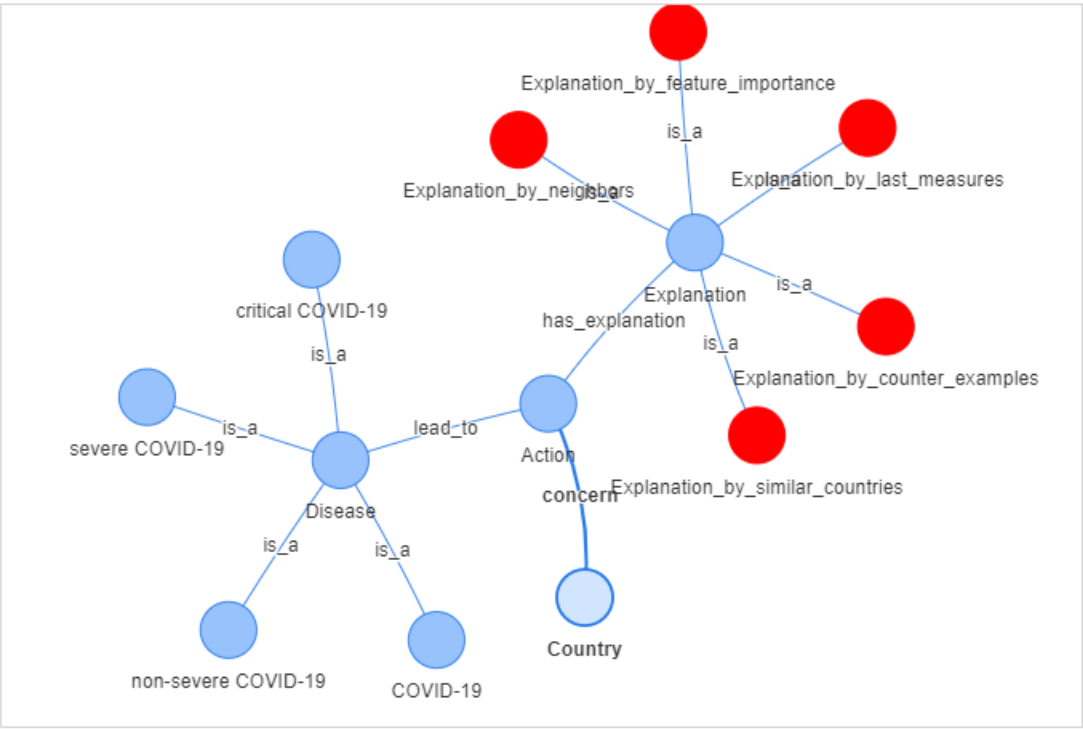


Figure 4.6: Example of explanation ontology after mapping with the COVID-19 disease classes from HDO

ing the constitution with scientific information, our approach queries OxCGRT platform to analyze the situation in (II) neighboring countries. The explanations for each country are presented as individuals of new classes to which each neighboring country belongs. Then, our approach queries OxCGRT platform to get (III) the last taken measures during a period defined by the user and creates subclasses of the last measures class in the explanation ontology. Similarly, it fetches in training set for (V) examples following the reasoning presented above and presents them as classes in the explanation ontology. Then, it applies the shapely method to constitute an explanation using (IV) feature importance. Figure 4.8 depicts an explanation via SHAP illustrating the contribution of features in the final recommendation.

4.4.2 Explanation subgraph extraction

Following the dynamic construction of the explanation ontology, the next step extracts and proposes the convenient subgraph for explanation according to user preferences. As

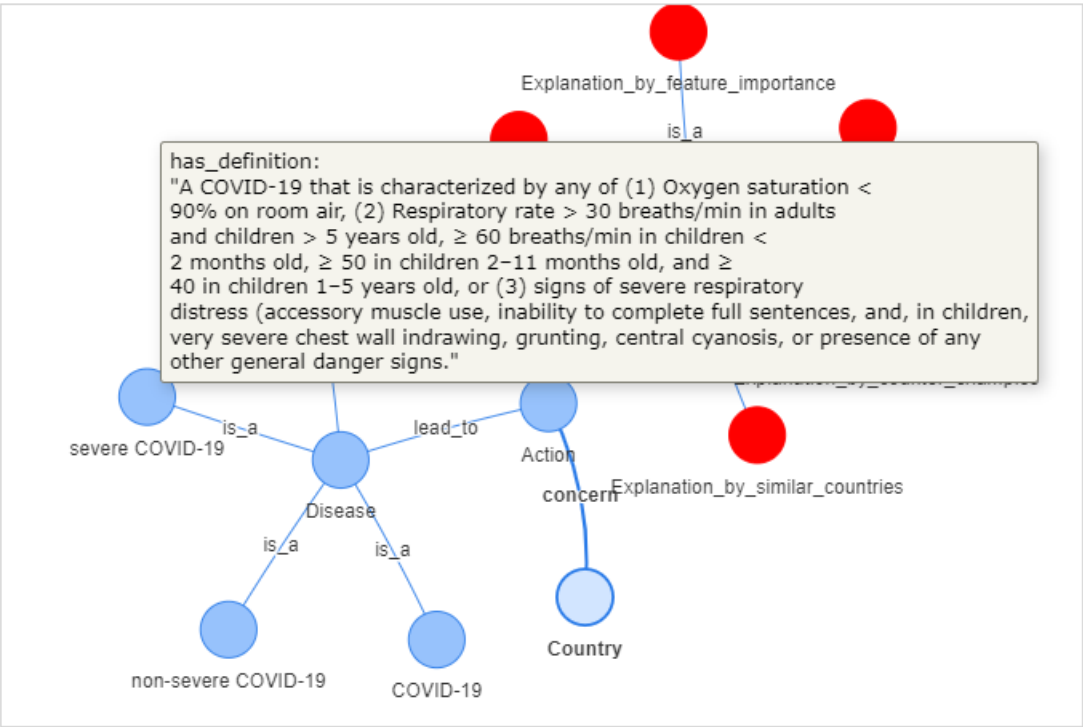


Figure 4.7: Example of explanation for severe COVID-19 case

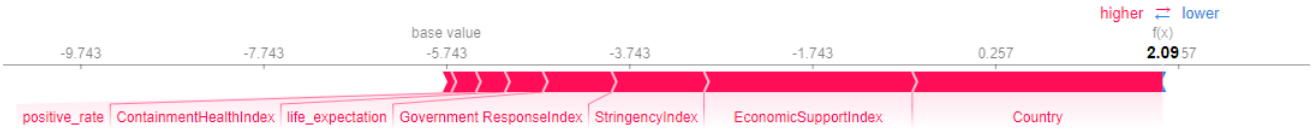


Figure 4.8: Example of feature importance explanation using SHAPely values

aforementioned, end users may have different interests in explanation types according to their roles and needs. Hence, we consider user preferences, roles, and decision context during the process of subgraph extraction from the explanation ontology. To do so, our explanation approach relies on the matrix-factorization method, one of the most efficient methods used in recommendation systems and adopted by online services like Netflix to speed up the search for user content recommendations [117]. This way, we implicitly infer user preferences by analyzing their interactions with the recommendation system. Then, we infer their preferences and propose a convenient explanation subgraph. As depicted in Figure 4.10, a user can hide or show an explanation that may give us information about his interests.

Moreover, we rely on user interactions to deduce user satisfaction regarding the proposed explanations. For instance, if the user chooses to hide the proposed explanations, our approach can infer that the user was probably not satisfied with them. Likewise, if she/he selected to explore other explanations not suggested by the explanation approach, we can

reason that our approach had probably not proposed the desired explanations. Thus, our explanation approach adapts the proposed explanations according to user needs and suggests alternative explanations.

For this purpose, our proposed approach creates a user-explanation matrix in which we saved the captured interactions. Then, it relies on the matrix factorization to map users and explanations to a joint latent factor space of dimensionality. The user-explanation interactions in this space are modeled as inner products as described in [118]. In other words, we defined two matrices, Q and P, that stand for users and explanations, respectively. Each matrix contains features identified via matrix features algorithms (e.g., Singular Value Decomposition (SVD), which is a factorization method that divides a matrix into three separate matrices). Similarly, it includes the ratings between the users, the explanations, and the features. Subsequently, we derived the user interest in an explanation by multiplying the Q and P matrices. Figure 4.9 illustrates the process of inferring user preferences regarding explanations using Matrix Factorization. As we considered different decision contexts in our approach, we defined multiple matrices of preferences devoted to each decision context. Hence, we inferred user preferences in different contexts to recommend distinct subgraphs for each user according to their decision context. For example, a user may prefer an explanation using the situation in neighboring countries during an ordinary situation. Nevertheless, during a crisis, they would like an explanation by counter-examples. Accordingly, our approach loads the matrix which corresponds to the context (i.e., ordinary situation) and applies the matrix factorization method, that multiplies the rows of the user matrix, related to the concerned user by the columns of the explanations matrix, related to the neighboring countries' explanation to infer the interest of the user in such explanation.
Illustration: 1^{st} row of the user matrix (0.5, 0.7, 0.8) X 1^{st} column of the explanations matrix = Inferred interest of the 1st user (1.17). After presenting our proposed explanation approach and detailing each step, we focus on the following section on the implementation and the experiments conducted to assess the effectiveness of our proposal.

4.5 Implementation and experimental results

In this section, we present and detail the implementation and the conducted experiments that focus on the deep learning-based recommendation model and the explanation approach. First, we introduce the experimental settings adopted to design and assess our contributions. Then, we present each conducted experiment and discuss the main results.

4.5.1 Implementation

We implemented the proposed contributions by performing three steps. First, we present the implementation of a multi-agent system that predicts the abnormal changes that may lead to the occurrence of crises. After identifying the risk, we employ the corresponding recommendation model we designed using deep learning, as described earlier. Then, we

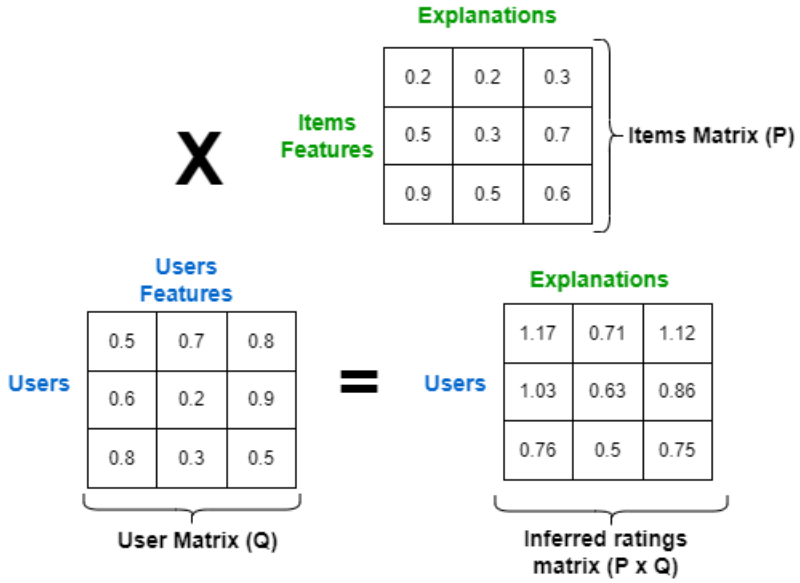


Figure 4.9: Illustration of user preference inference regarding explanations using Matrix Factorization

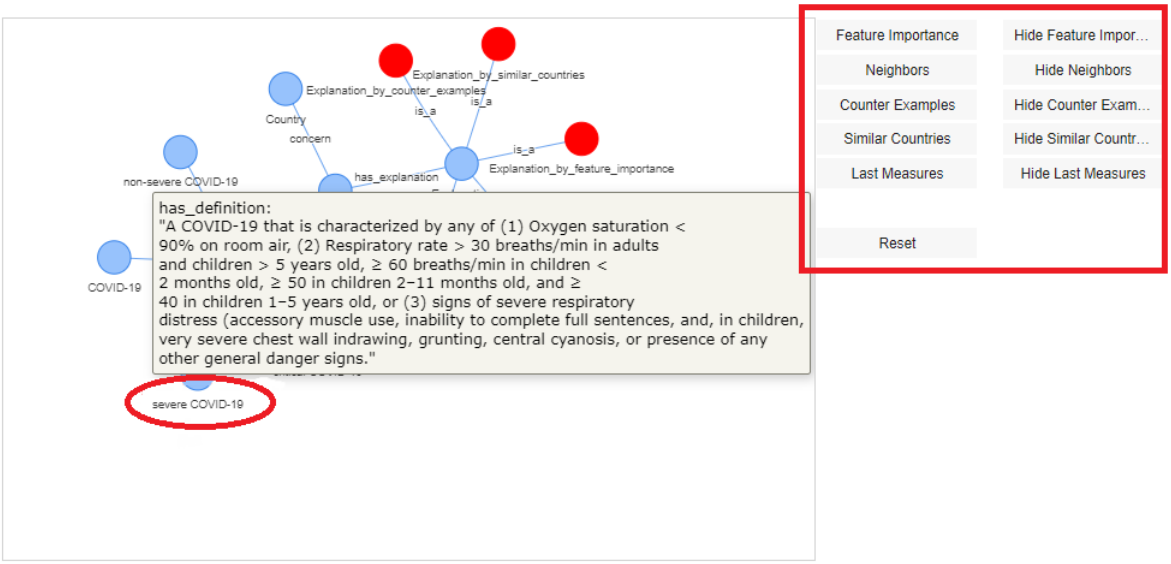


Figure 4.10: Interaction panel used to show/hide explanations

present the implementation of our proposed explanation approach, which explains our recommendation model choices.

In the present thesis, the focus is on data curation and the explainable recommendation of measures. Nevertheless, we implemented the aforementioned multi-agent system for the proof of concept and to link the data curation and the health measure recommendation steps, as depicted in Figure 4.11. The prediction step is performed via the analysis of weak

signals that could be early indicators of the occurrence of a crisis.

According to [119], disasters could be viewed as organized events requiring prolonged neglect of warning signs and signals of danger. Indeed, when risk signals are not noticed or misunderstood by organizations, safeguards, and agents, defenses against those risks become less efficient. Thus, identifying weak signals and interpreting early warnings is considered one of safety management's most challenging and critical aspects. Hence, the non-identification or misunderstanding of the warning signs can gradually accumulate and enlarge the disaster. However, this long and incremental process of disaster incubation gives us considerable time to detect a disaster before it becomes a significant care failure.

For this purpose, we designed a multi-agent system that identified and analyzed weak signals to predict the risk of crises and applied it to predict health crises. This system treats different signals, such as social networks (e.g., Twitter), mobility, travel, meteorological data, and search engine trends. These signals may provide a partial solution to the global health outbreak prediction problem. For this reason, we present a multi-agent system in which each agent resolves the problem from its own point of view as a convenient solution to overcome the prediction problem. As we aim to predict the crisis by analyzing various signal types (e.g., social media, travel data, etc.), we tackle the prediction problem by adopting the "divide and conquer" technique, which divides a problem into different sub-problems to be resolved individually. Hence, we found that a multi-agent system (MAS) could solve this problem by creating several agents devoted to resolving a sub-problem. As we presented in Chapter 2, an agent is characterized by several characteristics, such as independence, self-awareness, and autonomy, and no agent has a complete global view of the system or such knowledge to resolve a problem. These characteristics have encouraged us to adopt multi-agent systems as they fit the requirements of our context.

Moreover, we could take advantage of the social aspect of the MAS to resolve each sub-problem independently and then exchange partial solutions. Specifically, we designed a MAS consisting of different autonomous agents communicating to ensure different tasks, such as social network, search engine trends, mobility, travel, and meteorological data analysis agents. Our system also includes an agent that identifies diseases and their related information (e.g., symptoms) and another one that analyzes the outputs of all the agents of the system to make the final decision. Figure 4.12 depicts the designed multi-agent system for health outbreak prediction. We devoted the social network, mobility, and search engine trends agents to monitoring the changes in the environment. For instance, the social network agent surveys the trending words and hashtags (e.g., "crisis", "#COVID-19", "disease", "virus", etc.) on Twitter and also analyzes the possible changes in the users' sentiments. Similarly, the search engine trend agent monitors the trending searches on search engines like Google.This agent could interact with the disease identification agent, which employs a knowledge base including information about several infectious diseases to communicate and analyze whether the trending terms are related to a disease. In this case, the trend monitoring agent may identify a potential risk of infectious disease. The mobility agent, in turn, monitors the sudden changes in drug and grocery sales rates, which

may also indicate the occurrence of a disease (e.g., the sudden augmentation of sales of paracetamol and vitamin C during the COVID-19 outbreak). The mobility agent also surveys the movement of people to the park and the use of public transit. Clearly, the massive gathering of people in the same place can facilitate the transmission of diseases.

Nevertheless, our system ensures more than monitoring the movement of people at the national level by monitoring the intensity of exchanges in airports via a dedicated travel agent. Besides, no one can deny that weather conditions could impact the occurrence of crises (e.g., the transmission of infectious diseases, natural disasters). Therefore, we have defined the weather agent, which monitors the weather conditions to identify whether the environment could be favorable for the occurrence of crises. Hence, each of the presented agents can decide whether the situation is risky from its point of view. Then, all the agents' decisions are transmitted to the decision agent, which makes the final decision by employing fuzzy logic.

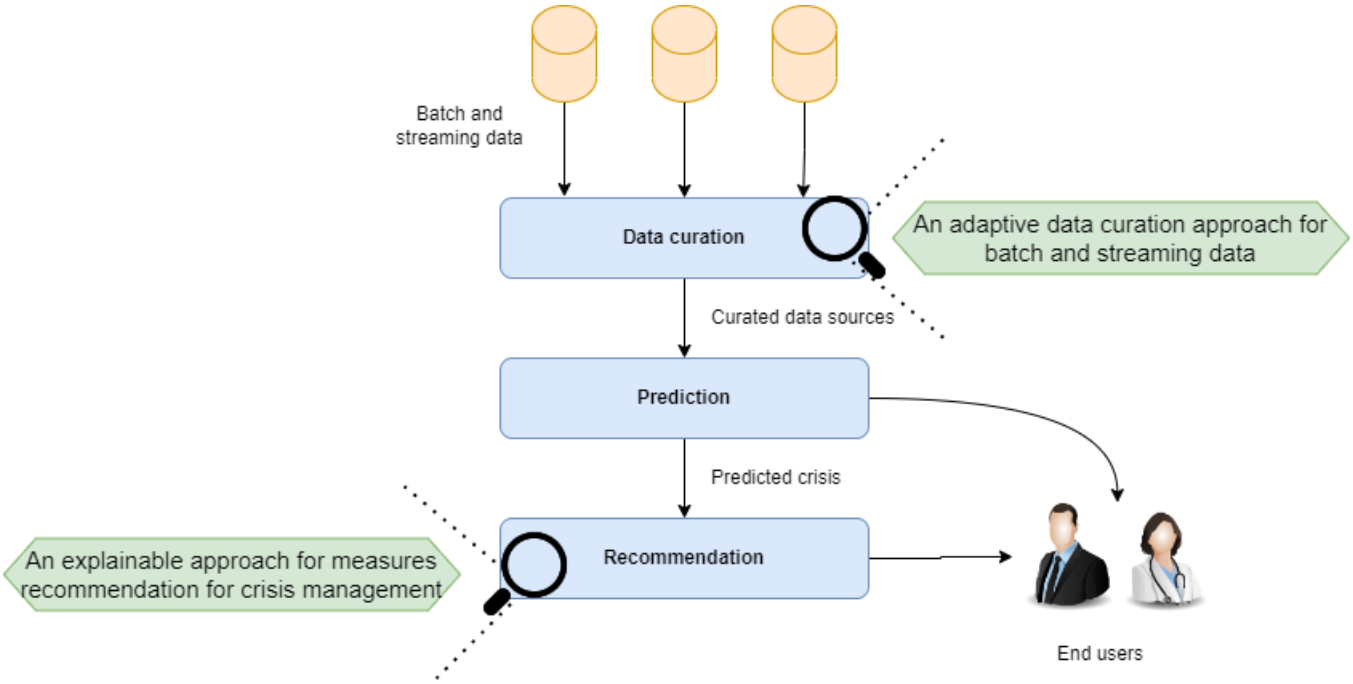


Figure 4.11: Summary of the contributions at each level (presented in green)

After implementing the multi-agent system for disease prediction, we designed the explainable recommendation step according to our proposed approaches. We used Python 3 and Tensorflow/Keras, an API, to design deep neural networks. Also, we relied on this API to design a multi-output deep learning model to recommend measures to manage the crisis. Then, we implemented the explanation approach by coding the presented algorithms. The implemented model and the explanation approach will be involved later to assess the performance of our proposals according to an experimental protocol that we detail in the

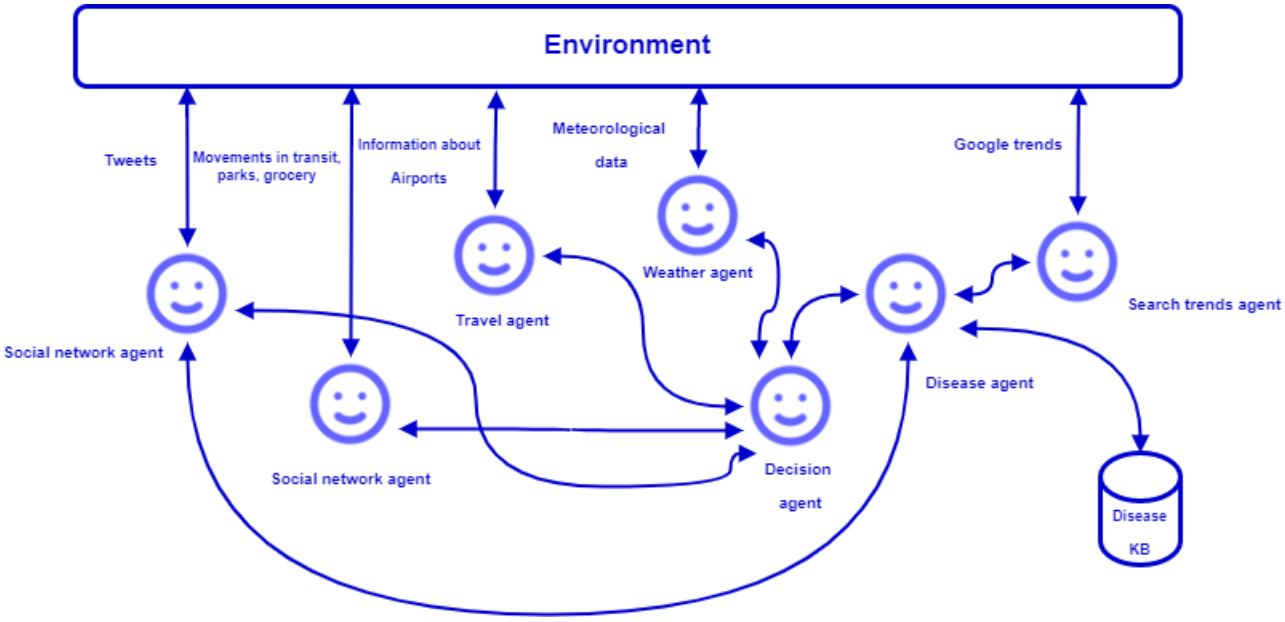


Figure 4.12: The designed multi-agent system for abnormal changes prediction

following section.

4.5.2 Experimental settings

We present in this section the experimental protocol that we adopted to evaluate the effectiveness of our contributions. To do so, we first assessed the performance of the recommendation model using dedicated performance metrics, namely recall, precision, accuracy, and F1-Score. Specifically, we trained and validated the deep learning model using the OxCGRT dataset, which encompasses the strategic measures taken by governments to combat the COVID-19 outbreak. We considered the different strategic measures, such as school closures, travel restrictions, and economic policies, collected from January 1, 2020, to January 31, 2022, and covered more than 180 countries.

Following the experimental protocol, we concentrate in what follows on the evaluation of the recommendation model and the explanation approach by assessing (1) the quality of the constituted explanation ontology, (2) the extracted sub-graph, and (3) the effectiveness of the recommendation mechanism.

4.5.3 Recommendation model performance

We firstly initiate the experimental protocol by measuring the evaluation metrics to assess the overall performance of the recommendation model, namely Recall (95%), Precision (96%), F1-Score (95%), and Overall Accuracy (95%). Thus, the recommendation model showed good performance in terms of recommendation as the performance measures exceed 95%. Then, we measured the accuracy of the recommendation of each output layer.

As illustrated in Figure 4.13, the proposed health measure recommendation model has ensured good accuracy, which varies between 91% and 99%.

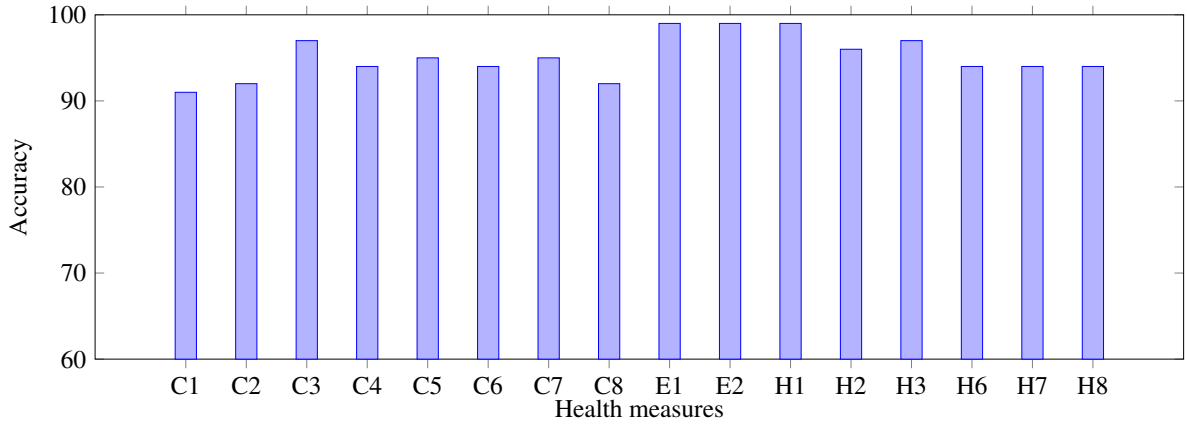


Figure 4.13: Accuracy of recommendation of each recommended measure

Figures 4.14-4.16 depict examples of the training/validation accuracy curves of some measures of each category. These figures reflect the proper functioning of the training and validation processes. Indeed, the combination of the training loss and validation loss metrics, as well as the accuracy, is one of the most widely used metrics to assess the effectiveness of the training process. Monitoring these two curves helps identify model overfitting or underfitting. Overfitting is one of the fundamental issues in machine learning that prevents generalizing a model to fit well observed data on training data, new data, and unseen data on the testing set due to several reasons, including noise and the size of the training set [120]. Underfitting is the opposite of overfitting, which occurs when the model cannot capture the data's variability [121]. Thus, if training loss is considerable compared to validation loss, we can identify an underfitting, else overfitting. As depicted in Figures 4.14 - 4.16, the accuracy curves in the training and validation processes are very close, which means that the proposed model is not under/overfitted. On the other hand, the curves presented in Figures 4.17 - 4.18 depict the effectiveness of the training/validation processes and the low loss value of less than 0.4. The loss is a metric that indicates the penalty due to a wrong prediction for a single example. Hence, the slopes of the presented curves show that the loss value quickly (i.e., from the 15th epoch) converges to approximately 0. Following these experiments and analyzing the generated performance curves, we deduce that our proposed model has performed well in recommending health measures.

4.5.4 Explanation approach performance

As described in the experimental protocol, the second step evaluates our proposed explanation approach's performance. We detail hereafter the experiments related to (1) the explanation ontology, (2) the subgraph quality, and (3) the evaluation of the recommendation mechanism.

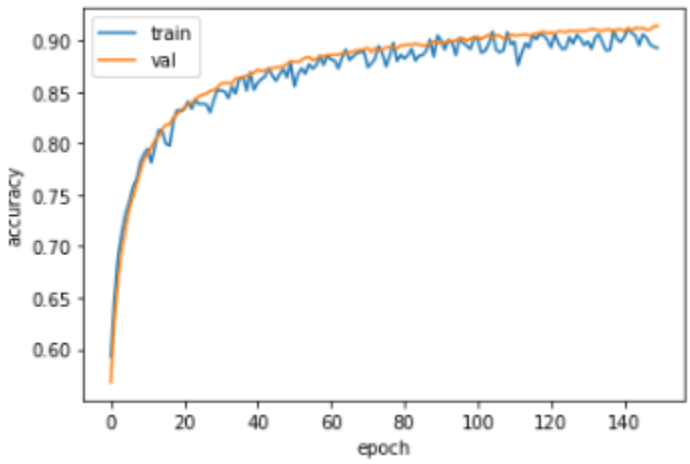


Figure 4.14: Training/validation accuracy curve of "school closures" measure

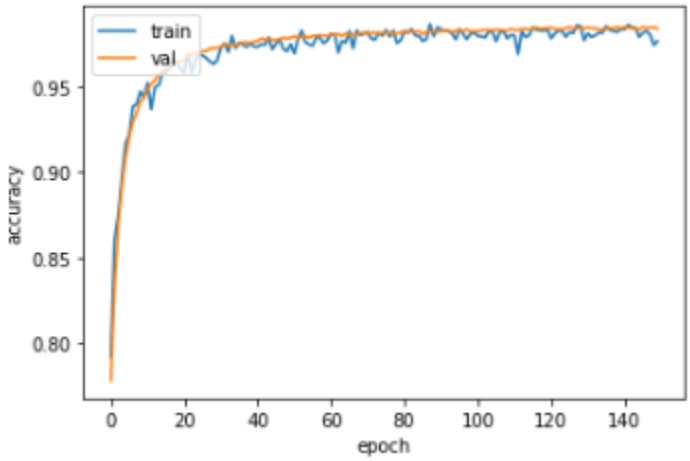


Figure 4.15: Training/validation accuracy curve of "income support" measure

Explanation ontology evaluation

We used ontology assessment metrics to assess explanation ontology quality. These metrics measure the ontology's structure (i.e., schema or graph) and the knowledge it encompasses. They are concerned with different quality evaluation perspectives, such as functional, analytical, pragmatic, syntactic, cognitive, semantic, social, and practical quality perspectives. For instance, structural characteristics of the ontology (e.g., size) may affect the process of ontology merging, alignment, and reuse. Hence, schema and graph evaluation metrics provide an overview of the structural quality of the ontology [95]. Similarly, knowledge base metrics assess the quality of an ontology in terms of its richness and the knowledge provided. We present the evaluation results in Table 4.2-4.4.

As depicted in Table 4.2, the structure of the explanation ontology is rich in inheritance

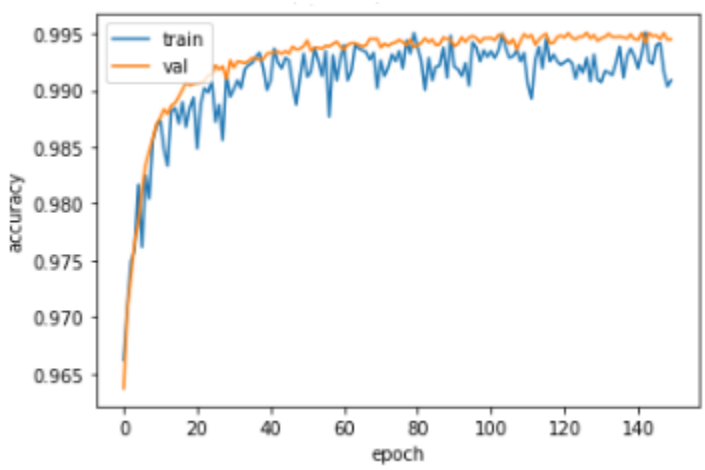


Figure 4.16: Training/validation accuracy curve of "Testing policy" measure

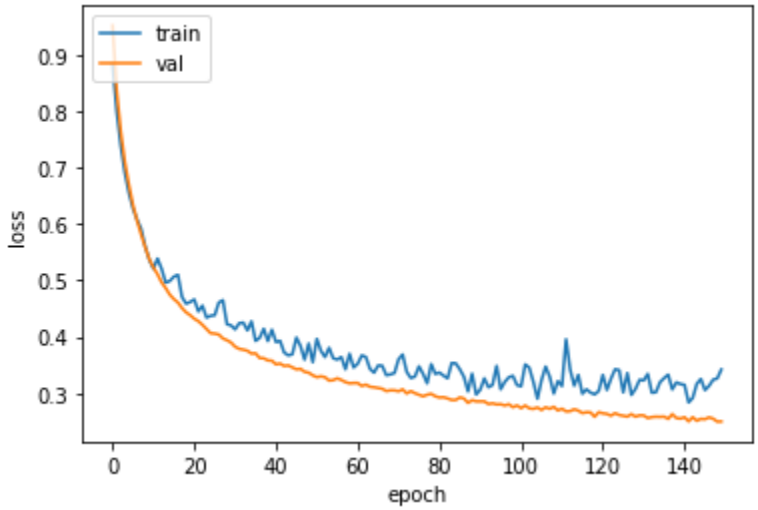


Figure 4.17: Loss curve of "school closures" measure

Table 4.2: Schema evaluation metrics

Inheritance richness	0.83
Relationship richness	0.16
Attribute Class Ratio	0.70
Axiom/Class Ratio	68.20
Class/Relation Ratio	1

and attributes, which have a ratio close to 1. The relationship richness is low because most of the relations in the constructed explanation ontology are inheritance relations (i.e., is-a). Moreover, the constructed ontology is balanced in terms of the distribution of attributes,

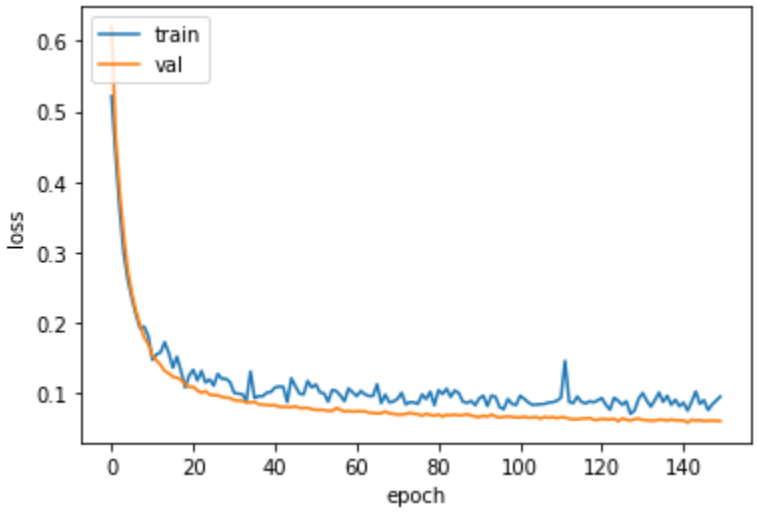


Figure 4.18: Loss curve of "income support' measure

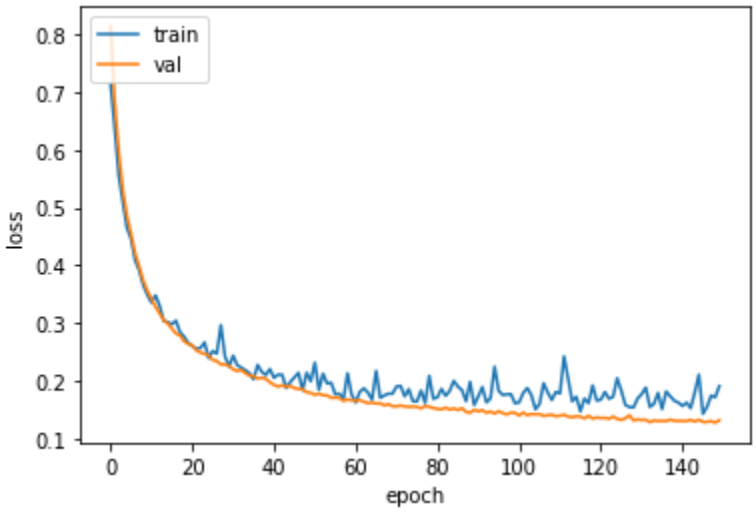


Figure 4.19: Loss curve of "testing policy" measure

Table 4.3: Knowledge metrics

Average population	4
Class richness	0.75

classes, and axioms, as the attribute, inheritance, and axiom/class ratio values are equal or close to 1. Hence, these results prove efficient knowledge presentation and the ease of merging with other ontologies and reusing.
Furthermore, the metrics presented in Table 4.4 depict the richness of our ontology and its simplicity. For instance, the maximal and average depth and breadth values show that the generated ontology is not complex. Nevertheless, it encompasses various classes de-

Table 4.4: Graph evaluation metrics

Absolute root node	1
Absolute leaf cardinality	19
Absolute sibling cardinality	19
Average depth	9
Maximal depth	4
Absolute breadth	3
Maximal breadth	4
Ratio of sibling fanoutness	0.79

scribing an entity from different knowledge dimensions. The knowledge metrics presented in Table 4.3 depict the high knowledge quality of the explanation ontology. Indeed, the explanation ontology is rich semantically since it contains high values of the average population (i.e., instance richness) and class richness ratio (i.e., 75%). Moreover, it provides a wide range of concepts that have an average depth of 9 concepts and 19 siblings.

Sub-graph evaluation

We evaluated the quality of the proposed explanation types by investigating the extracted subgraphs, much like we did for the overall explanation ontology. To do so, we assessed the quality of each recommended explanation subgraph via the same evaluation metrics as the explanation ontology (i.e., schema, knowledge, and graph evaluation). The results depicted in Table 4.5 prove the semantic richness of the explanation subgraphs, which are shown by the numbers and the distribution of attributes, classes, and relationships.
By analyzing the subgraphs' performance metrics, we found that feature importance and neighboring countries are the richest explanations in attributes and population, despite poor inheritance relationships.
The evaluation metrics also depict the structural balance of the explanation subgraphs. Indeed, the values of absolute siblings, leaf cardinality, and breadth show the horizontal and vertical balance of the subgraphs.
Moreover, the maximal depth of all subgraphs shows their semantic richness and the low complexity of the explanation content. Thus, richness metrics (i.e., attributes, relationships, and classes) depict the high cognitive and semantic quality of the recommended sub-graphs and their high practical quality (i.e., complexity of conceptualization and ease of use), which the moderate depth and breadth values can prove.

Recommendation evaluation

We assessed the performance of the explanation sub-graph recommendation step by investigating different recommendation algorithms. To be more precise, we assessed the performance of Matrix Factorization, KNN-based, and algorithms such as Slope One, Random

Table 4.5: Explanation sub-graph evaluation

Category	Metric	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

Recommendation, and Co-Clustering. These methods have recently become popular by combining good scalability with predictive accuracy. In addition, they offer much flexibility for modeling various real-life situations. As we described in the previous chapters, the basic idea of matrix factorization is to infer ratings from items to represent items and users using vectors of factors. In our context, we rely on the following Matrix Factorization[122] algorithms:

- SVD is a matrix decomposition technique that reduces the item-user rating matrix into two lower-dimensional matrices. Specifically, SVD reduces the dimensionality space of the problem from N to k, where k is less than N.
- Much like SVD, SVD++ reduces the user-item matrix into low-dimensional matrices. SVD++ optimizes the SVD algorithm by employing a biased regularized strategy to factorize the matrix into three low-rank matrices using implicit feedback information. This technique aims to extract the algebraic features, analyze the score of each factor, and then predict the results based on the analysis of the optimized data.
- Non-negative matrix factorization (NMF or NNMF) adopts the same reasoning as Matrix Factorization with the property that all the matrices have no negative elements to make the resulting matrices easier to inspect.

On the other hand, K-Nearest Neighbors (KNN) comprises a family of algorithms usually used to perform clustering via machine learning. Considering the recommendation problem, the K value specifies the number of nearest neighbors to be used in finding the missing ratings.
The k Nearest Neighbor approach encompasses two types, namely structure-based KNN and structure-less KNN. The structure-based method relies on the basic structure of the data, while the structure-less mechanism handles training data samples. Accordingly, the K nearest neighbors are the K samples having the minimum distance. In the present work, we consider the following KNN algorithms:

- "KNN Basic" is the basic KNN algorithm that predicts the ratings by measuring the distance between samples to identify the K nearest neighbors and, thus, to predict the ratings by performing majority voting.

- "KNN Baseline" relies on the baseline factor, which is a technique helping the learning models identify the functional relationships between the inputs (i.e., features) and the desired outputs (i.e., labels). Then, the KNN Baseline algorithm predicts the rating using a baseline rating.
- "KNN With Means" predicts the ratings by searching the nearest data points and using the arithmetic mean of the ratings of the K Nearest neighbors.
- "KNN With Z-Score" searches the nearest k data points, as described above, by considering the z-score normalization of each user.

While Slope One [123] is an easy-to-implement and accurate recommendation model, Co-clustering is used in cluster analysis to deal with high-dimensional and sparse data efficiently. Indeed, it ensures synchronous clustering of the columns and rows of a matrix.
The performance of the presented algorithms is assessed against synthetic health data derived from benchmark datasets widely used for evaluating recommendation system performance and to propose five types of explanations (i.e., explanation by similar examples, counter-examples, feature importance, neighboring countries, and last measures).
Specifically, we rely on measures such as precision, recall, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Indeed, Precision measures the relevance of the recommended items, while Recall assesses the proportion of the relevant items found in the recommendations.

On the other hand, RMSE and MAE measure the average magnitude of the error. As depicted in Table 4.6, SVD and SVD++ algorithms showed the best performance using the two datasets. KNN-based algorithms have also ensured good performance using the two datasets. Besides, the Non-Negative Matrix Factorization, the SlopeOne, and the Random algorithms showed good performance using the first dataset but were less performant with the second one. However, CoClustering is the least performant algorithm. Hence, the results prove the effectiveness of the Matrix Factorization paradigm that we adopted to recommend explanations.

Table 4.6: Comparison of performance of the recommendation algorithms

		Synthetic Dataset 1				**Synthetic Dataset 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

4.6 Conclusion

4.6.1 Summary

In the present chapter, we introduced a new model that recommends measures for crisis management and also an explainable approach that explains the choices of the model for multiple user roles. To do so, we relied on deep learning to design a multi-output model in which each output layer predicts the stringency of each health measure. The recommendation model proposes the future stringency of each measure. Since we implemented our contributions in the health crisis management fields, we used our model to recommend measures related to containment and closure, economics, and health system policies. Yet we stress that deep learning algorithms constitute black-box models. Thus, we proposed a semantic-based approach that explains and argues for the choice of the recommendations generated by the proposed model. The originality of our contributions relies on several points, namely the diversity of the proposed measures and explanation types such as explanation by examples, counter-examples, neighboring countries, feature importance, and last measures. Our explanation approach has the merit of providing various explanations according to the user's needs. To do so, we rely on dynamically designing an explanation ontology and employ matrix factorization techniques to extract from the constructed ontology the adapted explanation sub-graph for the user.
Moreover, we conducted several experiments to evaluate the performance and prove the effectiveness of our proposal:

1. We assessed the performance of the recommendation model by measuring accuracy, precision, recall, and the F1-Score. The results showed the robustness of the proposed measures recommendation model.

2. We applied schema evaluation, knowledge, and graph-evaluation metrics to evaluate the expressiveness and complexity of the explanation ontology and the extracted sub-graphs. Hence, the results showed the semantic richness and low complexity of both explanation ontologies and sub-graphs.

3. We compared the performance of different recommendation algorithms, such as Matrix Factorization variants, KNN-based algorithms, SlopeOne, etc. The results showed that Matrix Factorization algorithms, namely SVD and SVD++, outperform the other algorithms.

4.6.2 Limitations & Enhancement ideas

Privacy-preserving measures recommendation

The proposed model recommends several measures directly related to sensitive dimensions like people's movements (e.g., traveling, transit), health, economics, etc. However, such data may reveal sensitive information and make the model vulnerable to privacy attacks like property inference attacks. On the other hand, we noticed that the explanations proposed for the end-user might be considered a "double-edged weapon" since they may

be employed to perform model extraction attacks. Accordingly, we plan to analyze the potential sources of privacy leaks in our recommendation model and explanation approach to consider the privacy-preserving aspect while generating recommendations and explanations.

Considering user feedback and experiences

Considering our proposed explanation approach, our proposal ensures adaptivity via user interaction analysis. Based on matrix factorization techniques, our approach adapts the provided explanations to better align with the user needs. Nevertheless, the explanation approach implicitly captures the potential user feedback and relies on inference. To overcome this limit, we intend to consider explicit user feedback and focus more on user feedback and experiences (UX) by automatically analyzing their interactions in order to provide them with recommendations better aligned with their needs.

CHAPTER 5
Conclusion and future work

CHAPTER 5

Conclusion and future work

> No matter how much experience you have, there's always something new you can learn and room for improvement.
>
> Roy T. Bennett

Humanity has witnessed several phenomena that have disrupted daily life and impacted the ecosystem, such as Hurricane Katrina in the United States, the global financial crisis, and recently the earthquake in Turkey. Worldwide, the last five years have been marked by the COVID-19 epidemic. This outbreak was a concrete example to the public of crisis management by the various stakeholder communities, such as health and strategy experts, the economy, etc. Crisis management, in general, requires overcoming the problem from several angles (strategic, economic, etc.) by analyzing massive and heterogeneous data from various sources, such as IoT, social networks, and sensors, collected in batch and streaming. In addition, the users involved in this process may have different requirements in various contexts regarding the different steps of the data management process (i.e., data curation, analysis, prediction, recommendation, etc.).

By deeply examining the health crisis and big data management contexts, we tackled in this thesis essential research questions about generating recommendations based on massive, heterogeneous, and complex data. Hence, we focused on several challenges related to the adaptive data curation challenge and explainable health measure recommendation, which resulted in three main contributions:

- A literature study about big data management in data lakes/data lakehouses.
- A framework for multi-structured batch and streaming data curation built upon an

ontology (i.e., for data characterization and quality assessment) guiding service composition using the ACUSEC approach.

- A multi-target deep learning-based recommendation model and a semantic-based XAI approach that explains the choices of the recommendation model for multiple user roles.

We validated these contributions in the context of healthcare data curation and measure recommendation.

5.1 Key Findings & Contributions

5.1.1 What are the challenges that address data management in data lakes/lakehouses?

Chapter 2 addresses this question by presenting a literature study of the existing works addressing data management steps in data lakes/lakehouses. More precisely, we tackled this problem by proposing a systematic mapping that categorizes and discusses the existing works related to these steps and the open issues. This study helped us identify the underlying data management steps that could be employed in recommendation systems. Thus, in this chapter, we present and review proposed works related to different data management steps, such as data curation, data quality evaluation, data analysis, and prediction. Following this study, we revealed the related challenges, including the need for adaptive data curation and explainable recommendations. Hence, we overcame the identified scientific challenges through the contributions discussed in the following sections.

5.1.2 How to manage multistructured data collected from diverse sources in batch and streaming?

Chapter 3 tackles this research question in detail by focusing on data curation. Specifically, we proposed, DARQAN, a modular ontology that assesses the data quality to judge whether the data source needs to be curated. In that case, this ontology identifies the main data characteristics that impact the data curation process. Thus, the identified characteristics are involved in ACUSEC, an approach we propose to compose curation services adaptively according to multi-user requirements in different decision contexts. ACUSEC relies on a library of curation services that we design to extract, enrich, standardize, and assess data quality. To do so, we employed reinforcement learning techniques to compose curation services adaptively according to the user's functional and non-functional requirements. We proposed implementing the ACUSEC approach into an adaptive curation framework for batch and streaming data. We validated the effectiveness of our proposal using an experimental protocol that focuses on assessing scalability, adaptivity to changes, and alignment with user needs. Contrary to existing works that perform static data curation, the merit of our solution lies in its ability to consider multi-structured data and to adapt to the needs of several user roles by considering different needs simultaneously.

5.1.3 How to generate recommendations for multi-users with different needs?

Chapter 4 answers this question by proposing a deep learning-based cross-country recommendation model that considers several user needs. We use multi-output deep learning techniques to recommend measures devoted to managing crises. The proposed model is explained via a recommendation approach that we propose to provide various explanation types adapted to different user roles, such as explanation by similar and counter-examples, neighboring countries, feature importance, and the last measures taken for a given period. In addition to the various explanations proposed by our approach, the originality of our work consists in adaptively proposing the explanation sub-graphs according to the user's needs. To do so, we employed matrix factorization techniques to adapt the provided recommendations. We validated our contributions via machine learning evaluation metrics like accuracy, recall, precision, and F1-score. We also employed quality evaluation metrics for ontologies, which evaluate quality along several dimensions: schema, knowledge, and graph. We used these metrics to assess the quality of the constructed ontology and the extracted sub-graphs. We also relied on other performance metrics like precision, recall, mean absolute error, and root means squared error to evaluate the performance of the sub-graph recommendation mechanism. Contrary to the previous works of XAI proposed in the literature, our solution has the merit of proposing different types of explanations by considering the needs of several user roles.

5.2 Limitation and future work

These contributions help remarkably in data management, particularly in performing data curation and measures recommendations and explanations. However, it is still possible to enhance the data curation and recommendation processes, and these potential improvements are detailed hereafter.

5.2.1 Limitation

Evolution of the performance of the data curation approach in the new data era

We have witnessed a quarter-century of digital transformation that has ushered in the data age, from the introduction of email to big data analytics, cloud storage, and SaaS, and now we are at the dawn of a new era. The latter is characterized by an explosion in the size and quantity of data and the services that process that data. We state that a library of curation services could not reach more than 12,000 services. Hence, our composition approach remains valid and efficient for data curation. Indeed, our service-based approach for data curation relies on Q-Learning, which has proven its effectiveness and efficiency compared to First-visit Monte Carlo and Temporal difference algorithms and other composition algorithms, as illustrated through the experiments. However, the new data era may have different requirements, necessitating that we enhance our curation approach by relaying on other machine learning algorithms that scale correctly with large inputs.

Evolution of the size of the explanation ontology

We proposed a multi-target deep learning-based model for measures recommendation that is explained via a semantic-based approach. Our approach constructs an explanation ontology dynamically by importing fragments from datasets, frameworks, and knowledge graphs. Yet, the ontology construction could become greedy in terms of imported classes which may raise the risk of the increasing size of the structural graph. Even though we showed the simplicity of the constructed explanation ontology, increasing the graph's size may increase the risk that the querying and the management of the ontology become more complex.

5.2.2 Possible future research venues

Investigate the performance of reinforcement learning in the new data era

We intend to investigate the performance of other reinforcement learning approaches to enhance the data curation service composition in the new data era. Indeed, we think that reinforcement learning remains a convenient solution for the curation service composition problem since we are dealing with unsupervised learning. Yet, the Q-learning algorithm may become less effective due to the multiplication of curation services that meet new curation needs. Hence, we plan to investigate the performance of other reinforcement learning like Deep Q-Learning and Multi-agent determinantal Q-learning. Actually, such methods may be less effective in our current context because they are greedy regarding training data (i.e., services in our case). However, we think they may enhance our contribution to services composition in the future.

Prediction of future crises

Chapter 4 discussed the limitations related to crises prediction, particularly, in the healthcare field. Indeed, we take advantage of multi-agent system techniques to constitute a prediction system that analyzes weak signals that may lead to the occurrence of a crisis. Indeed, in the present thesis, we focus on data curation and the recommendation of health measures. Hence, we proposed this multi-agent system to link these two steps. Yet, we think that crisis prediction could be an interesting research track and can help enhance our proposals' effectiveness. Accordingly, we aim to review crisis prediction deeper to analyze the existing approaches, techniques, and methodologies. Thus, we intend to propose a crisis prediction approach that considers the presented user and decision context requirements.

Investigate the explanation ontology evolution

We intend to investigate the application of ontology evolution approaches to optimize the explanation ontology constructed via our explanation approach. For this purpose, several techniques were proposed to evolve the ontology's schema and semantics, such as word

embeddings and machine learning. In addition, we plan to investigate the metrics that measure ontology evolution in terms of structural, relative, and sub-classes additions/changes. For this purpose, we think a literature study covering these aspects would be an interesting future research venue to better understand this concept. Following this study, we aim to identify the requirements of the ontology evolution aspect to consider it in our explanation approach and to enhance its evolution.

References

[1] Stephen Morse, Jonna Mazet, Mark Woolhouse, Colin Parrish, Dennis Carroll, W. Karesh, Carlos Zambrana-Torrelio, W Lipkin, and Peter Daszak. Prediction and prevention of the next pandemic zoonosis. *Lancet*, 380:1956–65, 12 2012.

[2] 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.

[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] Giri Kumar Tayi and Donald P. Ballou. Examining Data Quality. *Communications of the ACM*, 41(2):54–57, 1998.

[5] François Conesa and Francis 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.

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

[7] 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.

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

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

[10] Hela Taktak, Khouloud Boukadi, Firas Zouari, Chirine Ghedira, Michael Mrissa, and Faiez Gargouri. A knowledge-driven service composition framework for wildfire prediction. *Cluster Computing*, pages 1–20, 04 2023.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

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

[18] Barbara Hayes-Roth, Lee Brownston, and Robert van Gent. *Multiagent Collaboration in Directed Improvisation*, page 141–147. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[19] Katia P. Sycara. The many faces of agents. *AI Magazine*, 19(2):11, Jun. 1998.

[20] Eric Werner. *Cooperating Agents: A Unified Theory of Communication and Social Structure*. 01 1989.

[21] Philip Lord, Alison Macdonald, Liz Lyon, and David Giaretta. From data deluge to data curation. In *In Proc 3th UK e-Science All Hands Meeting*, pages 371–375, 2004.

[22] 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.

[23] Rolf Sint, Stephanie Stroka, Sebastian Schaffert, and Roland Ferstl. Combining unstructured, fully structured and semi-structured information in semantic wikis. 01 2009.

[24] Peter Buneman. Semistructured data. 1997.

[25] Michael J. Cafarella, Jayant Madhavan, and Alon Halevy. Web-scale extraction of structured data. *SIGMOD Rec.*, 37(4):55–61, mar 2009.

[26] 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.

[27] 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.

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

[29] 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.

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

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

[32] 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.

[33] 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.

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

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

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

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

[38] 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, 2017.

[39] K.H. Roberts, Peter Madsen, and V. Desai. Organizational sensemaking during crisis. pages 107–122, 01 2007.

[40] Jonathan Bundy, Michael Pfarrer, Cole Short, and Timothy Coombs. Crises and crisis management: Integration, interpretation, and research development. *Journal of Management*, 43, 12 2016.

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

[42] 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.

[43] 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.

[44] Toph Allen, Kris Murray, Carlos Zambrana-Torrelio, Stephen Morse, Carlo Rondinini, Moreno Di Marco, Nathan Breit, Kevin Olival, and Peter Daszak. Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, 8:1124, 10 2017.

[45] Nils Jonkmans, Valérie D'Acremont, and Antoine Flahault. Scoping future outbreaks: a scoping review on the outbreak prediction of the who blueprint list of priority diseases. *BMJ Global Health*, 6:e006623, 09 2021.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] Haiqian Chen, Leiyu Shi, Yuyao Zhang, Xiaohan Wang, and Gang Sun. A cross-country core strategy comparison in china, japan, singapore and south korea during the early covid-19 pandemic. *Globalization and Health*, 17, 02 2021.

[52] 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.

[53] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. Recommender systems: An introduction. *Recommender Systems: An Introduction*, 01 2010.

[54] Masatoshi Nei and W. H. Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76 10:5269–73, 1979.

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

[56] Paul Jaccard. Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 01 1901.

[57] Rémy Chaput, Amélie Cordier, and Alain Mille. Explanation for humans, for machines, for human-machine interactions? 2021.

[58] Zhenhua Yu, Taher Nofal, and Joao Tavares. Explainability of neural network clustering in interpreting the covid-19 emergency data. *Fractals*, 30, 11 2021.

[59] P. Jonathon Phillips, Carina Hahn, Peter Fontana, David Broniatowski, and Mark Przybocki. Four principles of explainable artificial intelligence. 08 2020.

[60] 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.

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

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

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

[64] 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.

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

[66] 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.

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

[68] 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.

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

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

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

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

[73] 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.

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

[75] 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.

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

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

[78] Carlo Batini and Monica Scannapieco. *Erratum to: Data and Information Quality: Dimensions, Principles and Techniques*, pages E1–E1. 05 2016.

[79] Matthias Jarke, Manfred Jeusfeld, Christoph Quix, and Panos Vassiliadis. Architecture and quality in data warehouses. pages 93–113, 07 1998.

[80] Diane Strong, Yang Lee, and Richard Wang. Data quality in context. *Communications of the ACM*, 40, 08 2002.

[81] Leo Pipino, Yang Lee, and Richard Wang. Data quality assessment. *Communications of the ACM*, 45, 07 2003.

[82] Felix Naumann. Quality-driven query answering for integrated information systems. *Lectures Notes in Computer Sciences*, 2261, 01 2002.

[83] Juan Carlos Augusto and Andrés Muñoz. User Preferences in Intelligent Environments. *Applied Artificial Intelligence*, 33(12):1069–1091, 2019.

[84] 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.

[85] 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.

[86] Jiapeng Wang and Yihong Dong. Measurement of text similarity: A survey. *Information*, 11(9), 2020.

[87] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.

[88] Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. Hamming distance metric learning. 25, 2012.

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

[90] Hongbing Wang, Xuan Zhou, Xiang Zhou, Weihong Liu, Wenya Li, and Athman Bouguettaya. Adaptive service composition based on reinforcement learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6470 LNCS(60673175):92–107, 2010.

[91] Anja Strunk. QoS-aware service composition: A survey. *Proceedings - 8th IEEE European Conference on Web Services, ECOWS 2010*, (1):67–74, 2010.

[92] Matthieu Lauras, Sébastien Truptil, and Frédérick Bénaben. Towards a better management of complex emergencies through crisis management meta-modelling. *Disasters*, 39(4):687–714, 2015.

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

[94] Evren Sirin and Bijan Parsia. Pellet: An owl dl reasoner. *Description Logics*, pages 212–213, 01 2004.

[95] Ravi Lourdusamy and Antony John. A review on metrics for ontology evaluation. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 1415–1421, 2018.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] Ayush Singhal, Pradeep Sinha, and Rakesh Pant. Use of deep learning in modern recommendation system: A summary of recent works. *International Journal of Computer Applications*, 180:17–22, 12 2017.

[102] Yuhao Niu, Lin Gu, Yitian Zhao, and Feng Lu. Explainable diabetic retinopathy detection and retinal image generation. 07 2021.

[103] Peter Nsubuga, Ben Masiira, Christine Kihembo, Jayne Byakika-Tusiime, Caroline Ryan, Miriam Nanyunja, Raoul Kamadjeu, and Ambrose Talisuna. Evaluation of the ebola virus disease preparedness and readiness program in uganda: 2018 to 2019. *Pan African Medical Journal*, 38, 02 2021.

[104] 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.

[105] Rongna Zhang, Zuoru Liang, Mingfan Pang, Xinping Yang, Jiewen Wu, Yuansheng Fang, Hanran Ji, and Xiaopeng Qi. Mobility trends and effects on the covid-19 epidemic — hong kong, china. *China CDC Weekly*, 3:159–161, 02 2021.

[106] Shinya Kumagai, Tomomi Aoyama, Eri Ino, and Kenji Watanabe. Oxcgrt-based evaluation of anti-covid-19 measures taken by japanese prefectures. *Journal of disaster research*, 16:16–23, 2021.

[107] Joshua Choma, Fábio Mathias Corrêa, Salah-Eddine Dahbi, Kentaro Hayashi, Benjamin Lieberman, Caroline Maslo, Bruce Mellado, Kgomotso Monnakgotla, Jacques Naudé, Xifeng Ruan, and Finn Stevenson. Risk adjusted non-pharmaceutical interventions for the management of covid-19 in south africa, 07 2020.

[108] Dennis Wesselbaum and Paul Hansen. Lockdown design: which features of lockdowns are most important to covid-19 experts? *Journal of the Royal Society of New Zealand*, 0(0):1–11, 2022.

[109] 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.

[110] 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.

[111] 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.

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

[113] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[114] Harshana Liyanage, Paul Krause, and Simon de Lusignan. Using ontologies to improve semantic interoperability in health data. *Journal of innovation in health informatics*, 22:309–15, 07 2015.

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

[116] 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.

[117] 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.

[118] 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.

[119] Carl Macrae. Early warnings, weak signals and learning from healthcare disasters. *BMJ quality & safety*, 23, 03 2014.

[120] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168:022022, 02 2019.

[121] Haider Allamy. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). 12 2014.

[122] Taushif Anwar, Uma Vijayasundaram, Md Hussain, and Muralidhar Pantula. Collaborative filtering and knn based recommendation to overcome cold start and sparsity issues: A comparative analysis. *Multimedia Tools and Applications*, 81:1–19, 03 2022.

[123] Daniel Lemire and Anna Maclachlan. Slope one predictors for online rating-based collaborative filtering. *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005*, 5, 02 2007.