SEMAPRO 2024

The Eighteenth International Conference on Advances in Semantic Processing

ISBN: 978-1-68558-190-9

September 29 - October 03, 2024

Venice, Italy

SEMAPRO 2024 Editors

Shin-ichi Ohnishi, Hokkai-Gakuen University, Japan

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

SEMAPRO 2024

Forward

The Eighteenth International Conference on Advances in Semantic Processing (SEMAPRO 2024), held between September 29th, 2024, to October 3rd, 2024, in Venice, Italy, continued a series of international events that were initiated considering the complexity of understanding and processing information. Semantic processing considers contextual dependencies and adds to the individually acquired knowledge emergent properties and understanding. Hardware and software support and platforms were developed for semantically enhanced information retrieval and interpretation. Searching for Video, Voice and Speech [VVS] raises additional problems to specialized engines with respect to text search. Contextual searching and special patterns-based techniques are current solutions.

With the progress on ontology, web services, semantic social media, semantic web, deep web search /deep semantic web/, semantic deep web, semantic networking and semantic reasoning, SEMAPRO 2024 constituted the stage for the state-of-the-art on the most recent advances.

We take here the opportunity to warmly thank all the members of the SEMAPRO 2024 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to SEMAPRO 2024. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the SEMAPRO 2024 organizing committee for their help in handling the logistics of this event.

We hope that SEMAPRO 2024 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress in the field of semantic processing.

SEMAPRO 2024 Chairs

SEMAPRO 2024 Steering Committee
Fabio Grandi, University of Bologna, Italy
Tim vor der Brück, FFHS, Switzerland
Els Lefever, LT3 | Ghent University, Belgium

SEMAPRO 2024 Publicity Chairs
Laura Garcia, Universidad Politécnica de Cartagena, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

SEMAPRO 2024
Committee

SEMAPRO 2024 Steering Committee

Fabio Grandi, University of Bologna, Italy
Tim vor der Brück, FFHS, Switzerland
Els Lefever, LT3 | Ghent University, Belgium

SEMAPRO 2024 Publicity Chairs

Laura Garcia, Universidad Politécnica de Cartagena, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

SEMAPRO 2024 Technical Program Committee

Witold Abramowicz, Poznan University of Economics, Poland
Harry Agius, Brunel University London, UK
Jose María Alvarez Rodríguez, Carlos III University of Madrid, Spain
Abdel-Karim Al-Tamimi, Sheffield Hallam University, UK
Fernanda Baiao, PUC-Rio, Brazil
Arvind Bansal, Kent State University, USA
Giuseppe Berio, Université de Bretagne Sud | IRISA, France
Floris Bex, Utrecht University & University of Tilburg, Netherlands
Karsten Boehm, University of Applied Sciences FH Kufstein Tirol, Austria
Loris Bozzato, Fondazione Bruno Kessler, Trento, Italy
Zouhaier Brahmia, University of Sfax, Tunisia
Okan Bursa, Ege University, Turkey
Ozgu Can, Ege University, Turkey
Rodrigo Capobianco Guido, São Paulo State University (UNESP), Brazil
Damir Cavar, Indiana University, USA
David Chaves-Fraga, Universidad Politécnica de Madrid, Spain
Ioannis Chrysakis, FORTH-ICS, Greece / Ghent University, Belgium
Cláudio de Souza Baptista, Federal University of Campina Grande, Brazil
Christian Dirschl, Wolters Kluwer Deutschland GmbH, Germany
Milan Dojchinovski, InfAI | Leipzig University, Germany / Czech Technical University in Prague, Czech Republic
Julio Cesar Duarte, Instituto Militar de Engenharia, Rio de Janeiro, Brazil
Enrico Francesconi, IGSG - CNR, Italy
Rolf Fricke, Condat AG, Berlin, Germany
Panorea Gaitanou, Greek Ministry of Justice, Athens, Greece
Bilel Gargouri, MIRACL Laboratory | University of Sfax, Tunisia
Fabio Grandi, University of Bologna, Italy
Jingzhi Guo, University of Macau, Macau SAR, China
Bidyut Gupta, Southern Illinois University Carbondale, USA
P. K. Gupta, Jaypee University of Information Technology, India
Shun Hattori, The University of Shiga Prefecture, Japan

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A Semantic Data Model of Harmonized Survey on Households Living Standards

Marc Mfoutou Moukala
Faculty of Science and Technology
Marien Ngouabi University
Brazzaville, Congo
e-mail: moukmarc@yahoo.fr

Macaire Ngomo
Research and Innovation Department
CM IT CONSEIL
Romilly sur Seine, France
e-mail: macaire.ngomo@gmail.com

Régis Freguin Babindamana
Faculty of Science and Technology
Marien Ngouabi University
Brazzaville, Congo
e-mail: regis.babindamana@umng.cg

***Abstract*—Since the year 2018, French-speaking countries in West and Central Africa have adopted an identical methodology for collecting household living conditions data through a survey called Harmonized Survey on Households Living Standards (HSHLS). This survey aims specifically at gathering information on socio-economic conditions of households and communities, enabling public authorities and development partners to identify areas where necessary solutions can be provided. The related questionnaire for this survey, intended to be administered electronically in the form of Computer-Assisted Personal Interview (CAPI) using tablets and telephones, helps reduce errors during data collection because some data consistency checks are managed automatically through the data collection application used, based on a set of prior real-world information. However, during data cleaning works, discrepancies are sometimes observed between the methodology and the actual collected data. Moreover, data processing teams are often forced to manually check methodologic documents for analysis purposes. In this paper, we design a semantic model, which represents the information contained in the methodological documents related to the survey under study. This model, based on an ontology built using Resource Description Framework (RDF) language and its extension, RDF Schema, allows documenting knowledge related to this survey in the form understandable by computers and easily queryable.**

Keywords-knowledge documentation; semantic data modeling; RDF ontology; Methodological information; Harmonized Survey on Households Living Standards.

I. INTRODUCTION

The Harmonized Survey on Households Living Standards (HSHLS) [1] is a household survey consisting of a set of modules or sections, each section dealing with a given theme. Among the topics addressed there are: the socio-demographic characteristics of households and their members as well as information on education, health and employment of household members. Information about each section is collected through a specific questionnaire administered using a computer application. The collected data is usually stored in a tabular structure in a relational database. After the data collection operations are executed, they are retrieved in Excel-type files for possible analyses. During data cleaning operations, discrepancies are often found between the methodology and the data actually collected, as some of the collected data do not comply with the conditions or rules defined in the methodology. Moreover, since the methodological information is not automated, data processing teams are often forced to manually consult the related documents; this sometimes causes delays in the data processing procedures. It is in this context that the present research work is conducted, aiming to develop a semantic modeling approach to represent the methodological information of the survey under study. Ultimately, the developed model will facilitate the search for methodological information and serve as a foundational tool for automatic quality control of the real data from this survey. Since the HSHLS is a major statistical survey conducted by several African countries, our motivation is to document and disseminate knowledge related to this survey to make it accessible to a wide audience.

Our paper is structured as follows: In Section II, we define the concept of semantic data modeling and its advantages. In Section III, we present the literature review related to the semantic modeling in statistical surveys and highlight some limitations of the state of the art. Section IV presents the adopted methodology and tools. In Section V, we present our model and the implementation details. In Section VI, we present and discuss the results. Finally, we end with a conclusion along with an outline of potential future work, in Section VII.

II. SEMANTIC DATA MODELING

Data semantics is the meaning given to that data; it is its significance. It encompasses all the information that can be gathered about the data with respect to a specific objective and a particular reality. For example, regarding the data point Age, the following information can be inferred: Age is a property of a person, that defines their current lifespan, expressed in years, which is the mathematical difference between the year of birth and the current year, taking into account the date of the person's most recent birthday. This lifespan is an integer between 0 (minimum duration) and 120 (maximum duration). The year of birth and the current year

 ISBN: 978-1-68558-190-9

are also properties of a person, of integer type. Semantic data modeling involves representing the data and their semantics as well as the relationships between them.

Semantic data models play a crucial role in the modeling of complex knowledge. They allow representing relationships and interactions between concepts in an accurate and structured way. These models are used in various application scenarios: representing relationships between concepts, encapsulating business knowledge, data search and analysis, integrating heterogeneous data sources, and supporting artificial intelligence and machine learning.

III. LITERATURE REVIEW

In the context of semantic modeling applied specifically to statistical survey data, the following research has strongly influenced our work.

The work in [2] addresses the challenge faced by users who need to write complex database queries to retrieve information, given their limited understanding of both the structural and semantic complexities of databases. It focuses on improving this process through the use of ontologies to facilitate better knowledge representation and interactive query generation.

Thirumahal et al. use an ontology-based approach to develop a semantic model for harmonizing and integrating population health data from heterogeneous sources [3]. Following a presentation of the ontology literature, the authors of [3] identified key concepts and relationships between population health data. Then, they used this information to develop an XML schema-based semantic ontology to harmonize and integrate population health data from different sources (Excel, SQL Server and MongoDB) for early detection of COVID-19. The authors state that the model designed allows data to be inserted, updated and deleted without anomaly as the data mapping is based on schema and not on data. The authors also state that in the future, their method could be extended by creating ontologies in RDF/Turtle formats.

Berges et al. propose an approach to improve the semantic interoperability of electronic health records using ontological management of domain ontology evolution [4]. The researchers first developed a domain ontology representing concepts and relationships relevant to the domain of electronic health records. Then, they proposed methods to manage the evolution of this ontology over time, taking into account changes in the electronic health domain and new data requirements.

In [5], Nicholson et al. use an ontology-based approach to ensure a good level of data quality for cancer-related information in registries, in order to accurately compare indicators related to this disease on regional and national scale based on harmonized rules.

The work in [13] introduces a generic ontology designed to represent questionnaires in a machine-readable format. This ontology aims to enhance decision support systems and smart environments by facilitating automatic processing of questionnaire data, which has become more abundant and cost-effective due to mobile devices. It addresses the challenge of managing and reasoning about large volumes of collected information to gain deeper insights.

Considering the literature review, we notice that there is a great deal of similar work in semantic data modeling. However, to the best of our knowledge, there is no application of these research studies to the statistical harmonized housing surveys on household living standards. Access to methodological information is manual and the majority of data quality control is conducted manually, leading to significant delays in data processing. Therefore, our contribution lies in the application of this work in the context of documentation and popularization of knowledge relating to the HSHLS survey and consists in proposing a semantic data model whose use would among other things, facilitate access to related knowledge and help speed up data processing.

IV. METHOD AND TOOLS

A. Methodology

The literature review led us to adopt an ontology-based semantic modeling approach to represent and structure knowledge semantically. This choice is justified by the fact that:

- Ontologies enable complex logical relationships between concepts to be defined formally. Axioms and rules can be used to express logical conditions of dependency between survey questions such as IF-THEN conditions, validation constraints, hierarchical relationships, etc.
- Ontologies provide a standardized, shared data model, promoting interoperability between different systems and enabling easier data integration. This can be particularly useful in a survey context, where data needs to be collected, stored and analyzed in a consistent and standardized way.
- Ontologies enable the complexity of dependencies between survey questions to be managed in a structured way. Concepts can be organized into classes and sub-classes and properties and restrictions can be defined, making it easier to manage and understand the relationships between different questions.
- Ontologies provide a solid basis for managing the evolution and maintenance of the semantic data model. Concepts and relationships can be easily added, modified or deleted without compromising model consistency and compatibility.

B. Tools

Although there are many works available in the literature that use other representation ways, to build our semantic model, represent concepts and their relationships, the Resource Description Framework (RDF) language is used in this paper.

RDF is a language standardized by the World Wide Web Consortium (W3C) to represent information on the Web in a

structured and interoperable way. It provides a simple data model based on subject-predicate-object assertions, also known as RDF triples. Each triple describes a relationship between two resources [6].

RDF Schema (RDFS) is an extension of RDF that provides a vocabulary for describing schemas and ontologies. This enables the definition of classes, properties and relationships between RDF resources [7].

To learn about the various concepts related to the survey under study as well as the relationships between the data, this research relies on methodological documents including household questionnaires, interviewer manuals and survey data dictionaries.

Our semantic model is built using RDF and RDF Schema. We propose to use metamodeling techniques to implement the models needed to manipulate the elements of the Harmonized Survey on Households Living Standards questionnaire and the survey data consistency control. The following section presents our semantic model.

V. MODELIZATION AND IMPLEMENTATION

A. Definitions of key concepts

1) Harmonized Survey on Households Living Standards(HSHLS): HSHLS is the main harmonized statistical survey conducted by French-speaking countries in West and Central Africa to capture household living conditions.

2) Section: The HSHLS survey is composed of sections. A section is named according to the topic addressed: Socio-demographic characteristics, Education, Health, etc.

3) Questionnaire: Every section has a single questionnaire. A questionnaire captures the main information of surveyed households related to a given section.

4) Question: A questionnaire is composed of questions.

5) Household: This is the main statistical unit on which information are gathered.

6) Household member: A person belonging to a particular household.

B. Presentation of the HSHLS metamodel

Our model consists of a resource class HSHLS representing HSHLS surveys, an instance of the rdfs:Class class of the RDF metamodel. A survey is made up of sections. A section contains a questionnaire. A questionnaire concerns a household, and a questionnaire is of a certain type, depending on the information collected. This may be information characterizing household members or common household characteristics. These characteristics are variables represented by questions. Each question is a property in our model and concerns a household member or a household as a whole. A question belongs to a type (integer, float, string) depending on the nature of the expected response. Each question is linked to a questionnaire. A question may depend on other questions and may have constraints.

The general metamodel includes the class declaration metamodel (Figure 1) and the property declaration metamodel (Figure 2). More details of the semantic model are given in the Appendix. In the class declaration, the declaration C rdf:type C' means that the class C is an instance of the class C'. For example, hshls:HSHLS rdf:type rdfs:Class states that the HSHLS is an instance of rdfs:Class. In the property declaration, a triple of the form: P rdfs:domain C declares that P is an instance of the rdf:Property class, that C is an instance of the rdfs:Class class, and that the resources indicated by the subjects of triplets whose predicate is P are instances of the C class. This implies that hshls:hasQuestion rdfs:domain hshls:Questionnaire states that hasQuestion is an instance of rdf:Property class, Questionnaire is an instance of the rdfs:Class class, and that the resources indicated by the subjects whose predicate is hasQuestion are instances of the class Questionnaire. The triple P rdfs:range C means that P is an instance of the rdf:Property class, that C is an instance of the rdfs:Class class, and that the resources indicated by the objects in the triple whose predicate is P are instances of the C class. In this case, hshls:hasQuestion rdfs:range hshls:Question means that hasQuestion is an instance of the rdf:Property class, Question is an instance of the rdfs:Class class, and that the resources indicated by the objects in the triple whose predicate is hasQuestion are instances of the Question class.

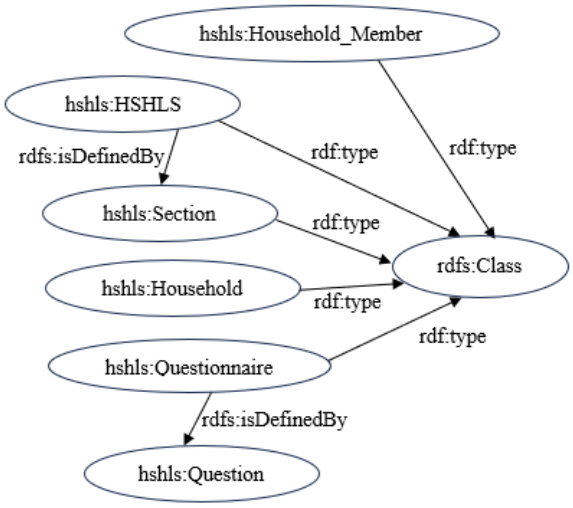

Figure 1. HSHLS RDFS metamodel with class declaration.

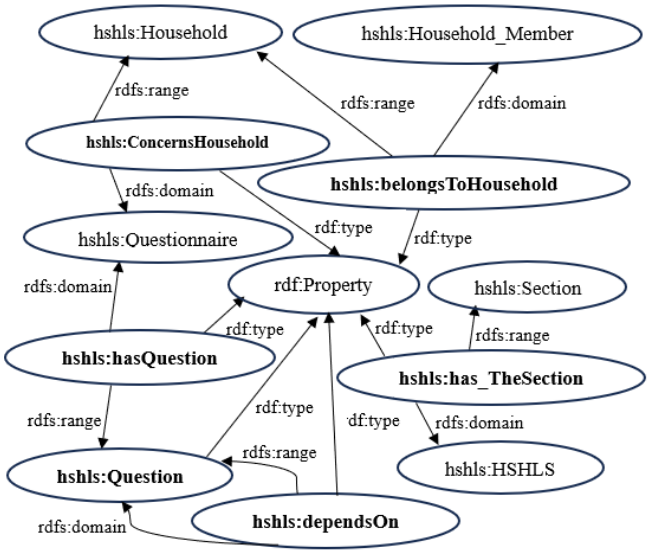

Figure 2. HSHLS RDFS metamodel with property declaration.

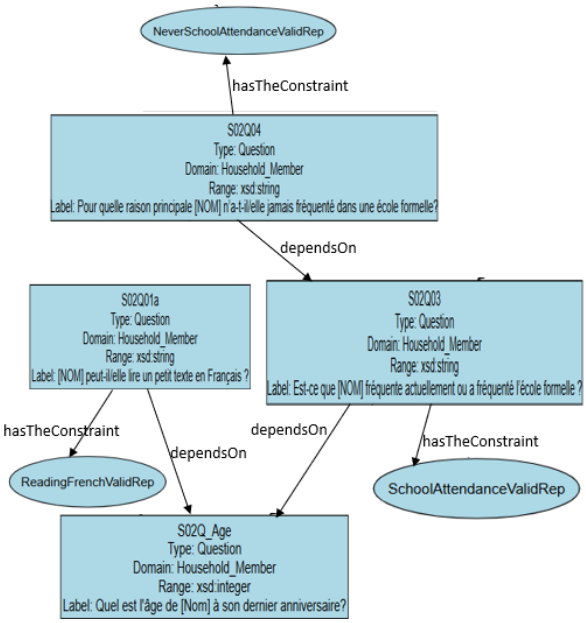

Figure 3. The HSHLS RDFS Ontology with actual questions.

C. *HSHLS model specialization*

The HSHLS has 21 sections. To serve as a proof of concept, we propose a specialization of our metamodel, considering personalized methodological information used during the first edition of this survey in Congo. We do consider a particular section of the survey: the Education section. The specialization is as follows:

1) HSHLS ontology with actual questions: Figure 3 illustrates the map of HSHLS ontology with actual questions and their dependencies and constraints, for the section which captures education's information, coded S02. This section has a single questionnaire, Questionnaire2. Questionnaire2 will be instantiated for every surveyed household. The education information concerns household members of three (3) years old or above. So the entry in this questionnaire depends on the response to the question on the age of the corresponding household member surveyed, here coded S02QAge. The question S02Q1a captures whether the surveyed household member can read a little text written in French. S02Q1a depends on the question S02QAge. An illustration of dependency conditions between questions is presented in Figure 5. The question S02Q03, which also depends on the question S02QAge, captures whether the surveyed household member is currently attending or have attended a formal school. The question S02Q04 captures the reason why the surveyed household member has never attended a formal school. S02Q04 depends on the response to the question S02Q03 which can be "Oui" (for Yes) or "Non n'a jamais fréquenté" meaning that the concerned surveyed has never attended a formal school. S02Q04 is asked only if the response to S02Q03 takes the second valid value. An illustration of constraints specification on questions is given in Figure 4.

2) HSHLS ontology constraints specification: The triples hshls:ReadingFrenchValidRep rdf:type hshls:Constraint and hshls:SchoolAttendanceValidRep rdf:type hshls:Constraint mean that ReadingFrenchValidRep and SchoolAttendanceValidRep are instances of Constraint class. The constraint ReadingFrenchValidRep specifies that the valid values of the related question (S02Q01A) are "Oui" (for Yes) or "Non" (for No). The constraint SchoolAttendanceValidRep indicates the valid values of the related question (S02Q03).

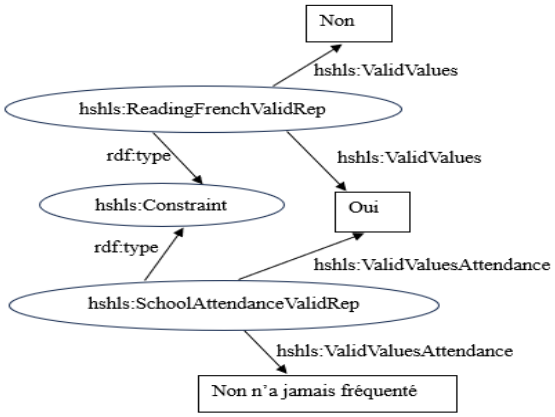

Figure 4. HSHLS RDFS Ontology constraints specification.

3) HSHLS ontology dependency specification: Figure 5 indicates that the question S02Q01a depends on the question S02QAge and the dependency condition is that the value of S02QAge (which precedes S02Q01a) must be greater or equal to 9. That means, to ask the question S02Q01a, the concerned surveyed must be 9 years old or above. If this condition is not satisfied, then the property S02Q01a must be empty for the concerned surveyed.

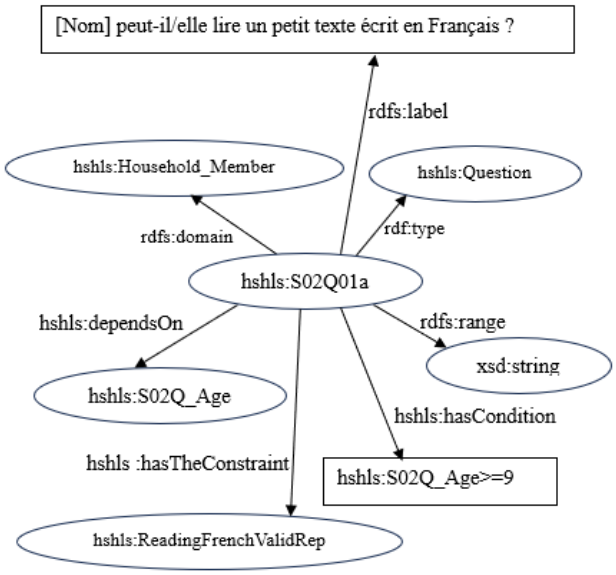

Figure 5. HSHLS ontology dependency specification.

We built the proposed model using the Python RDFLib package [8], in a Jupyter notebook environment. The model is saved in Turtle (ttl) format and can be exploited as an RDF graph. To make it possible to get access to the model in a persistent way, we defined an International Resource Identifier (IRI) for the model. We also developed an HTML ontology documentation file using PyLODE [9]. We created a public Github repository [10] and saved the rdf graph and its HTML documentation in it.

VI. RESULTS AND DISCUSSION

This model makes it possible to store the methodological information of the Harmonized Survey on Households Living Standards in such a way that it can be understood by the computer and retrieved automatically. By specifying the semantics of the questions addressed, this model helps to better understand the meaning of the data manipulated in this survey as well as the semantic relationships that exist between these data. With the specification of some constraints and conditions on related questions, this model can serve as a fundamental tool for data quality control on actual data during data collection and processing. Since the model is saved in a persistent repository, one can easily get access and perform some retrievals and analysis requests using SPARQL [11] or any appropriate data analysis tool. Also, a large audience can get access and learn related knowledge. Therefore, the project will not only help improving the efficiency during the survey data collection and processing activities, but also contributes to the dissemination of the survey knowledge.

VII. CONCLUSION AND FUTURE WORK

In this work we propose an approach to semantic data modeling of the Harmonized Survey on Households Living Standards. The model built makes it possible to store the semantic information contained in the methodology of this survey so that it can be consulted automatically. The results of this work can therefore be exploited as part of the automatic retrieval of methodological information. Generally speaking, this work completes the state of the art and serves as a proof of concept to demonstrate the feasibility of documenting the knowledge contained in a statistical survey questionnaire through ontology-based semantic modeling. Researchers from a variety of backgrounds will find it a source of information when it comes to design approaches requiring ontology-based semantic modeling of data, whether statistical survey data or not.

This model highlights semantic information derived from the methodology of the Harmonized Survey on Households Living Standards. To enable automatic data quality control based on this model, an extension of the model will be developed in the future with complex constraints and conditions, using OWL2 [12] or another equivalent language that we will study, which is complementary to and interoperable with RDF. An automatic reasoning engine will be built for the purpose of anomaly detection in actual data.

ACKNOWLEDGMENT

Marc Mfoutou Moukala thanks the representative of the Agence Universitaire de la Francophonie (AUF) in Congo, for the seminar on writing a scientific paper.

REFERENCES

[1] Harmonized Survey on Households Living Standards 2018-2019, Food and Agriculture Organization of the United Nations, 2022, https://microdata.fao.org/index.php/catalog/2355/study-description (consulted on May 10, 2024).

[2] K. Munir et al., The use of ontologies for effective knowledge modelling and information retrieval, Applied Computing and Informatics, Vol. 14, N°2, pp. 116–126, 2018.

[3] R. Thirumahal, G. Sudha Sadasivam and P. Shruti, Semantic Integration of Heterogeneous Data Sources Using Ontology Based Domain Knowledge Modeling for Early Detection of COVID-19, SN Computer Science, Vol. 3, No. 428, 2022, https://doi.org/10.1007/s42979-022-01298-4.

[4] I. Berges, J. Bermúdez and A. Illarramendi, Towards Semantic Interoperability for Electronic Health Records, IEEE Trans Inf Technol Biomed, Vol. 16, N°3, pp. 424-431, 2012, doi: 10.1109/TITB.2011.2180917.

[5] N. C. Nicholson et al., An ontology-based approach for developing a harmonised data-validation tool for European

cancer registration, Journal of Biomedicam semantics, Vol. 12, N°1, pp. 1-15, 2021.

[6] RDF 1.1 Primer, W3C Working Group Note 24 June 2014 https://www.w3.org/TR/rdf11-primer/, (consulted on May 10, 2024).

[7] RDF Schema 1.1, W3C Recommendation 25 February 2014, (Document updated on December, 1th 2023), https://www.w3.org/TR/rdf-schema/, (consulted on May 10, 2024)

[8] RDFLib. https://pypi.org/project/rdflib/, (consulted on May 10, 2024).

[9] PyLODE. https://pypi.org/project/pylode/, (consulted on August 20, 2024).

[10] M. Mfoutou Moukala, HSHLS survey ontology Web Page, https://moukmarc.github.io/.

[11] SPARQL 1.1 Query Language, W3C Recommendation 21 March 2013, https://www.w3.org/TR/sparql11-query/, (consulted on May 10, 2024).

[12] OWL2 Web Ontology Language, W3C Recommendation 11 December 2012, Document Overview (Second Edition), https://www.w3.org/TR/owl2-overview/, (consulted on May 10, 2024).

[13] A. V. Borodin and Y. V. Zavyalova, An ontology-based semantic design of the survey questionnaires, 19th Conference of Open Innovations Association (FRUCT), 2016, Jyvaskyla, Finland, pp. 10-15.

APPENDIX

Here is a part of the semantic model of the Harmonized Survey on Households Living Standards:

```
@prefix hshls: <http://w3id.org/HshlsOnto/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix  rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd:<http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix vann: <http://purl.org/vocab/vann/> .
hshls: a owl:Ontology;
   rdfs:seeAlso "https://github.com/moukmarc/HshlsOnto" ;
   dcterms:creator "Marc Mfoutou Moukala" ;
   dcterms:title "Harmonized survey on household living standards Ontology (HshlsOnto)" ;
   vann:preferredNamespacePrefix "hshls" .
# Core classes declaration
hshls:HSHLS a rdfs:Class ;
   rdfs:isDefinedBy hshls:Section ;
   rdfs:label " Harmonized Survey on Households Living Standards" .
hshls:Section rdf:type rdfs:Class ;
        rdfs:label " A section or module of the survey " .
hshls:Questionnaire rdf:type rdfs:Class ;
        rdfs:isDefinedBy hshls:Question .
hshls:Household a rdfs:Class ;
   rdfs:label " A household " .
hshls:Household_Member rdf:type rdfs:Class ;
   rdfs:label " A household member surveyed " .
hshls:Constraint rdf:type rdfs:Class ;
      rdfs:label " Constraint on the question " .
# Property declaration
hshls:ccontainsQuestionnaire a rdf:Property ;
 rdfs:domain hshls:Section ;
 rdfs:range hshls:Questionnaire ;
   rdfs:label "contains questionnaire" .
hshls:ConcernsHousehold a rdf:Property ;
   rdfs:domain hshls:Questionnaire ;
   rdfs:range hshls:Household ;
   rdfs:label " Concerns household " .
hshls:typeOfQuestionnaire a rdf:Property ;
   rdfs:domain hshls:Questionnaire ;
   rdfs:range rdfs:Literal ;
   rdfs:label " type of questionnaire " .
hshls:Question a rdf:Property ;
   rdfs:label " A question relating to the survey " .
hshls:belongsToHousehold a rdf:Property ;
   rdfs:domain hshls:Household_Member ;
   rdfs:range hshls:Household ;
   rdfs:label " belongs to household " .
hshls:hasTheSection rdf:type rdf:Property ;
        rdfs:domain hshls:HSHLS ;
        rdfs:range hshls:Section ;
        rdfs:label " Includes section " .
hshls:hasQuestion rdf:type rdf:Property ;
        rdfs:domain hshls:Questionnaire ;
        rdfs:range hshls:Question ;
        rdfs:label " Includes question " .
hshls:hasTheTypeOfResponse rdf:type rdf:Property ;
        rdfs:domain hshls:Question ;
        rdfs:range xsd:string ;
        rdfs:label " has the question answer type " .
hshls:hasTheConstraint rdf:type rdf:Property ;
      rdfs:domain hshls:Question ;
      rdfs:range hshls:Constraint ;
      rdfs:label " has the constraint " .
hshls:dependsOn rdf:type rdf:Property ;
      rdfs:domain hshls:Question ;
      rdfs:range hshls:Question ;
      rdfs:label " depends on " .
hshls:hasCondition rdf:type rdf:Property ;
      rdfs:domain hshls:Question ;
     rdfs:range xsd:string ;
     rdfs:label " has the condition " .
# Constraint Classes
hshls:NumericConstraint rdf:type rdfs:Class ;
        rdfs:subClassOf hshls:Constraint ;
        rdfs:label "Numeric constraint" .
# Constraint Properties
hshls:minValue rdf:type rdf:Property ;
      rdfs:domain hshls:NumericConstraint ;
      rdfs:range xsd:float ;
      rdfs:label "Minimum value" .
hshls:maxValue rdf:type rdf:Property ;
      rdfs:domain hshls:NumericConstraint ;
      rdfs:range xsd:float ;
      rdfs:label "Maximum value" .
# Model specialization
```

```
hshls:HSHLS-C1 rdf:type hshls:HSHLS ;
    hshls:hasTheSection hshls:S00, hshls:S01, hshls:S02, hshls:S03, hshls:S04, hshls:S05, hshls:S06, hshls:S07,
    hshls:S08, hshls:S09, hshls:S10, hshls:S11, hshls:S12, hshls:S13, hshls:S14, hshls:S15, hshls:S16, hshls:S17,
    hshls:S18, hshls:S19, hshls:S20, hshls:S21 ;
    rdfs:comment "The different sections of HSHLS survey for the first edition in Congo named here HSHLS-C1" .
hshls:S02 a hshls:Section ;
hshls:ccontainsQuestionnaire hshls:Questionnaire2 ;
rdfs:comment "Every HSHLS section contains one questionnaire, which will be instanciated for every surveyed household" .
# Questionnaires in section S02
hshls:Questionnaire2 a hshls:Questionnaire ;
hshls:typeOfQuestionnaire "Household member's education information" ;
hshls:ConcernsHousehold hshls:Household ;
hshls:hasQuestion hshls:S02Q_Age, hshls:S02Q01a, hshls:S02Q01b, hshls:S02Q01c, hshls:S02Q01d,
hshls:S02Q02a, hshls:S02Q02b, hshls:S02Q02c, hshls:S02Q02d, hshls:S02Q03a, hshls:S02Q03b, hshls:S02Q03c, hshls:S02Q03d, hshls:S02Q04, hshls:S02Q05, hshls:S02Q06, hshls:S02Q07, hshls:S02Q08, hshls:S02Q09, hshls:S02Q10, hshls:S02Q11, hshls:S02Q12, hshls:S02Q13, hshls:S02Q14, hshls:S02Q15, hshls:S02Q16,
hshls:S02Q17, hshls:S02Q18, hshls:S02Q19, hshls:S02Q20, hshls:S02Q21 ;
rdfs:comment "The Questionnaire2 captures household member's education information for members of 3 years old and more" .
# Questions of Questionnaire2
hshls:S02Q_Age a hshls:Question ;
rdfs:domain hshls:Household_Member ;
rdfs:range xsd:integer ;
rdfs:label "Quel est l'âge de [Nom] à son dernier anniversaire?" ;
rdfs:comment "The question S02Q_Age captures the age of the surveyed household member" ;
rdfs:subClassOf [
a rdfs:Datatype ;
rdfs:subClassOf xsd:integer ;
rdfs:minInclusive "0"^^xsd:integer ;
rdfs:maxInclusive "120"^^xsd:integer
] .
hshls:S02Q01a rdf:type hshls:Question ;
   rdfs:domain hshls:Household_Member ;
    rdfs:range xsd:string ;
    rdfs:label "[NOM] peut-il/elle lire un petit texte en Français ?" ;
    hshls:hasTheTypeOfResponse "String" ;
    hshls:dependsOn hshls:S02Q_Age ;
    hshls:hasCondition "( hshls:S02Q_Age >= 9)" ;
    hshls:hasTheConstraint hshls:ReadingFrenchValidRep ;
    rdfs:comment "This question captures whether the surveyed household member can read a little text written in french" .
hshls:S02Q03 rdf:type hshls:Question ;
    rdfs:domain hshls:Household_Member ;
    rdfs:range xsd:string ;
    rdfs:label "Est-ce que [NOM] fréquente actuellement ou a fréquenté l'école formelle ?" ;
    hshls:hasTheTypeOfResponse "String" ;
    hshls:dependsOn hshls:S02Q_Age ;
    hshls:hasCondition "( hshls:S02Q_Age >= 3)" ;
    hshls:hasTheConstraint hshls:SchoolAttendanceValidRep ;
    rdfs:comment "This question captures whether the surveyed household member is currently attending or have attended a formal school" .
hshls:S02Q04 rdf:type hshls:Question ;
    rdfs:domain hshls:Household_Member ;
    rdfs:range xsd:string ;
    rdfs:label "Pour quelle raison principale [NOM] n'a-t-il/elle jamais fréquenté dans une école formelle?" ;
    hshls:hasTheTypeOfResponse "String" ;
    hshls:dependsOn hshls:S02Q03 ;
    hshls:hasCondition "(hshls:S02Q03 = Non n'a jamais fréquenté)" ;
    hshls:hasTheConstraint hshls:NeverSchoolAttendanceValidRep ;
    rdfs:comment "This question captures the reason why the surveyed household member has never attended a formal school" .
# Constraint for valid responses for the question S02Q01a
hshls:ReadingFrenchValidRep a hshls:Constraint ;
hshls:ValidValues "Oui","Non" .
# Constraint for valid responses for the question S02Q03
hshls:SchoolAttendanceValidRep a hshls:Constraint ;
hshls:ValidValuesAttendance"Oui","Non n'a jamais fréquenté" .
# Constraint on S02Q04 for valid responses why the surveyed never attended a formal school
hshls:NeverSchoolAttendanceValidRep a hshls:Constraint ;
hshls:ValidValuesNeverAttendance "Trop jeune (moins de 6 ans)", "Pas d'école, école trop éloignée", "Refus de la famille","Préférence pour un emploi", "Travaux champêtres/pastoralisme", "Travaux domestiques", "Mariage", "Renvoi", "Frais de scolarité élevés", "Manque de moyens financiers", "Etudes non adaptées", "Etudes peu utiles", "Malade", "Pas d'acte de naissance", "Handicap", "Insécurité", "Autre (à préciser)" .
```

An Architecture for Ontology-based Semantic Reasoning Using LLMs in Healthcare Domain

Müge Oluçoğlu
Department of Computer Engineering
Izmir Bakircay University
Izmir, Turkey
muge.olucoglu@bakircay.edu.tr

Okan Bursa
Department of Computer Engineering
Izmir Bakircay University
Izmir, Turkey
okan.bursa@bakircay.edu.tr

***Abstract*—This study presents research that involves examining various methods used in the field of health services and proposing a new architecture for more precise diagnosis from health records. Traditional and modern methods such as Electronic Health Records (EHR), Clinical Decision Support Systems (CDSS), Natural Language Processing (NLP)-based analytics, and Machine Learning (ML) techniques are discussed, highlighting their advantages, disadvantages, and usage areas. Based on these evaluations, a new method involving ontologies and Large Language Models (LLMs) has been developed to provide a more effective solution for healthcare informatics. The proposed approach is a candidate solution to achieve higher accuracy, speed, and flexibility by integrating ontologies, LLM and reasoners.**

Keywords-semantic reasoning; ontology; healthcare; knowledge graph; large language model.

I. INTRODUCTION

Health informatics is currently undergoing a significant transformation with the integration of big data and artificial intelligence technologies. At the center of this transformation are Large Language Models (LLMs) and Knowledge Graphs (KGs). LLMs have made groundbreaking advances in Natural Language Processing (NLP) techniques and can extract meaningful and contextual information from large datasets [1]. KGs, on the other hand, improve knowledge integration and extraction processes by representing the relationships between data at a semantic level.

Healthcare is an emerging field where LLMs can provide significant advantages in data analysis and interpretation. These models can extract meaningful information from large and diverse data sources such as electronic health records, clinical notes and medical literature, and provide valuable insights to healthcare professionals in decision support systems [2]. However, the complexity and multidimensionality of health data require the use of advanced techniques to interpret and process these data accurately and efficiently. Knowledge graphs play an important role in addressing this complexity.

Knowledge graphs are structured data models used for semantic modeling and identifying relationships between data elements. In healthcare, knowledge graphs integrate different data sources such as patient data, genetic information, medical literature and clinical trials, and establish semantic links between them [3]. This integration enables healthcare professionals and researchers to better understand and analyze complex data relationships. On the other hand, this integrated representation of the knowledge and data makes the processing of the domain knowledge and claims of the rules of the domain from the data more efficient. This extraction is crucial to submerge the existing knowledge and rules from the data to provide more precise definitions for the healthcare domain.

The relationship between LLMs and KGs has great potential for the processing and analysis of health data. The aim is to create models that improve healthcare outcomes, obtain personalized and accurate findings, and support human decision-making processes [4][5]. When the NLP capabilities of LLMs are combined with knowledge graphs, it becomes possible to analyze and extract the semantics of health data more effectively. This integration enables the development of knowledge-based decision support systems in healthcare and potential to provide more accurate results in clinical decision processes [6].

In this regard, the aim of this study is to investigate semantic reasoning processes in the healthcare domain using LLMs and KGs. Existing approaches in the literature for semantic modeling, integration and analysis of health data will be summarized to create an architecture to merge not only knowledge graphs but also ontologies with the LLM to obtain a framework to build semantic-aware LLM applications. Furthermore, a new reasoning structure is introduced with the usage of Semantic Web Rule Language (SWRL) [23] for domain knowledge and Simple Protocol and Resource description framework Query Language [24] (SPARQL) to query the inferred entities.

In the following sections of the study, the literature review will be described in Section 2, and the methods, databases and ontologies currently used for selected studies in different healthcare fields will be explained in Section 3. In Section 4, the difference between the architecture created and the existing studies will be discussed. We conclude the article in Section 5.

II. LITERATURE REVIEW

In this part of the article, as a result of the literature review, studies on different health fields carried out using LLMs and KGs will be explained. Since current content was desired to be included, studies between 2020 and 2024 were in the literature review. The research summarized in these papers focuses on leveraging various artificial intelligence and semantic technologies to enhance healthcare predictions, diagnosis, and management, particularly in the domains of chronic diseases, mental health disorders, and Alzheimer's disease.

The study from [7] introduces dynamic and adaptive approaches, such as the dynamic fuzzy rule-based inference system for Alzheimer's disease diagnosis and the ostensive

information architecture for enhancing semantic interoperability in healthcare information systems. The authors describe a real-world case study utilizing an Alzheimer's Disease (AD) diagnosis system that is built on fuzzy, dynamic, and semantic decision criteria. The study's dataset was compiled from medical records of patients in the USA and Canada. Semantic data, including genetics, screening findings, treatments, etc. from patient diagnostic results is used to develop a recommended system and do a comparative study. Machine learning and ontology-based systems were compared using the same data.

To handle uncertainty in medical information, particularly in relation to the diagnosis of mental health issues, the combination of fuzzy logic and ontologies is advocated. The complexity and uncertainty associated with mental health illnesses are addressed in this study by proposing a machine learning model that integrates fuzzy logic and ontology with Mamdani inference in Fuzzy Ontology Web Language (OWL) for Protegee for the diagnosis of Major Depressive Disorder (MDD) [8]. Nine input variables—such as mood, sleep, hunger and weight, joy and pleasure, fatigue, guilt, memory, and psychomotor impairment—are linked to symptoms of MDD. To account for the uncertainty in characterizing symptoms, each of these variables is mapped to Type-2 fuzzy sets with trapezoidal membership functions. Based on input symptoms, the suggested fuzzy inference system applies a set of rules to determine the degree of depression.

Many studies integrate ontologies, knowledge graphs, and semantic reasoning to improve the accuracy of clinical prediction models using Electronic Health Record (EHR) data. In another study [9] that integrates semantic reasoning using ontology-based decision support and recommendation systems for diabetes nutrition treatment and diagnosis, fuzzy medical rules are used. The study in [10], on the other hand, offers a nutrition recommendation system. The authors of [11] have developed a system using the symptoms and complaints of diabetic patients in Fast Health Interoperability Resources (FHIR). In another study, the authors created a prediction system with personal health information in the EHR system [12].

Recent developments in LLMs are increasing interest in their potential applications in medicine. Studies evaluate the performance of LLMs on a variety of health-related tasks, including clinical language understanding, new drug discovery, and health prediction, using wearable sensor data. Challenges such as hallucinations and the need for task-specific orientation strategies are also addressed. [13] developed an LLM-based query-answer system for drug discovery in cancer research. It validates gene-disease associations using machine learning techniques to analyze multimodal data [14]. [15] performs semantic reasoning using past patient records. The system proposed by the authors asks and receives informative questions and answers regarding the clinical scenarios at hand. The authors used National Center for Biotechnology Information (NCBI-Disease), BC5CDR-Chemical, i2b2 2010-Relation, SemEval 2013-DDI, BIOSSES, MedNLI, i2b2 2006-Smoking, BioASQ 10b-Factoid, BioASQ 10b-Factoid databases. LLMs have also been used in medical decision-making. The authors of [16] used decision trees to produce more reliable medical answers. The data attributes used by the authors of [17] are user demographics, health information, stress, readiness, fatigue, activity, calories, sleep quality, sleep disorder, anxiety, and depression. This variety of patient data attributes in [17] enables LLMs to make inferences about health based on semantic reasoning.

In addition to all these benefits that LLMs provide in the field of healthcare, situations that may cause data security, ethical handling, etc. and many other problems are discussed in [18] and [19]. It is mentioned how and what solutions will be produced against these problems.

The proposed architectural solution leverages ontologies to reason over the collected data, while also enabling queries with well-defined parameters. This approach has the potential to identify irregularities in the LLM's responses, leading to more accurate and precise answers for the patient over health records.

III. METHODOLOGY

In this section, existing methodologies in different healthcare fields will be discussed in detail.

In this part of the study, there are many different studies conducted in different fields of health services in the literature. The methods and ontological languages used in these studies are explained in more detail in Section 2.

In [7], dynamic and adaptive methodologies are introduced, such as a complete information architecture to improve semantic interoperability in health information systems and a dynamic fuzzy rule-based inference system for Alzheimer's disease detection. The dataset for the research is composed of medical records from patients in the USA and Canada, derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the World Health Organization (WHO). After patient information was collected, data preprocessing steps were used. In the Semantic Reasoning stage, neuropsychological cognitive data, Cerebrospinal Fluid (CSF) data, MRI-PET, physical examination, demographic data, and genetic data obtained from patients are divided into numerical and categorical classes. These classes are later accepted as semantic features for inference. Thus, semantic reasoning is derived by performing ontological reasoning. The results obtained here are made from fuzzy reasoning by applying fuzzification. This system, created by processing semantic data such as symptoms, allows early diagnosis of Alzheimer's disease.

In [8], a study was conducted for the diagnosis of MDD, which is complex and unclear, related to mental health and illness. An anonymous medical dataset of 90 people with depression between the ages of 25 and 65, selected from a mental health center in Iran, was used. 9 parameters were selected for diagnosis in line with the standards recognized by The Diagnostic and Statistical Manual of Mental Disorders (DSM). These parameters; (1) mood, (2) sleep, (3) change in appetite & weight, (4) joy & pleasure, (5) feeling tired, (6) feeling guilty, (7) memory, (8) the dominant force

in suicide and (9) psychomotor disturbance. After the fuzzifier step is applied to the findings obtained, the knowledge base and rules base steps are taken for definitive inference. Defuzzification is applied to determine the severity of depression. The result of the decision is given to the application user and doctor. There are 4 modules in the recommendation system for the diagnosis phase of the study. Module 1 is the patient's electronic record repository, which contains the patient's history of health problems and personal information such as name, gender, and age. Module 2 is an ontological database where all variables of the patient's medical care (Dopamine, Cortisol, Growth Hormone, Norepinephrine, Thyroid) and psychological symptoms are semantically explained and stored. Module 3 is the implementation of the fuzzy inference system that allocates input variables to output variables based on rules and regulations defined in the ontology database. Module 4 is the control and decision-making layer.

In [9], a diagnostic and nutritional recommendation system for diabetes was designed. It creates a linguistic fuzzy rule base that integrates information from the EHR and domain experts with information obtained from training data and information extracted from the semantic model. The authors developed a fuzzy knowledge-based system on pre-processed data using machine learning algorithms. Mamdani inference engine was applied to the results of these pre-processing using fuzzy rules.

The authors of [11] have developed a system for detecting complaints and symptoms of diabetic patients. The authors started from the problems in implementing FHIR. A Semantic Engine with FHIR knowledge graph as the core was built using Neo4j to provide semantic interpretation and examples. The authors obtained the data in Medical Information Mart for Intensive Care (MIMIC) and diabetes datasets. In the architecture that had been created in [11], the components of the semantic engine are the FHIR knowledge graph and transformation components. FHIR knowledge graph consists of three layers. The first layer consists of local health information systems, which act as sources for collective health data. The second layer, known as the transformation layer, functions as an integrator between the first and third layers, utilizing a query processor, linker, and mapping connector to process the data. Once connected, the health data is transferred to the third layer, the semantic interpretation layer. This layer is responsible for interpreting the semantics by defining the lexical meaning of nodes and their relationships, as established in the second layer. As a result, the FHIR Knowledge Graph is constructed in the final layer.

In [13], the authors developed the Knowledge Graph Based Thinking (KGT) model by combining LLM's knowledge graphs to reduce real errors in reasoning. KGT uses LLMs to create the optimal subgraph based on important information extracted from the question.

TABLE I. LITERATURE REVIEW

Ref	***Year***	***Domain***	***Method***	***Ontology Language***	***Dataset***
[7]	2024	Alzheimer's disease diagnosis	Semantic features were created using ontological reasoning. Fuzzification was applied to the results	Fuzzy OWL	Alzheimer's Disease Neuroimaging Initiative (ADNI)
[8]	2023	Diagnosed with major depressive disorder	After the fuzzification process, knowledge base and rule base steps were used. Depression severity was clarified	Fuzzy OWL	Anonymous patients' records
[9]	2020	Treatment diagnosis for diabetes	Creating a linguistic fuzzy rule base for the information coming from the semantic model	Fuzzy OWL	Electronic Health Records (EHR)
[11]	2024	Local health information system diabetes	The components of the semantic engine consist of knowledge graph of the FHIR data and its transformation components	OWL	Medical Information Mart for Intensive Care (MIMIC)
[13]	2024	Drug discovery for cancer research	Creates the optimal subgraph using important information	-SynLethKG -SDKG -SOKG	--
[14]	2023	Gene-disease relationship	Cancer information is integrated into the ontology. Final fine-tuning with LLMs	Ontological rules	-OncoNet Ontology (ONO) -Scientific Article

It facilitates the discovery of new uses of existing drugs through drug-cancer relationships. The authors use the SmartQuerier Oncology Knowledge Graph (SOKG) ontology, a pan-cancer knowledge graph.

The Onco Net ontology (ONO) used in [14] is created with scientific articles and real information obtained from various sources. The obtained cancer-related information is integrated into the ontology. Final fine-tuning is being done with LLM's as it has a more comprehensive knowledge graph. Thus, it is aimed to confirm the gene-disease diagnosis relationship.

Table 1 provides information on current methods and practices used in healthcare in different fields for selected studies. It includes summary information about the year, method, ontology used, study area and databases of the studies. Studies in different healthcare fields, mostly published in 2023 and 2024, are included.

IV. ARCHITECTURE

In this chapter, we introduce an architecture for ontology-based semantic reasoning using LLMs in the healthcare domain. Various methods and ontologies used in health services have been examined to develop a robust framework. In previous studies, different datasets and analytical techniques were applied across diverse healthcare fields. For example, dynamic methodologies were developed for the early diagnosis of Alzheimer's using ADNI and WHO data, leveraging ontological reasoning. In another study on major depressive disorder (MDD) diagnosis, DSM parameters were applied to assess depression severity. For diabetes diagnosis and nutrition recommendation systems, a Mamdani inference engine utilizing fuzzy rules based on EHR data was implemented. A similar approach was used to identify diabetic symptoms through a semantic engine built with FHIR knowledge graph data. These efforts exemplify how reasoning models, such as the Reasoning-Based Thinking (KGT) model, integrate LLMs with knowledge graphs for enhanced diagnostic precision. The architecture presented here builds on these foundational works, aiming to offer a scalable and efficient solution for semantic reasoning in healthcare using ontologies and LLMs.

The architecture created in this study was designed as a system that determines possible treatment methods by performing semantic reasoning in the field of healthcare. The steps of the architecture created for the study are clearly seen in Figure 1.

The first step is to collect data from the patient. Patient data can be in text, image or video format. To process this multi-modal data, it is first tokenized into small pieces. On the other hand, the same data is also combined with the prompting process of a domain expert or a doctor defined with SWRL rules.

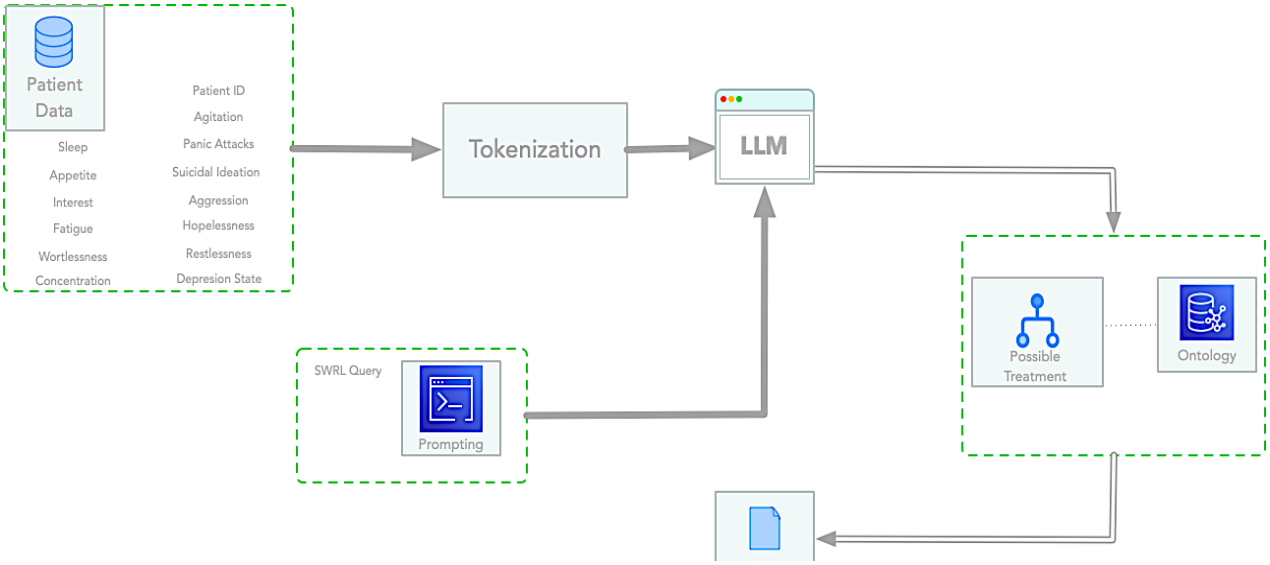

Figure 1. Proposed Model Architecture.

Figure 2 shows an example SWRL rule. It states that if a particular patient (person) uses certain substances and has certain symptoms (dry mouth and muscle tension), that patient should seek care from a doctor and receive psychotherapy treatment. The first step is to collect data from the patient. Patient data can be in text, image or video format. To process this multi-modal data, it is first tokenized into small pieces. On the other hand, the same data is also combined with the prompting process of a domain expert or a doctor defined with SWRL rules. SWRL is a rule definition language that allows us to create rules using entities in RDF datasets. These rules enable patient data to be made more meaningful and enriched with semantic features. Patient data and the data obtained by the prompting process are fed into

 ISBN: 978-1-68558-190-9

an LLM. This model is trained with a broad knowledge base and language knowledge, allowing it to understand patient data more deeply.

```
PERSON(?PATIENT), MAKESUSE(?PATIENT,
?SUBSTANCE), DOCTOR(?DOCTOR),
PSYCHOTHERAPY(?TREATMENT),
SYMPTOMISASSOCIATEDWITH(?CASE, ?SYMPTOM),
DRYMOUTH(?SYMPTOM),
MUSCULARTENSION(?SYMPTOM) ->
TAKESCARE(?DOCTOR, ?PATIENT),
TREATS(?PATIENT, ?TREATMENT)
```

Figure 2. Example of SWRL.

Data analyzed by LLM is used to identify potential treatment methods. LLM evaluates patient data and applies SWRL rules to analyze tokenized healthcare information. Based on this evaluation, it generates and promotes the most appropriate treatment options tailored to the patient's specific needs. However, in order to further optimize the response, knowledge graphs are used to determine treatment methods. The response will be reconstructed within the knowledge graph to find the possible treatments are valid due to SWRL rules. In order to check the valid responses, knowledge graphs are queried similar to graphs to create a subgraph of the treatment response of the LLM. A valid response means that a sub-graph can be extracted, and this shows that the response is a known method in the healthcare domain.

The possible treatment methods identified by the LLM are determined as the final results to be presented to the user and the doctor. These results are presented to the user in an understandable format. It provides information about the details and applicability of treatment methods.

This architecture aims to make sense of patient data and determine the most appropriate treatment methods by performing semantic reasoning in the field of healthcare. The system consists of the steps of tokenizing patient data, enriching it with prompting and SWRL rules, analyzing it using LLM, and checking its validity with help of knowledge graphs.

The dataset chosen for this study is related to mental health. The dataset was obtained from Kaggle [20]. In the dataset, there are 13 personal attributes of information about the patients: frequency of sleep disturbance, change in appetite, loss of interest in activities, fatigue or low energy, feeling of worthlessness or guilt, difficulty concentrating, physical agitation, suicidal ideation, aggression, experiencing panic attacks. There is information about despair, restlessness and general depression. Except for the general depression state, the values of other parameters are between 1 and 6. These values are 1: Never, 2: Always, 3: Often, 4: Rarely, 5: Sometimes, 6: Not at all. Values for the General Depression State parameter are classified as: "No depression", "Mild", "Moderate", and "Severe".

The ontology constructed in the study [21] includes concepts and features used to risk classification in mental health. The study aims to determine the patient's risk level by scoring signs and symptoms. The ontology consists of 332 classes, 82 individuals and 37 properties.

V. CONCLUSION

This study examines the effectiveness of a new architecture built with Large Language Models (LLMs) using ontology-based semantic reasoning in the healthcare field. The mental health dataset and the ontology in this field were used as a sample study area. Dynamic and adaptive methodologies in the literature have been compared with systems used in the diagnosis and treatment of various health problems such as Alzheimer's disease diagnosis and major depressive disorder.

In the proposed architecture, patient data is collected and made analyzable by tokenization, and a prompting process enriched with SWRL rules is applied. This data is processed using LLM and potential treatment methods are identified through knowledge graphs. This approach enables more accurate and comprehensive decisions in healthcare, thanks to the combination of semantic reasoning and natural language processing techniques. Applications made on the mental health dataset can be used in the diagnosis of psychological disorders, especially depression and anxiety, and in the treatment recommendation system.

In the future, expanding and modifying this architecture to entail additional healthcare domains presents a possible avenue to enhance healthcare quality. Text-based data is the main focus of the current system. In order to achieve better prediction, image and video data will be integrated to create a multimodal structure. This makes it possible to assess patients' body language, facial expressions, and other visual cues to deliver more thorough and precise diagnostic and treatment suggestions. Sentiment analysis will be used to assess the emotional tone and substance of the patients' expressions in order to gather more detailed information on the mood and emotional state of the patients. The most suitable course of treatment will be selected by optimizing the system's learning capabilities by considering an integration of the patient's unique medical history, genetic information, and preferences.

REFERENCES

[1] M. P. Joachimiak et al., “The Artificial Intelligence Ontology: LLM-assisted construction of AI concept hierarchies”, arXiv preprint arXiv:2404.03044, 2024.

[2] B. Li, T. Meng, X. Shi, J. Zhai, and T. Ruan, “MedDM: LLM-executable clinical guidance tree for clinical decision-making”, arXiv preprint arXiv:2312.02441, 2023.

[3] S. Singh, “Natural language processing for information extraction”, arXiv preprint arXiv:1807.02383, 2018.

[4] D. -Q. Wang et al., “Accelerating the integration of Chat-GPT and other large-scale AI models into biomedical research and healthcare”, MedComm – Future Medicine, vol. 2(2), https://doi.org/10.1002/mef2.43, 2023.

[5] K. Singhal et al., “Towards Expert-Level Medical Question Answering with Large Language Models”, arXiv.2305.09617, 2023.

[6] Q. C. Ong et al., “Advancing Health Coaching: A Comparative Study of Large Language Model and Health Coaches”, Available at SSRN 4784958, pp. 1-22, 2024.

[7] N. Shoaip, S. El-Sappagh, T. Abuhmed, and M. Elmogy, “A dynamic fuzzy rule-based inference system using fuzzy inference with semantic reasoning”, Scientific Reports, 14(1), pp. 1-17, 2024.

[8] A. Ghorbani, F. Davoodi, and K. Zamanifar, “Using type-2 fuzzy ontology to improve semantic interoperability for healthcare and diagnosis of depression”, Artificial Intelligence in Medicine, vol. 135, pp. 1-17, 2023.

[9] N. Shoaip, S. El-Sappagh, S. Barakat, and M. Elmogy, “A framework for disease diagnosis based on fuzzy semantic ontology approach”, International Journal of Medical Engineering and Informatics, 12(5), pp. 475-488, 2020.

[10] D. Spoladore, M. Tosi, and E. C. Lorenzini, “Ontology-based decision support systems for diabetes nutrition therapy: A systematic literature review. Artificial Intelligence in Medicine”, 102859, pp. 1-21, 2024.

[11] H. Guo, M. Scriney, and K. Liu, “An ostensive information architecture to enhance semantic interoperability for healthcare information systems”, Information Systems Frontiers, 26(1), pp. 277-300, 2024.

[12] P. Jiang, C. Xiao, A. R. Cross, and J. Sun, “GraphCare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs”, arXiv preprint arXiv:2305.12788, 2023.

[13] Feng, Y. et al., “Knowledge Graph-based Thought: a knowledge graph enhanced LLMs framework for pan-cancer question answering”, bioRxiv, pp. 1-10, April 2024.

[14] M. R. Karim et al., “From Large Language Models to Knowledge Graphs for Biomarker Discovery in Cancer”, ArXiv, abs/2310.08365, 2023.

[15] Y. Wang, Y. Zhao, and L. Petzold, “Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding”, ArXiv, abs/2304.05368, 2023.

[16] B. Li, T. Meng, X. Shi, J. Zhai, and T. Ruan, “MedDM: LLM-executable clinical guidance tree for clinical decision-making”, arXiv preprint arXiv:2312.02441, 2023.

[17] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, “Health-LLM: Large language models for health prediction via wearable sensor data”, arXiv preprint arXiv:2401.06866, 2024.

[18] I. L. Alberts et al., “Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?”, European journal of nuclear medicine and molecular imaging, 50(6), pp. 1549-1552, 2023.

[19] R. Poulain, H. Fayyaz, and R. Beheshti, “Bias patterns in the application of LLMs for clinical decision support: A comprehensive study”, arXiv preprint arXiv:2404.15149, 2024.

[20] Kaggle (Accessed: September 2024), Available at: https://www.kaggle.com/datasets/hamjashaikh/mental-health-detection-dataset/data

[21] Bioportal (Accessed: September 2024), Available at: https://bioportal.bioontology.org/ontologies/ONTRISCAL

[22] ADNI (Accessed: September 2024) Available at: https://adni.loni.usc.edu/

[23] SWRL, (Accessed: September 2024) Available at: https://www.w3.org/submissions/SWRL/

[24] SPARQL, (Accessed: September 2024) Available at: https://www.w3.org/TR/sparql11-query/