A Biomedical Ontology for Liver Diseases using UMLS as Data Source

Author

Kanwal Wahab

Reg. Number

00000205598

Supervisor

Dr. Usman Qamar

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

September 2019

A Biomedical Ontology for Liver Diseases using UMLS as Data Source

Kanwal Wahab

Reg. Number

00000205598

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Software Engineering

Thesis Supervisor:

Dr. Usman Qamar

Thesis Supervisor's Signature:

_____

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

July 2020

DECLARATION

I certify that this research work titled "A biomedical ontology for liver diseases using UMLS as data source" is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged / referred.

Signature of Student

Kanwal Wahab

MS – 17 – CSE

PLAGIARISM CERTIFICATE (TURNITIN REPORT)

This thesis has been checked for Plagiarism. “Turnitin” report endorsed by Supervisor is attached.

Signature of Student
Kanwal Wahab
00000205598

Signature of Supervisor
Dr. Usman Qamar

COPYRIGHT STATEMENT

ACKNOWLEDGEMENTS

I am extremely thankful to **ALLAH** Almighty for his bountiful blessings throughout this work. Indeed this would not have been possible without his substantial guidance through every step, and for putting me across people who could drive me though this work in a superlative manner. Indeed none be worthy of praise but the Almighty.

I am profusely thankful to my supervisor **Dr. Usman Qamar** for his generous help throughout my thesis, and for being available even for the pettiest of issues. My thanks for a meticulous evaluation of the thesis, and guidance on how to improve it in the best way possible.

I am profusely thankful to **Dr. Farhan Hussain** and **Dr. Wasi Haider Butt** for an excellent guidance throughout this journey and for being part of my evaluation committee.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

To my loving family and friends

Thank you for your constant support and encouragement.

ABSTRACT

The target of this research is creating a biomedical ontology for liver diseases using UMLS as data source. Due to extensive available information in the biomedical domain its unstructured format makes the access of that data to users difficult. Development in biomedical domain lies within its publications that is its literature. Ontologies help create structured knowledge base of the available information for easy access and use. In the recent years, ontology development has extended to desktops of experts from artificial intelligence. Ontologies now are common over the internet ranging from “Yahoo!” to “Amazon”. Resource description framework is developed by the WWW consortium to make the knowledge machine-readable. Protégé is tool developed by Stanford university to create and edit ontologies. It uses RDF triplets to model the schemas of an ontology by linking the concepts their properties and relations to concepts. The disease ontology is created using UMLS as data source.

Existing biomedical ontologies focus either on diseases and its hierarchy or symptoms only. No existing ontology draws the link between diseases and its symptoms. The existing ontologies has structured vocabulary like SNOWMED or has a semantic network like UMLS. The ontology created in this research focuses on liver diseases with hierarchy of the diseases and its semantic network based on UMLS concepts whereas the symptoms are based on the ontology created by Institute of Genome Sciences namely, “Symptom ontology”.

The relation between diseases and its respective symptoms are evaluated with the help of experts and expressed in graphical representation.

Key Words: *Ontologies, Unified Medical Library System, Resource Description Framework, Liver Disease, Protégé, Data Mapping, Object Based Data Access.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

CHAPTER 1: INTRODUCTION

Exponential growth of information particularly unstructured is leading to new influential technologies necessary to explore the available biomedical literature [1]. Smart-Health care system's such as medical diagnostics, decision support systems or health management are used to deduce information based on the available data using machine learning. An ontology is a machine understandable knowledge for a domain of interest [2].

Ontologies organize, integrate and helps derive useful conclusions with the information available after it is converted from unstructured information to structured data. The created data from ontology is in a format understandable by machines [3]. Due to the unstructured nature of biomedical data ontologies have become the most frequently explored technique to represent the available data [4]. Ontology is "Explicit specification of conceptualization" according to Gruber. Ontology represents a domain and the relation between concepts of that domain and their properties. These ontology forms the basis of semantic web [5]. Medical diagnostics related information is acquired from multiple sources and they consistency is the most important factor to contribution of accurate diagnosis. Many sources are considered in different papers among which clinical procedure guidelines (CPGs) are used to design for the usage of a doctor to make effective decision making according to the circumstances. Their use increases the quality and contributes to the effectiveness of the decisions providing a positive impact to quality of care and reduction in cost of the prescribed treatment [6]. This approach was first introduced in the western countries and its progress was monitored over the past two decades. Due to the positive progress they are still followed in those countries. These are usually in the form of paper booklet. Due to its paper format information review and retrieval of data from it is hard [7]. To identify all the information related to a specific problem these instructions need to be automated. Therefore, people are switching from booklets to automated procedures to overcome this boring and dry task of searching information within a book [8].

The conversion of this booklet's free text to terminologies and creating a vocabulary was the most important issue. As the data was about medical diseases and information was created by the people within that domain. These domain experts had the knowledge to make decisions based on their experience from the free text of the book in a very short time [9].

Additionally, the possibility of changing terminologies for a concept or same concept with multiple terminologies was the biggest test in automating the information. These were a number encountered difficulties in using CPGs as a source for clinical decision support systems [10].

The concept of ontologies comes in handy when the CPGs automation was ineffective due to its structured schemas as end-result. The ontologies are domain specific and creates objects on real world entities based called concepts. Properties are assigned to those concepts and their relationships with other concepts are defined, forming a structure for the information that was unstructured when started working with. To ensure data consistency and accuracy and already evaluated data source is used for the research in this thesis. UMLS integrates multiple medical knowledge bases. Liver diseases are our main focus in which the research will be conducted because liver disease are the most commonly occurring diseases in Pakistan as well as throughout the world.

Every year 2 billion people consume alcohol among which 75 million people suffer from alcohol induced liver diseases. Approximately 2 million deaths per year occur due to liver diseases. A million to liver cirrhosis complication, which is the 11^{th} leading cause of death worldwide and a million due to hepatocellular carcinoma as well as viral hepatitis. Liver cancer is the sixteenth leading cause of death globally. In total accounting for 3.5% of the worldwide death cause. Cirrhosis leading to disabilities weighs 1.6% making it to the top 20 causes of disabilities. 400 million have diabetes leading to overweighing lifestyle and obese which are caused by hepatocellular carcinoma and non-alcoholic fatty liver disease. Therefore, the focus of this research is liver disease [Arsani et, al. 2018].

1.1 Background, Scope and Motivation

The demand of software-based solutions has increased very rapidly since last two decades. As well as the biomedical literature has evolved drastically. It was initially thought that this huge amount of textual biomedical information could be proficiently controlled through automated textual extraction approaches. However, it soon became difficult to obtain the required results. The information available in bio-medical area is unstructured in publications. Thus, the problem cannot be simply solved by the use of the search engines. The absence of semantic association of available data on internet is the prime reason of incapability of the search engines. Ontology by definition means "explicit formal specifications of the terms in the domain and relations among them".

Ontology's scope includes several questions that it helps answers from a knowledgebase. The scope of our ontology deals with diseases, its symptoms, definitions, relations to other diseases and its synonyms. The relations consist of narrow and broad relations based on its hierarchy in the ontology. There exist particular ontologies that deal with either

diseases and its concepts or symptoms and its concepts but no existing ontology deals with both diseases and symptoms as well as their relations.

Due to lack of existing ontologies to depict the relation among diseases and its symptoms this topic has vast chances of improvement and research. The human body is a complex system and each organ in itself contains a lot of information with respect to medicine. Therefore, the organ under consideration for this thesis is “Liver” due to common occurrence of liver diseases in Pakistan.

1.2 Aims & Objectives

Major objectives of this research are as following:

- Extract data from UMLS
- Create a database for the extracted data
- Import the data into Protégé
- Create Liver disease ontology
- Merge Symptoms ontology and Liver diseases ontologies
- Create mapping for the data linking the diseases to its symptoms.
- OWL file containing the ontology and OBDA file containing the mappings of data into RDF triplets.

1.3 Structure of thesis:

Chapter 1: Includes an overview of the thesis and introduction to background knowledge and domain information, motivation for topic selection, goals of research and structure of this thesis. Chapter 2: includes the detailed systematic literature review whereas the Chapter 3 highlights the proposed methodology and implementation of our proposed idea. Chapter 4: includes a discussion of the achieved results and validation of these results from domain experts. Chapter 5: includes conclusion that concludes the thesis along with ideas for further improving the existing work under the heading future work.

Chapter 2

Systematic Literature Review

CHAPTER 2: LITERATURE REVIEW

2.1 Overview

Within the last two decades, the term “Ontology” originally used in philosophy, has gained extensive growth. This popularity is due to the use of ontologies for dealing difficulties of using computer for human purposes. Ontologies are found everywhere. They are trending these days in peer-to-peer systems, database integration, semantic web services, e-commerce, natural language processing, knowledge management and social networks. Managing ontologies and labelling them to associate semantics to the data throughout its life cycle. Ontology development in biomedicine has multiple ontologies. The biggest advantage is the domain knowledge in represented in a structured way creating a formal specification [12]. It helps understand the concepts specific to a domain and their relation to each other. Every concept is identified through an object and linked via properties to other object or values [13]. The available information for a domain is converted into a domain specific ontology and that in term becomes machine readable information [14]. Although ontology’s advantages out weigh’s its disadvantages, the biggest disadvantage is it’s development for a specific purpose or need through a specific person’s perspective. They can capture the domain information but can not generalize it for all activities in that domain [15]. Another disadvantage is their effectiveness. An ontology must be rigorously verified within the domain using the information of which it is created [16]. Assessing, verifying, validating and evaluating an ontology are the most important steps to ensure it’s correctness, integrity and preciseness [17]. A biomedical ontology is considered rich only if the information available is semantically correct and includes information from in dept that is sub domain linked to the specific domain are also encompassed within the ontology. For example is a disease is under consideration, the symptoms, synonyms, causative factors and risks should be included within the ontology domain [18]. The data source selection is the most important factor contributing to a consistent and accurate ontology. However evaluating the ontology from a domain expert is also very important in identifying the shortcomings within an ontology along with ensure its correctness and effectiveness [19]. UMLS (Unified Medical Language System) is known for its consistent and accurate data integrating multiple biomedical libraries like MeSH and SNOMED CT [20]. First order logic form the basis of ontologies that are declared, composed and context independent [21]. It consists of an object, relation, function triplet commonly known as

Resource Description Framework (RDF) triplet [22]. The RDF triplet is graph based model creating subject-predicate-object schemas for the relations defined for concepts [23].

Protégé toolkit is the user friendly for creating, editing and accessing ontologies. It is useful for understanding a domain easily through its visualization features [24]. CDSS based on ontology of infectious diseases and prescribing antibiotics based on machine learning techniques and AI models are developed [25]. A drug ontology is created for the sole purpose of obtaining detailed, accurate and consistent information related to drugs [26]. It is created on the drug terminology standard by national Library of Medicine (NLM) [27]. Another ontology is created namely human disease ontology which contains information related to diseases. They used web ontology language for its creation in protégé [28]. Ontology is vital way to represent domain specific information among users and researchers [30].

2.2 Ontology:

Ontology is a structural frameworks for organizing information to build a shared vocabulary that defines relations among concepts, its properties and types associated within a specific domain[31]. Ontology draws the domain in a realistic manner also capturing the restrictions set on a domain and model them. It specifies the domain knowledge and models them in a way to be used by individuals in that domain [32].

2.2.1 Components of an Ontology

Usually ontologies are defined through “Subject predicate object” which describes objects (instances), classes (concepts) and properties. Relations connecting concepts are also defined. The most common components of ontology are described below.

- **Individuals**: These are Instances or objects of a specific domain.
- **Classes**: These are collection of concepts dividing them into classes on basis of similarity and differences like in programming.
- **Properties**: Properties define aspects, features and characteristics of a specific nature associated with a domain.
- **Relationships**: Relationships are ways in which classes and individuals are connected to one another.
- **Axioms**: Assertions including rules in a logical form that comprising the overall ontology together.

The leading reasons for developing ontology is to reuse the information of that domain and share it within people related to that domain or software agents to use that information due to its machine readability. It makes explicit the assumptions and operational knowledge and separate the domain knowledge to analyze it.

2.2.2 OWL Language

Ontology based web language OWL is a recently developed language among standard languages for ontologies. It was developed according to WWWC standards and rules. The following are basic components of an OWL ontology:

- **Individuals:** Individuals are also known as instances. They represent objects of a domain that is under consideration in a specific ontology. They relate to properties and form a knowledge base.
- **Classes:** Classes are also called concepts, used describe a concept within that specific domain. OWL classes are used to differentiate between individuals on the basis of their characteristics. Similar individuals are grouped together. Their instances are created on the basis on their semantic classifications. Classes can have child classes or sub classes. Classes can also have parent classes or super classes.
- **Properties:** properties describe relations and roles within different objects. They are also called roles or slots. There are different types of properties most common among with are two types namely object properties and datatype properties.
    1. Object property:
       It relates instances of classes to each other. Describes a relation between different classes or concepts.
    2. Datatype property:
       It relates instances or individuals of OWL classes to its literal values.

Developing an ontology in practical terms includes creation of classes in a ontology and defining them on the basis of their hierarchy or taxonomical structure. Then properties are defined and values are assigned on the basis of datatype property or relations are established on the basis of object properties.

2.2.3 SPARQL Query:

SPARQL queries play an important role in usage of ontology. The ontology is built in a way that semantic association is built within the concepts and query answers questions with

respect to creating easy for decision support systems and evaluation systems for ontology. Using SPARQL queries we can know how well an ontology answers the questions corresponding it is queried for by the users. Using queries we could know how well the ontology is answering the question of the users. In this thesis SPARQL queries are used to query diseases related to liver their relations to each other and symptoms corresponding to each of the disease.

2.2.4 Resource Description Framework:

RDF (Resource Description Framework) provides a way to represent data in the form of triplets which in a standard model to interchange data on the web. A triplet defines a relation between two things for example “Dengue” has a symptom “Fever”. RDF triplet works on subject object predicate logic and SPARQL language is used to access these triplets which stands for “SPARQL Protocol And RDF Query Language”. Other languages like The Turtle language can also be used as alternative to SPARQL.

2.2.5 Protégé 5.5.0

Protégé is an open source free platform for ontology creation, editing and knowledge management along with supporting tools for OWL. The software is developed in JAVA and plugins are available to make the ontology accessible within applications for that specific domain[34]. Protégé allows creation and editing of ontologies for semantic web in correspondence to W3C standards for Web Ontology Language. It allows ontology visualization and graphical view. It allows users to load data to edit ontology as well as save newly created ontologies, it also provides reasoners to describe logic based classifiers.

2.3 Biomedical ontologies:

Prior towards the emergence of computer applications the enormous knowledge of ecology and biomedicine was stored in amorphous practice there were no doubts in the complexity of knowledge in unstructured form. Nonetheless it is astonishing that scholars who work on domain of biomedicine treaded and began locating ways to resolve difficulties and also started to epitomize information in systematized method. There are a portion of exertions in this domain and scientists schematize the biomedicine information as well as defined the touchstone concepts. The organization and calibration of terms in biomedicine have stemmed in the classification of diseases, repository control, thesaurus as well as vocabularies or ontologies.

The usage of ontologies in the biomedical domain is growing with time. For example a vocabulary of the biomedicine ontologies, contains more than three hundred ontologies as well as meticulous repositories in BioPortal2. Biologists used ontologies for dealing with the huge amount of data biomedicine ontologies. Other usage ares of biomedical ontologies are the hospitals as well as associated individuals use them for chronicling data approaching encounters in the clinics, also use in CDSS (Clinical Decision Support Systems) etc. Biomedicine scholars use ontologies concepts as well as KBs in boosted data amalgamation as well as discoveries in translation. Domineering flank of biomedical information discovery is the design of ontology as well as the reusability, which is of utmost apprehension which can be found by stage of compatibility amongst the ontology terms as well as amongst the schemes in the domain of biomedicine which they direct. Ontologies are castoff and they appear as empowering possessions in diverse biomedicine systems with the passage of time.

2.3.1 FMA:

FMA is an acronym for Foundation Model of Anatomy. The FMA was founded at Washington University and it rosette from previous exertion for supplement the structural substance of the UMLS. Emphasizing absolutely on the interpretation of structural organization, FMA gaze onward and appears as citation ontology. At start fringed for grossing the anatomy this anatomical ontology confines almost seventy thousand impressions stretched with cellular as well as sub cellular singularities. Protégé also has support of FMA.

2.3.2 OrphaData:

A scientific community providing comprehensive datasets with high quality relating to rare diseases and orphan drugs together forms “OrphaData” knowledge base. The data is in reusable formats, providing a structured vocabulary for rare diseases derived from orphaned knowledge base. It was established in France in 1997. It offers 33 different classifications of rare diseases each of which focuses on one organ system like rare cardiac diseases, rare neurological diseases.

2.3.3 Diseasome:

This is a dataset of diseases with genetic variation. It is a database constructed using an automating and integrating pipeline system. It contains data on 124,000 diseases, 12,000,000 single nucleotide polymorphism (SNP) markers, and about 38,500 genes, as well as more than 14,000 SNP records and 109,000 gene records associated with human diseases. This database

provides access to disease and its association studies through identification of candidate genes linked with disease and its specific genetic variation.

2.3.4 SNOMED CT:

It is an abbreviation of "Systematized Nomenclature of Medicine Clinical Terms". College of American Pathologists developed this biomedical ontological repository. It was built in instinctive explanation rationality formalism as well as it entails a huge amount of bio-medical terms and concepts. The latest version of SNOMED contains 269,864 terms as well as 407,510 names [50]. Now SNOMED-CT is accessible within UMLS knowledge sources without any cost. Only the requirement of signing the license covenant with the UMLS SNOMET-CT classes, concepts, names as well as pyramids are accessible consequently SNOMED-CT is probably consumed inside diverse bio-medical systems.

2.3.5 Unified Medical Language System (UMLS):

UMLS integrates biomedical standard domains and major vocabularies including SNOMED CT, MeSH, ICD and others. UMLS consists of three components a Metathesaurus, Semantic Network and Specialist lexicon. Metathesaurus is a vocabulary containing a million biomedically unique concepts linked to around 5 million labels derived from 100 terminologies. It further classified the relationship between these concepts roughly resulting into 17 million relationships. Each concept is associated with a unique identified called CUI that is concept unique identifier linking similar concepts from different ontologies and vocabularies. The Semantic Network categorizes the concepts in Metathesaurus and a semantic type is assigned to the concept. It also highlights the relations that can be used between the concepts through different semantic types. There are 133 semantic types such as disease or syndrome, anatomical structure, organism, clinical findings, symptoms and among the 133 semantic types 54 defines semantic relationships in the Semantic Network. Each concept from the Metathesaurus is assigned at least one semantic type.

2.4 Coverage of existing ontologies:

Ontology	**Scope**	**Number of Concepts**
SNOMED CT	Clinical medicine	310,400
FMA	Human Anatomical structure	73,000
GENE ONTOLOGY	Functional Annotation of Gene products	22,500

DISEASOME	Gene and its associated diseases	125,000
MESH	Biomedicine	12,350
UMLS	Terminology integration in life sciences	1.5 M

Table 1 "Existing ontologies and their respective domains"

The above tables lists few of the most commonly used ontologies around the world in biomedical domain. The table mentions the respective ontology's specific domain and the number of concepts covered by the ontology. The biggest of which is Unified Medical Library Systems (UMLS) in terms of coverage of concepts [33].

2.4.1 Existing Disease-Symptom Ontologies:

D-S relations are of the most importance in bioinformatics but the databases that catalog these relationships are incomplete in contrast to the available state of the art biomedical literature [Mohsen Hassan et al., 2015]. D-S relations provide details of diseases that can provide guidance to diagnosis systems for clinical care but the databases cataloging this relation are incomplete. Common among the databases that catalogs D-S relationship are "OrphaData" and "OMIM" [Köhler et al., 2014].

2.4.2 Symptoms Ontology (SYMP):

SYMP was founded by University of Maryland's institute of Genome Sciences in 2005. Currently it consists of 900 plus symptoms. Its hierarchy is categorized through broad headings like physical symptoms, digestive symptoms and so on and ending with root level symptoms related to the diseases from the broad headings.

2.5 Related Work:

[1] Cho-Tsan Bau and his colleagues came up with a decision support system for diabetic patients in a hospital setting through reasoning rules and surgery domain based ontology. Protégé was used as the editor to implement the ontology. It was based developing technique for ontology that included steps: conceptualizing, formulation, implementing, evolving and maintaining the created ontology. The reasoning rules were translated into JENA format and were used to recommend best procedures for a specific patient's clinical situation. It formed a hybrid system based on evidence from the ontology and evaluation through the decision support system.

[2] Yiqun Chen and his colleagues focused on the problem regarding data harmonization using ontology through expert validation. It states that ontologies should be created with reference to experts due to change in standard by crossing jurisdiction as well legacies and interpretation of the data. This research focuses on the data analytics and its necessity in standard ontology formation for the sake of reusability. Focusing on importance of experts involvement in defining concepts and linking them to each other.

[3] Wang Y. and his colleagues worked on drug-induced liver injury (DILI) to identify chronic and acute liver diseases that had no standard concepts. They identified drugs extracting them clinically and created their histopathological description by conducting biopsies of around 1080 samples. They used Metamap from UMLS to analyze these findings. And were mapped into SNOMED CT.

[4] Sadia Hafeez and her colleagues presented an ontology on the breast cancer domain with UMLS as data source. They focused on basic concepts and relations between different types and subtypes of breast cancer, with its coverage ranging between 14 main types and 92 subtypes of breast cancer. They targeted 101 broad relations and 254 more specific relations in their research work.

[5] Bureera and her colleagues created an ontology for urinal tract infection. Their research focused solely on development of an ontology with UMLS as data source. Their coverage included definition, relations, semantic types and synonyms. Their ontology's further evaluation was carried out but domain experts to measure its correctness and effectiveness.

[6] Reham Faisal Alhari and his colleagues worked decision support system based on ontology for diabetes diagnosis. This research resulted in a treatment plan on the basis of diabatic type of a patient on the basis of their personal information, risk factors, signs or symptoms and lab results in accordance to Clinical Practice Guidelines (CPGs). Their ontology is developed in OWL-DL and the rules are written in SWRL with JESS used for execution.

[7] T. H. Akila and her colleagues worked on Service Discovery on the basis of ontology for Intelligent Big Data Analytics. This ontology-based generation method focuses on automating by considering the planning stage to identify applicant services for abstract tasks in specific workflow.

[8] Osama Mohammed and his colleagues worked on the formation of a D-S ontology for medical diagnosis. Their algorithm focuses on merging existing disease and symptom ontology to link them formulating a core “diseases symptoms ontology (DSO)” scaling up to a number of symptoms and diseases. At present the research is only limited to a few diseases and their respective symptoms.

[9] Heiner Oberkampf and his colleagues worked on creating a missing link between symptoms and diseases. They built a model for D-S relations. By reusability of the existing ontology mappings, propagating the semantic type information of disease and symptom through ontologies. Similar semantic type diseases and symptoms were clustered and object properties are mapped for those clusters based on their relations.

[10] XueZhong Zhou and his colleagues worked on a human symptoms–disease network to study the relation between different diseases having similar symptoms and investigation into the connection between clinical occurrence of the diseases and their basic molecular connections. The data source used was PubMed bibliographic ontology.

[11] Le-thuy T. Tran and his colleagues worked on manipulating UMLS Metathesaurus for extraction and categorization of concepts showing signs and symptoms that automatically relate to a specific organ systems. The data source of this research is semantic type “Sign or Symptom” field present within the UMLS Metathesaurus.

[12] Mohsen Hassan and his colleagues worked on mining D-S relations from dependency graphs through syntactic pattern learning. It introduces a way to automate the mining of D-S relation from text known as SPARE. It was tested on a corpus consisting of abstracts from PubMed related to rare diseases.

[13] MICHAEL J. GROVE used data set UMLS and built an Ontology for Rehabilitation of Traumatic Brain Injury. This ontology contains data from UMLS to the domain of brain injury

[14] John D Osborne et.al castoffed Unified Medical Language System for discovering the genetic and the disease relations. They established the research paper containing the clarifications that the developed ontology has an evaluated accurateness in identifying the disease with the help of matching synonym.

[15] Marc Ehrig with his colleague worked on mapping ontologies in this paper. They determined similarity through rules that were encoded by ontology experts. Those rules were combined for one overall result.

Chapter 3

Proposed Methodology & Implementation

CHAPTER 3: PROPOSED METHODOLOGY & IMPLEMENTATION

The research aims to develop an ontology for liver diseases using UMLS as data source. The goal of this research is to link liver diseases to its respective symptoms and define relation among different diseases, definitions and identifying their synonyms available in multiple medical databases integrated within UMLS. This ontology uses RDF triplets to create links between concepts and define their relations and properties. Protégé is used for the creation and visualization of this ontology for liver diseases.

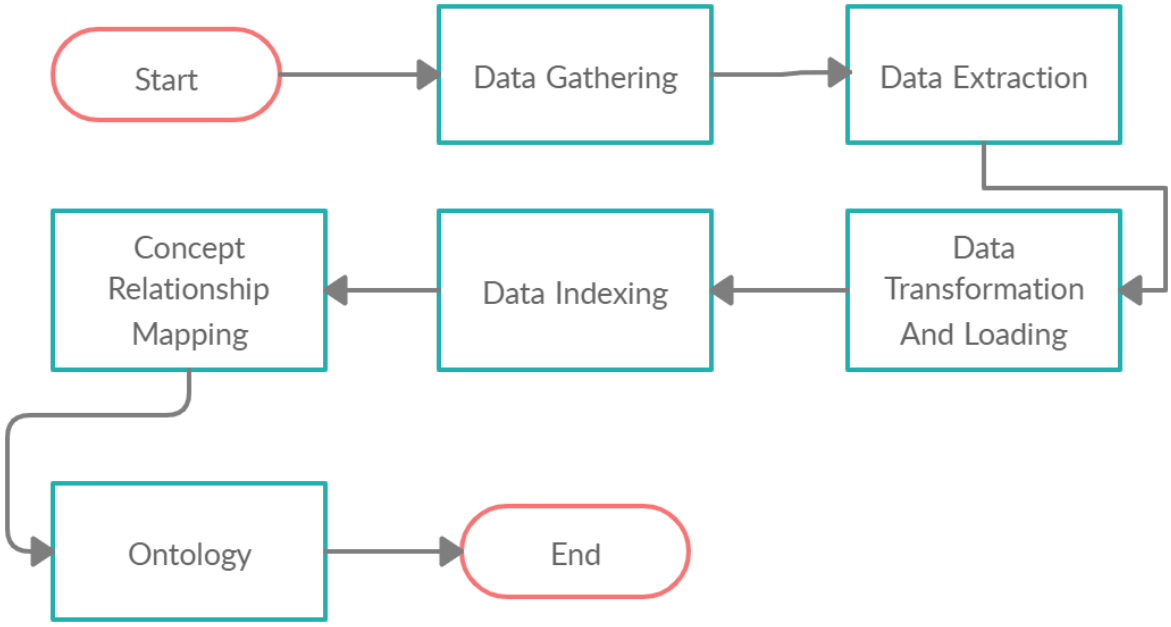


Figure 1 "Flow chart of the methodology"

The figure above shows the steps that should be followed in order to create an ontology. First the data is selected. Then the data is extracted from that data source and loaded into the database. The data is then indexed and loaded into a tool for further processing. After the data is loaded into ontology creating tool, its concepts are mapped into the classes and properties with respect to the ontology domain. The resultant ontology is formed in the form of OWL files and mappings in the form of OBDA files.

3.1 UMLS as data source:

The following steps are carried out in order to extract data from UMLS in MySQL database:

3.1.1 Data Gathering, Extraction & Loading

The data is extracted from UMLS repository and converted to Rich Release Format (RRF) files in the steps below:

1. Download zipped file of UMLS full release that include Metathesaurus, semantic network and lexical specialist.
2. Configure MetamorphoSys to extract data.
3. Select rich release format as output for relation Metathesaurus release format.
4. Copy database load scripts to hard drive for generation of textual files to load them into database.

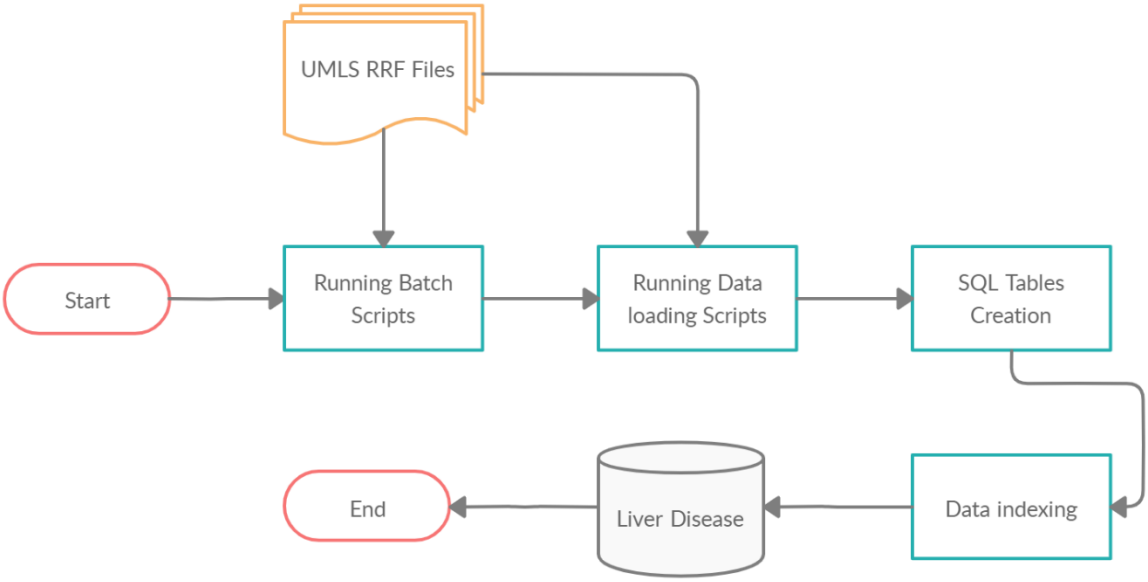


Figure 2 "Process of extracting and loading data into database"

The figure above explains the process of extracting data from UMLs and loading it into the database. The data is loaded into tables through batch scripts and indexed .

3.1.2 Data Indexing

Once the data is extracted from the UMLS files in rich release format and data is loaded into the database using batch scripts within these RRF files, the next step in to index this data. To index the database tables, SQL batch scripts available in the RRF files are executed. This increases the overall efficiency of the database and quick response of the tables to SQL queries.

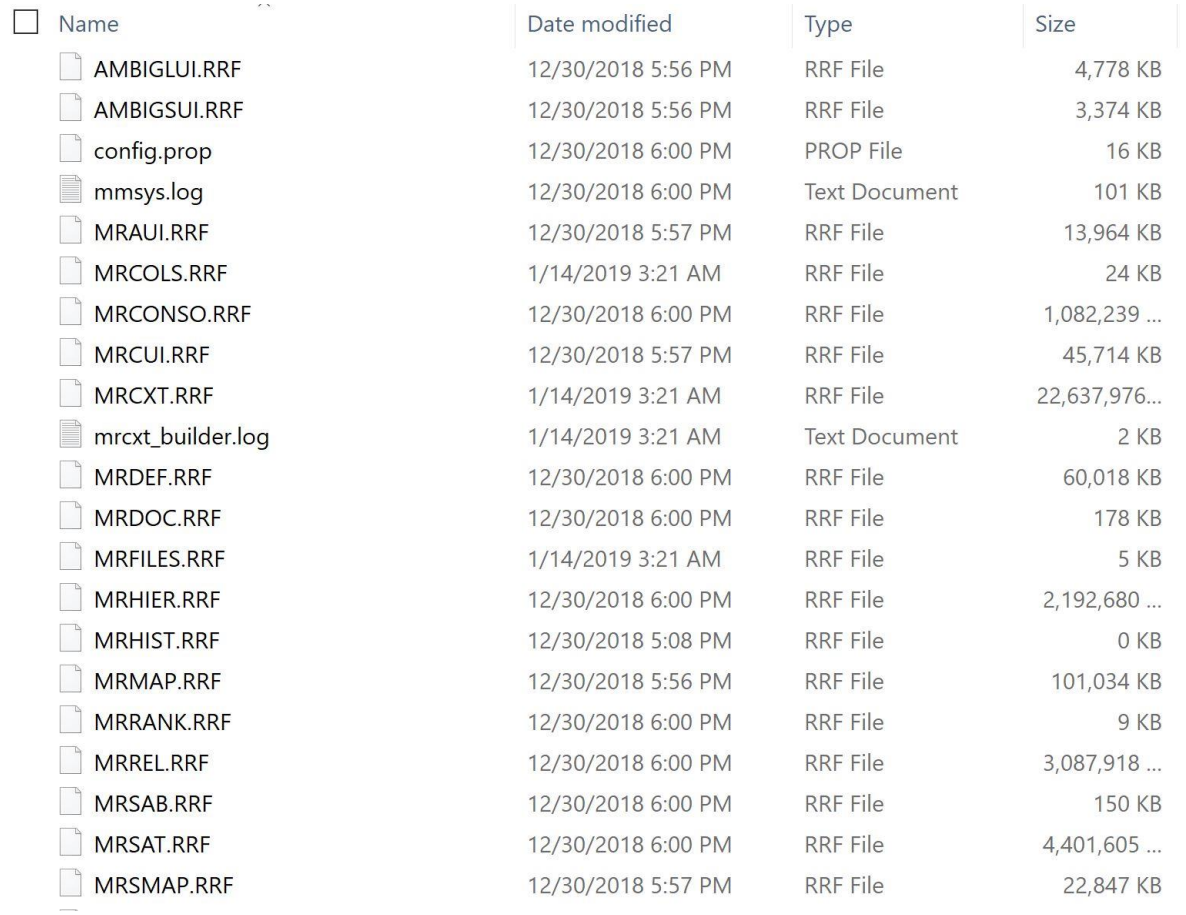
```
load data local infile 'LRAGR' into table LRAGR fields terminated by '|' lines terminat
create index LRAGR_EUI on LRAGR(EUI);
create index LRAGR_STR on LRAGR(STR);
create index LRAGR_SCA on LRAGR(SCA);
```

Figure 3 "Batch script for indexing"

The figure above shows a batch script for indexing the table “LRAGR”. It shows the indexing scripts for each attribute within the table.

3.1.3 Relationship Mapping & Ontology Creation

We use the concept unique identifiers (CUI) associated to each disease to query the large amount of data available within the UMLS. These CUIs are used to identify the relations, definitions and synonyms for the diseases but they are also used to find the semantic types for the concept. Semantic type classifies the concepts into categories depending on their characteristics. A concept with semantic type T033 is a sign/symptom, whereas a concept with semantic type T047 is a disease or syndrome. After the concepts are identified they are mapped according to the properties and relations to other concepts through queries within protégé. The mapping identifies all the concepts that are linked to liver diseases narrow or broad. The data in UMLS is divided into three different parts based on their format namely: Metathesaurus, semantic network and lexicon specialist. When the data is extracted from UMLS three directories are formed each containing RRF files of their own. The directories are META, NET and LEX.



Name	Date modified	Type	Size
AMBIGLUI.RRF	12/30/2018 5:56 PM	RRF File	4,778 KB
AMBIGSUI.RRF	12/30/2018 5:56 PM	RRF File	3,374 KB
config.prop	12/30/2018 6:00 PM	PROP File	16 KB
mmsys.log	12/30/2018 6:00 PM	Text Document	101 KB
MRAUI.RRF	12/30/2018 5:57 PM	RRF File	13,964 KB
MRCOLS.RRF	1/14/2019 3:21 AM	RRF File	24 KB
MRCONSO.RRF	12/30/2018 6:00 PM	RRF File	1,082,239 ...
MRCUI.RRF	12/30/2018 5:57 PM	RRF File	45,714 KB
MRCXT.RRF	1/14/2019 3:21 AM	RRF File	22,637,976...
mrcxt_builder.log	1/14/2019 3:21 AM	Text Document	2 KB
MRDEF.RRF	12/30/2018 6:00 PM	RRF File	60,018 KB
MRDOC.RRF	12/30/2018 6:00 PM	RRF File	178 KB
MRFILES.RRF	1/14/2019 3:21 AM	RRF File	5 KB
MRHIER.RRF	12/30/2018 6:00 PM	RRF File	2,192,680 ...
MRHIST.RRF	12/30/2018 5:08 PM	RRF File	0 KB
MRMAP.RRF	12/30/2018 5:56 PM	RRF File	101,034 KB
MRRANK.RRF	12/30/2018 6:00 PM	RRF File	9 KB
MRREL.RRF	12/30/2018 6:00 PM	RRF File	3,087,918 ...
MRSAB.RRF	12/30/2018 6:00 PM	RRF File	150 KB
MRSAT.RRF	12/30/2018 6:00 PM	RRF File	4,401,605 ...
MRSMAP.RRF	12/30/2018 5:57 PM	RRF File	22,847 KB

Figure 4 "List of Metathesaurus RRF files"

3.1.4 Data Access

Data is access into the ontology through user interface in protégé. The database is connected to protégé and queries are used for data retrieval based on the CUIs for liver diseases. The queries include retrieval of definitions, relations, symptoms and synonyms with the help of CUIs.

3.2 Importing data into Protégé and ontology creation

The steps to import data from database into protégé are as following:

3.2.1 Making connection of database and protégé:

The database is connected through a tab in the windows named ontop mapping. "**Window->Tabs->Ontop Mappings".** In the data source manager view values are assigned

to the parameters in order to create a working connection. Following attached screenshot shows a working connection of database to protégé.

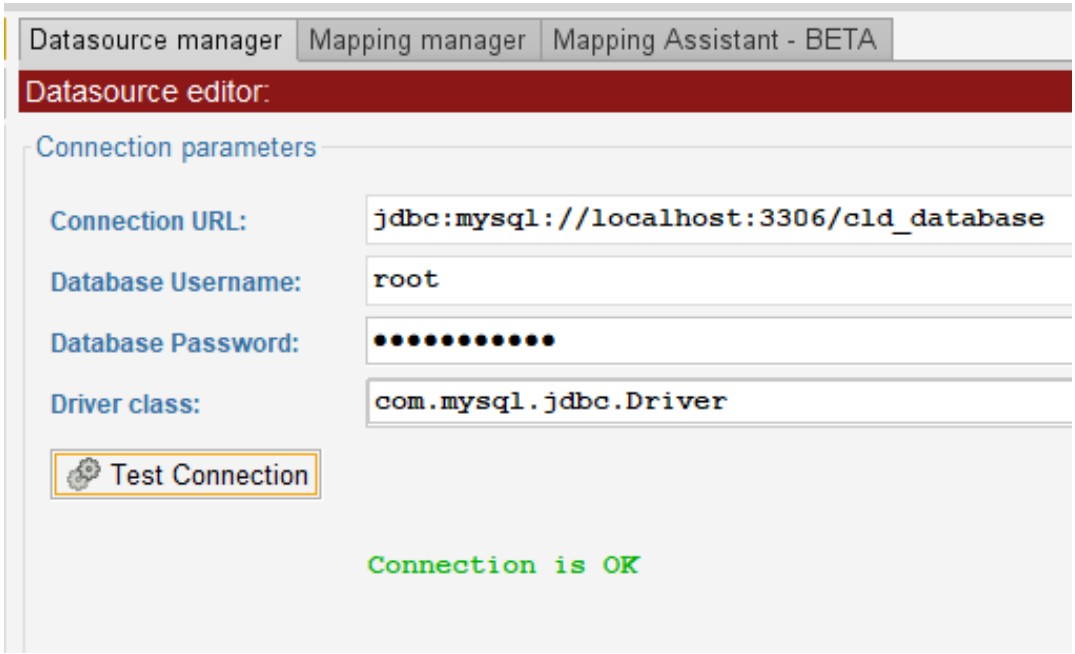
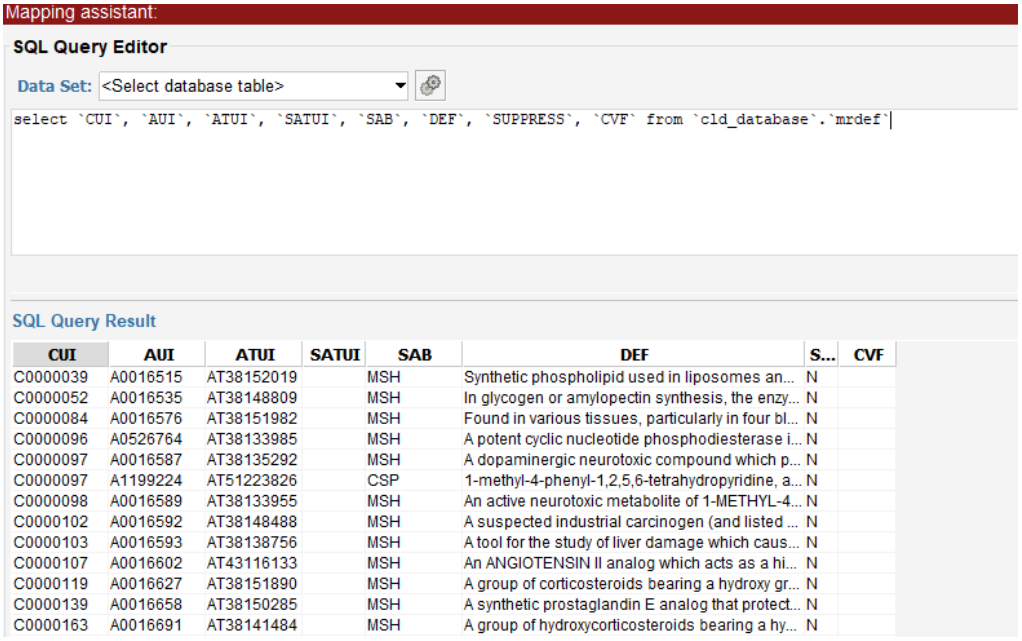


Figure 5 "Establishing connection to database in Protégé"

The above figure shows that a connection is established between the database named "cld_database" and protégé tool. A message is displayed after successful connection.

3.2.2 Querying database in protégé:

The data is access through queries by selecting the tables of database in the mapping assistant tab of ontop mappings. The fields required from the tables are selected and triplets are created according to usage of the data. The following screenshots explains the steps to data access from the database table "MRDEF" in the SQL query editor. Results for the query are displayed in SQL Query Result tab and it



CUI	AUI	ATUI	SATUI	SAB	DEF	S...	CVF
C0000039	A0016515	AT38152019		MSH	Synthetic phospholipid used in liposomes an...	N	
C0000052	A0016535	AT38148809		MSH	In glycogen or amylopectin synthesis, the enzy...	N	
C0000084	A0016576	AT38151982		MSH	Found in various tissues, particularly in four bl...	N	
C0000096	A0526764	AT38133985		MSH	A potent cyclic nucleotide phosphodiesterase i...	N	
C0000097	A0016587	AT38135292		MSH	A dopaminergic neurotoxic compound which p...	N	
C0000097	A1199224	AT51223826		CSP	1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine, a...	N	
C0000098	A0016589	AT38133955		MSH	An active neurotoxic metabolite of 1-METHYL-4...	N	
C0000102	A0016592	AT38148488		MSH	A suspected industrial carcinogen (and listed ...	N	
C0000103	A0016593	AT38138756		MSH	A tool for the study of liver damage which caus...	N	
C0000107	A0016602	AT43116133		MSH	An ANGIOTENSIN II analog which acts as a hi...	N	
C0000119	A0016627	AT38151890		MSH	A group of corticosteroids bearing a hydroxy gr...	N	
C0000139	A0016658	AT38150285		MSH	A synthetic prostaglandin E analog that protect...	N	
C0000163	A0016691	AT38141484		MSH	A group of hydroxycorticosteroids bearing a hy...	N	

Figure 6 "Querying database in protégé"

The above figure shows a query editor where database can be accessed via SQL queries.

3.2.3 Creating mapping triplets:

The resultant data from the query is converted into triplets in the mapping assistant tab. The fields are assigned values as per the screenshot attached.

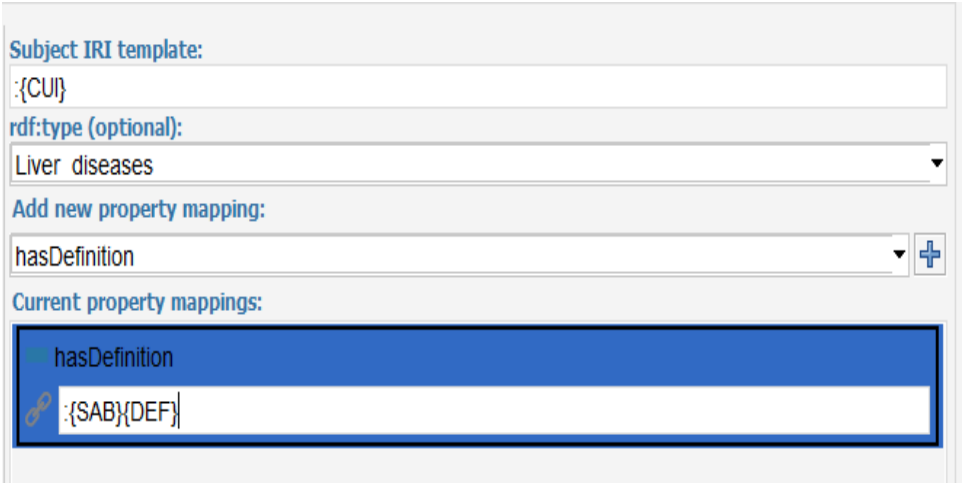


Figure 7 "Mapping data into RDF Triplets"

In the above figure the subject IRI template contains the CUI for the targeted disease. The rdf:type (optional) contains the targeted class. Add new property mappings contains the property and the fields that contains values to those properties. The created mappings are stored in "OBDA" file.

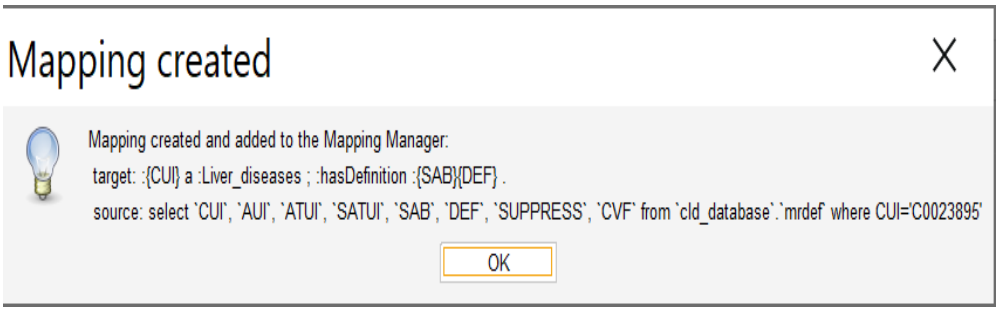


Figure 8 "Resultant RDF Mapping"

The figure above shows that a mapping is created successfully. In this mapping a triplet is created where the targeted field is CUI on basis of which the mapping was created whereas, a definition and the source of that definition are saved in the triplet corresponding to it. It also shows that the mapping belongs to liver diseases class.

Chapter 4

Results And Discussion

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Extracting data from UMLS data:

4.1.1 Connecting databases to extract data:

```
:: Database connection parameters
:: Please edit these variables to reflect your environment
::
set MYSQL_HOME=<C:\Program Files\MySQL\MySQL Server 5.6>
set user=<root>
set password=<flutus122A.>
set db_name=<cld_database>
```

Figure 9 "Assigning Values To Database Parameters"

The above figure shows the list of parameters that are edited in the batch script within the UMLS extracted data to create a connection to an already existing database.

4.1.2 Executing batch scripts to populate tables:

The database is populated with extracted files from the UMLS are 48.03 GB in size which is further populated in 72 tables in the database. The files are in Rich Release Format (RRF) files in three different directories namely LEX, META and NET.

LEX directory contains lexical specialist related files. The lexical tables populated onto the database tables are in the file named "mysql_lex_tables.sql". After executing the batch script 14 tables are created in the database.

```
125 •  DROP TABLE IF EXISTS MRDEF;
126 •  CREATE TABLE MRDEF (
127        CUI char(8) NOT NULL,
128        AUI varchar(9) NOT NULL,
129        ATUI    varchar(11) NOT NULL,
130        SATUI   varchar(50),
131        SAB varchar(40) NOT NULL,
132        DEF text NOT NULL,
133        SUPPRESS    char(1) NOT NULL,
134        CVF int unsigned
135    ) CHARACTER SET utf8;
136
137 •  load data local infile 'MRDEF.RRF' into table MRDEF fields terminated by '|' ESCAPED BY
138    (@cui,@aui,@atui,@satui,@sab,@def,@suppress,@cvf)
139    SET CUI = @cui,
140    AUI = @aui,
141    ATUI = @atui,
142    SATUI = NULLIF(@satui,''),
143    SAB = @sab,
144    DEF = @def,
145    SUPPRESS = @suppress,
146    CVF = NULLIF(@cvf,'');
```

Figure 10 "Batch script from mysql_tables.sql populating MRDEF table"

The above figure shows load scripts for creating tables in database. This specific batch script creates MRDEF table from MRDEF.RRF file and every attribute of the table is created and values are assigned from the file.

The META directory contains files related to the Metathesaurus. Loading and indexing of these files is through running the batch scripts within a file named "mysql_tables.sql" in the same directory. The tables created in the database are 51.

The NET directory contains files related to the semantic network. Loading and indexing of these files is through running the batch scripts within a file named "mysql_net_tables.sql" in the same directory. The tables created in the database are 7.

4.2 Classes and Class Hierarchy:

The following figure shows class hierarchy of liver diseases:

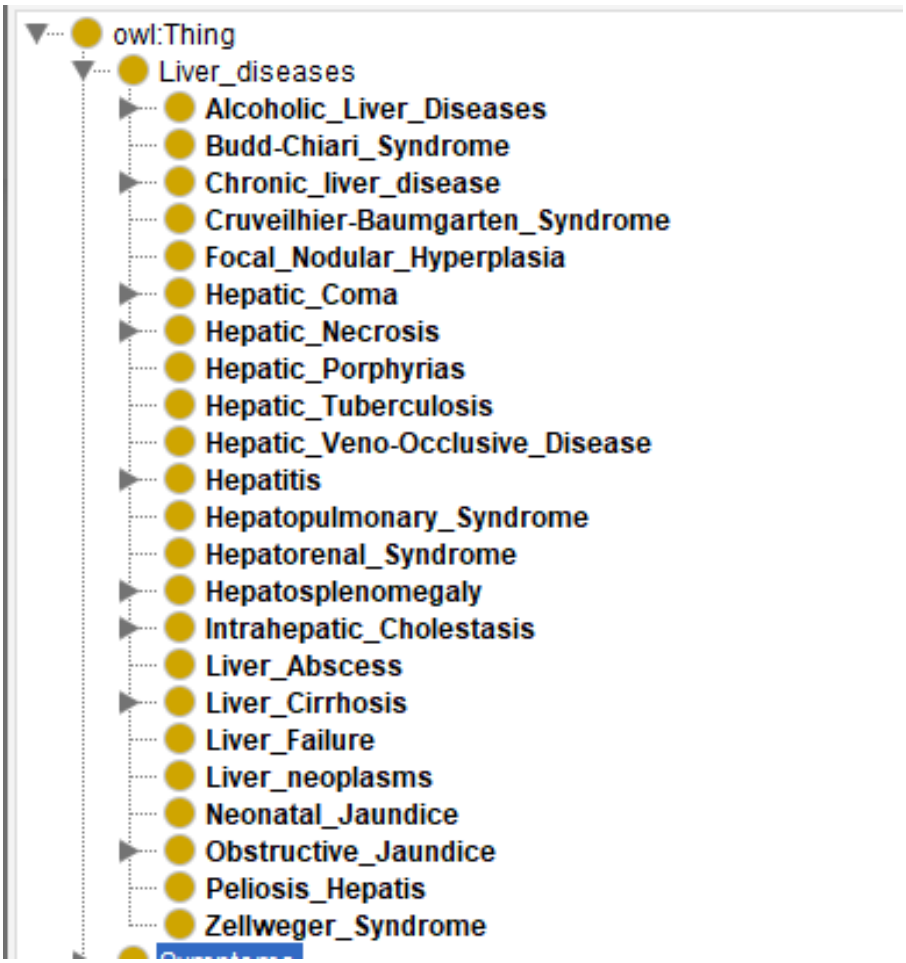


Figure 11 "Classes Liver Diseases"

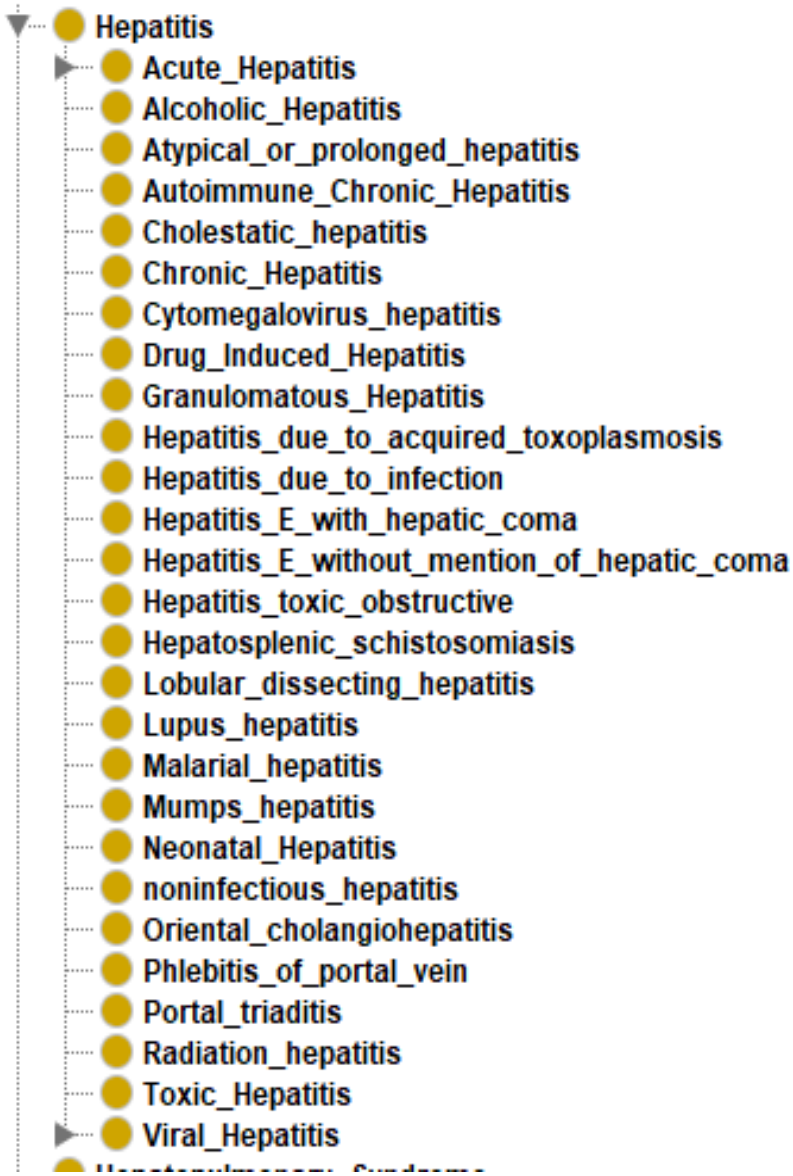


Figure 12 "Sub-Classes of Liver Diseases"

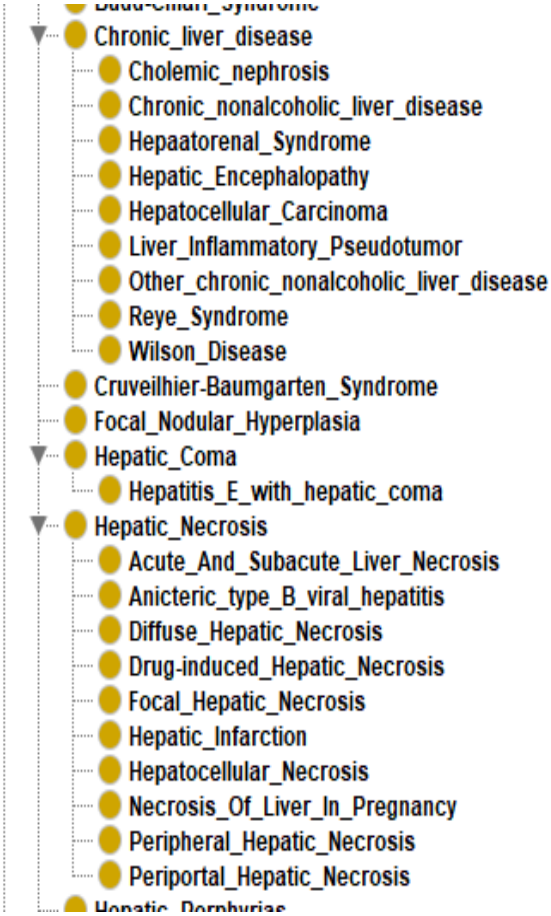


Figure 13 "Sub-Classes of Liver Disease"

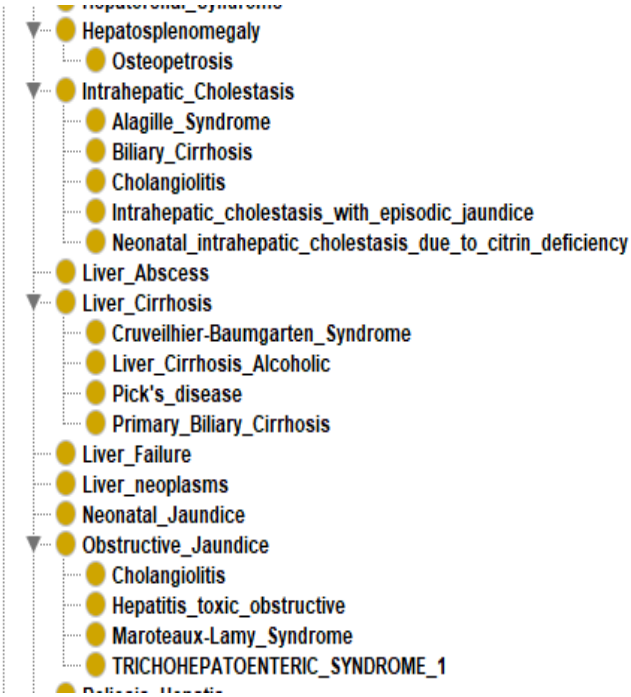


Figure 14 "Sub-Classes of Liver Disease"

The above figure explains the hierarchy of liver diseases and its subtypes. The diseases are classified based on their relation as child node or parent node to CUIs. A single disease can be classified into more than one class and therefore it can be added to multiple classes.

For symptoms connecting to specific liver disease focus in on two things mainly, first, a symptoms ontology containing all the symptoms and secondly, experts to link the diseases to the respective symptoms. The symptoms ontology of Institute for Genome Science, School of Medicine under University of Maryland is used in this thesis.

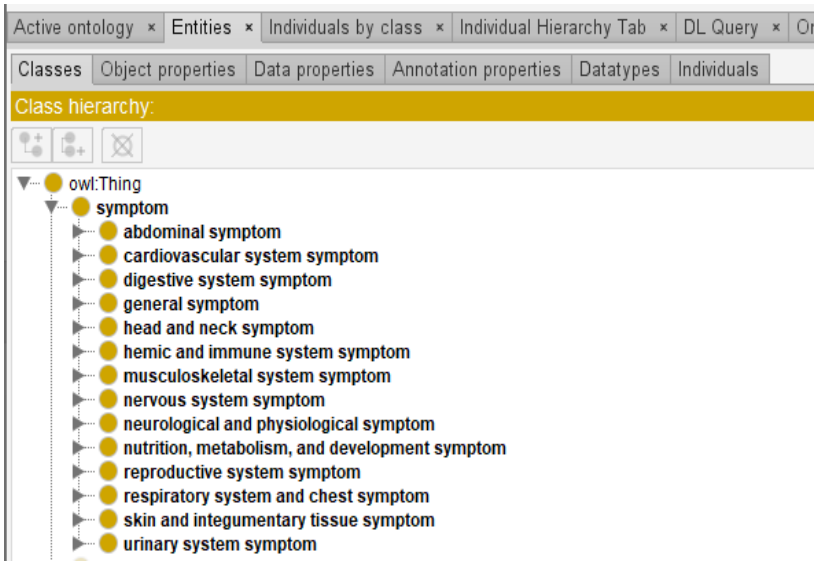


Figure 15 "Symptoms Ontology Generic View"

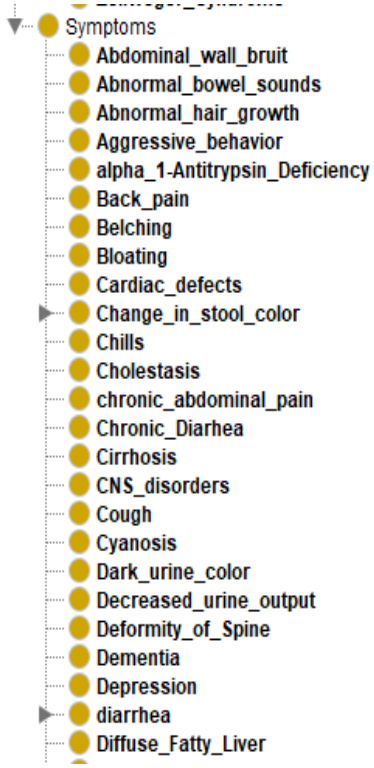


Figure 16 "Symptoms Sub-Classes"

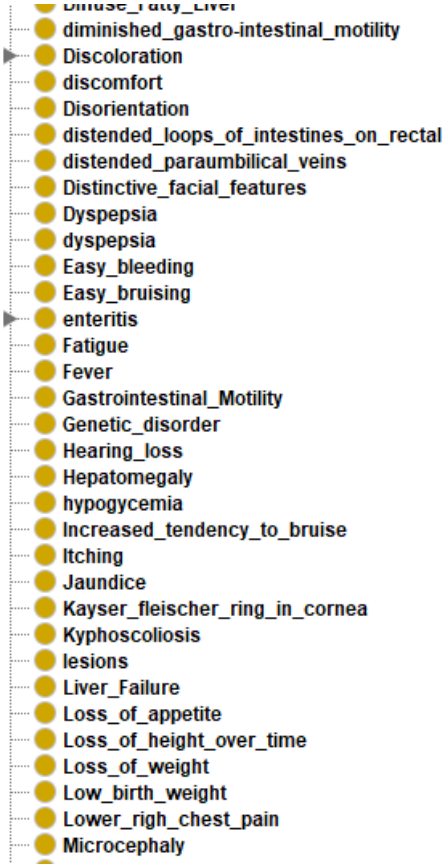


Figure 17 "Symptoms Sub-Classes"

Figure 18 "Symptoms Sub-Classes"

This figure above shows symptoms ontology hierarchy of the symptoms ontology created by institute of genome sciences in Maryland university. The symptoms are classified on the basis of human body systems and networks as well as general symptoms. The reason being its classification containing liver diseases symptoms as well as general symptoms. The hierarchy of general symptoms in the symptoms ontology. It shows the classes and subclass based on generic symptoms that are commonly occurred with other symptoms of a specific disease. This figure shows the existence of liver diseases symptoms in the disease ontology and their presence within the digestive system symptoms.

4.3 Merging ontologies:

Following are the steps for merging ontologies in Protégé. Second ontology being the symptoms ontology and first ontology be the liver disease ontology. The resultant ontology has liver diseases class as domain and symptoms class as rang.

1. Open Second Ontology (Ontology to be merged) in tool(Protégé 5.2.0).
2. Open First Ontology (Ontology in which merging has to be done) in the same window in Protégé 5.2.0.
3. Check for similarities.
4. Select 'Merge Ontologies' in Refactor Menu.
5. Then, select 'Merge into existing ontology' radio button.
6. Select First Ontology as Target Ontology.
7. Resolve inconsistencies in Resultant Merged Ontology by changing the Full URIs of conflicting classes and individuals.
8. Check for consistency of Resultant Merged Ontology using a reasoner.

Figure 19 "Steps for merging ontologies"

The figure above shows steps to follow in protégé to merge to existing ontologies and their resultant ontology is a merged ontology with a single file. The OBDA files for both ontologies must be present.

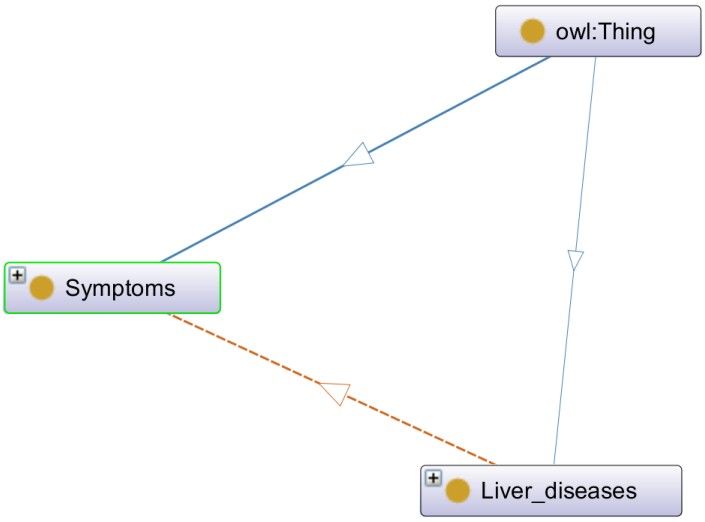


Figure 20 "Merged Liver Ontology and Symptoms Ontology"

The figure above shows a resultant ontology after the merge. The red line shows that symptoms class has liver diseases class as its domain.

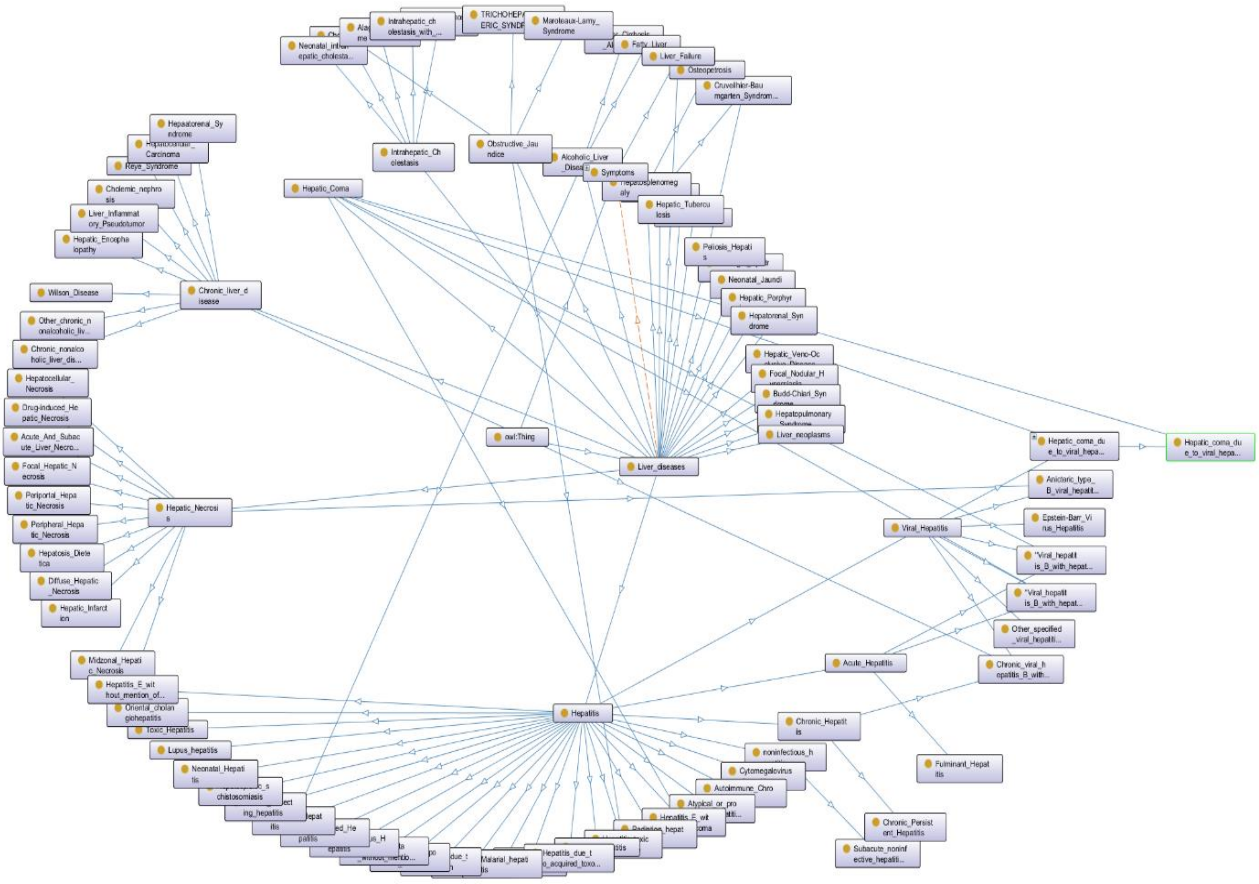


Figure 21 "Liver Diseases and its Subtypes"

The figure above shows classes and subclasses of liver diseases and their relation to each other. The overlapping lines show that the subclass belong to more than one class.

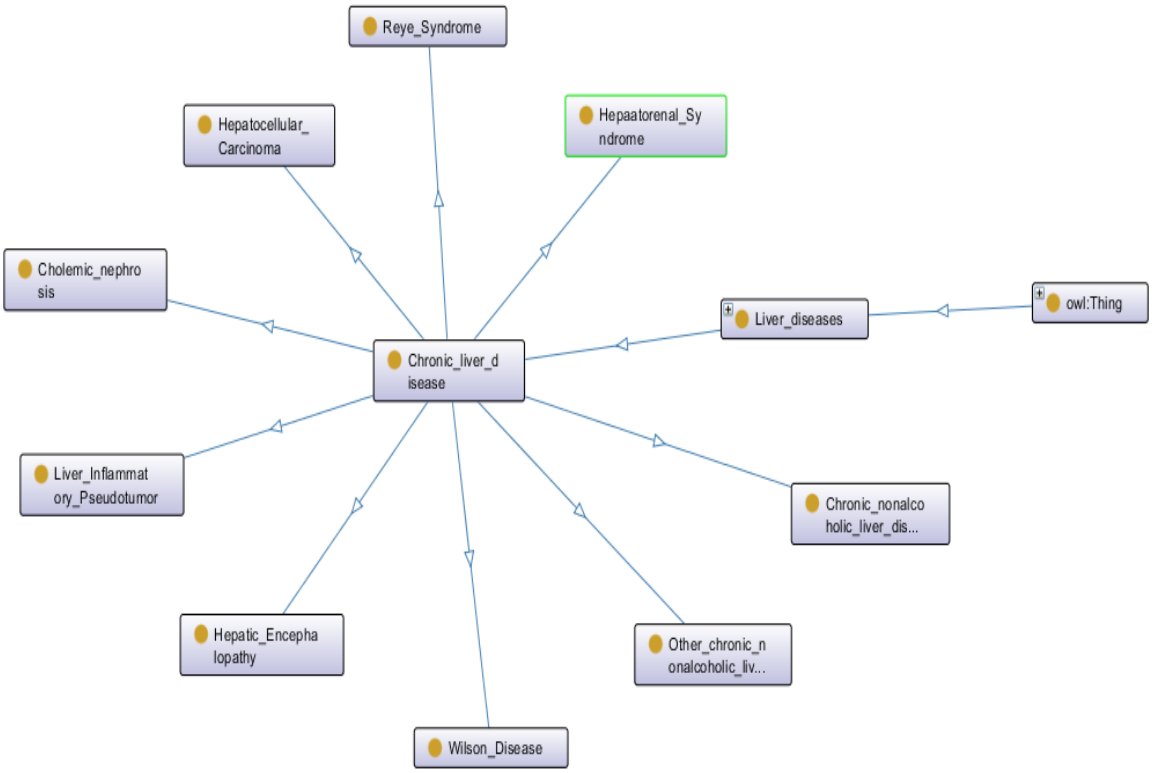


Figure 22 "Chronic Liver Diseases"

The figure above shows a view of the chronic liver diseases that is a subclass of the liver diseases ontology. It shows the diseases that can be classified as chronic liver diseases.

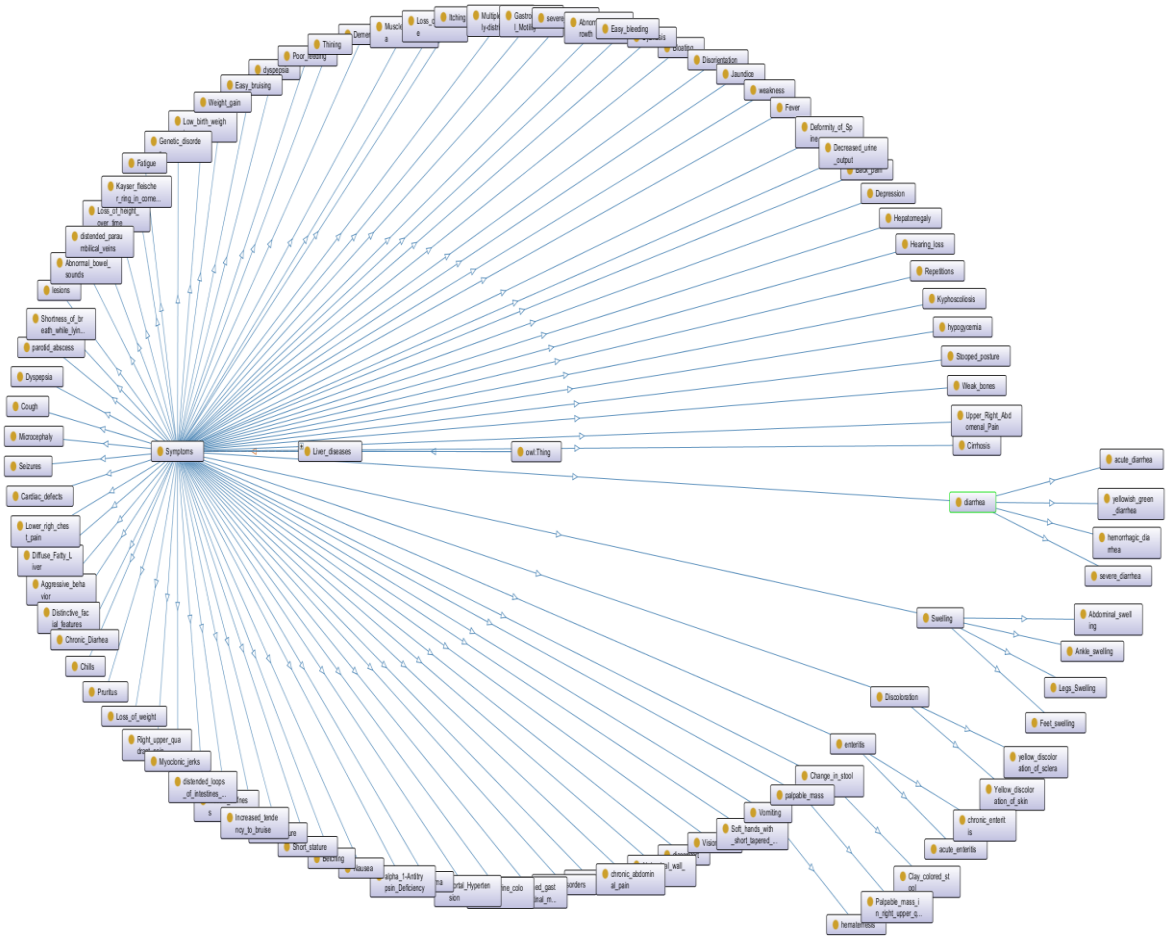


Figure 23 "Symptoms for Liver Diseases"

The figure above shows symptoms general as well as liver diseases that are associated with the 88 types and subtypes of liver diseases. These symptoms are divided into categories based on their nature.

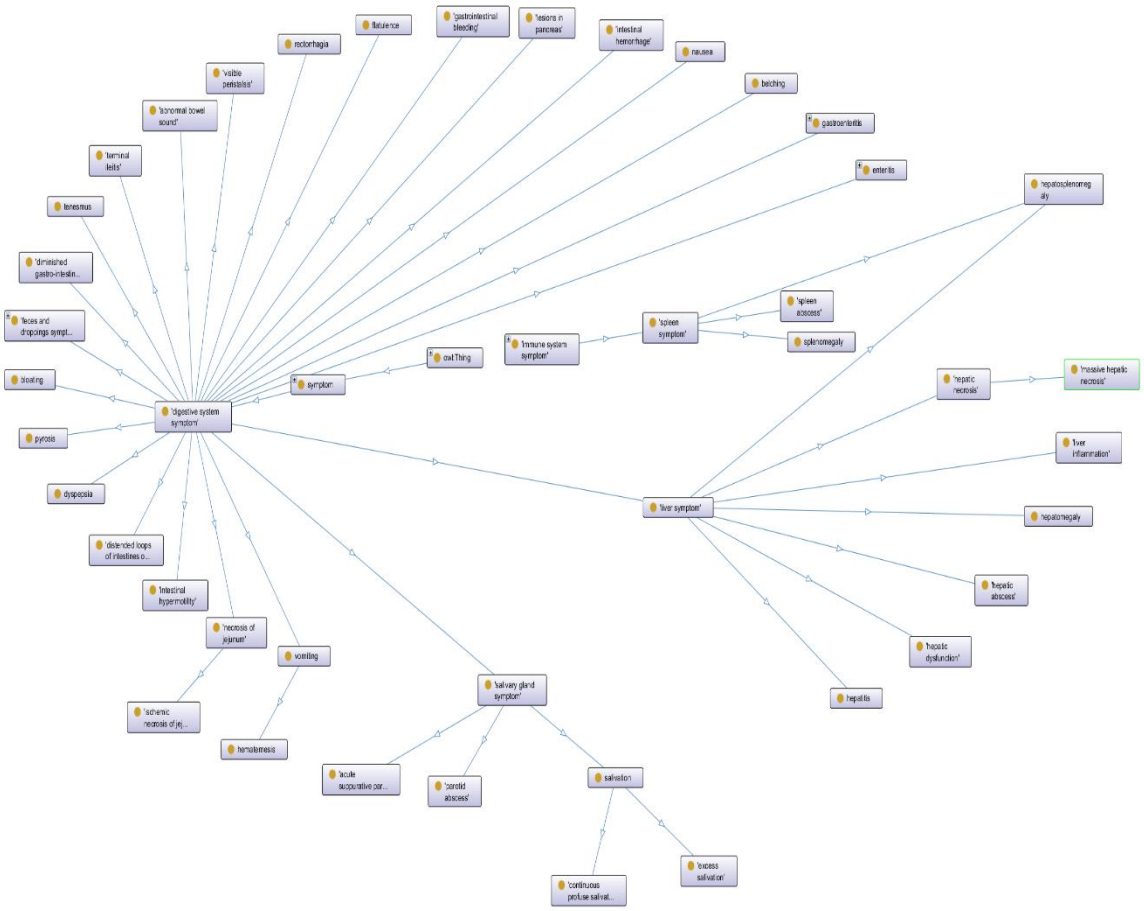


Figure 24 "Symptoms only associated with Liver Diseases"

The figure above shows symptoms that are only associated with liver diseases and not linked to diseases of other human bodies apart from liver.

4.4 Coverage:

The below queries specify coverage of the ontology by highlighting the relations between liver diseases among themselves as well as to other diseases. Narrow relations refer to children of that node, parent of that node and siblings of the node. Where as broad relations refer to extended nodes within the tree.

4.4.1 Broad Relations:

The query below when executed results in records of data based on attribute REL that is relation and value is set to RB that is broad relation.

```
SELECT `CUI1`, `AUI1`, `STYPE1`, `REL`, `CUI2` from `cld_database`.`mrrel` WHERE CUI1='C0001308' AND REL='RB'
```

4.4.2 Narrow Relations:

The query below when executed results in records of data based on attribute REL that is relation and value is set to RN that is narrow relation.

SELECT `CUI1`, `AUI1`, `STYPE1`, `REL`, `CUI2` from `cld_database`.`mrrel` WHERE CUI1='C0001308' AND REL='RN'

4.5 SQL Queries and their mappings:

The SQL queries mentioned below refers to object properties of the classes.

4.5.1 Query for Definition:

The query below after execution results in a list based on the CUIs. It is used to retrieve definitions for a specific disease as well as the source of that definition.

SELECT `CUI`, `SATUI`, `SAB`, `DEF` from `cld_database`.`mrdef` WHERE CUI=' C0010398'

4.5.2 Query for Synonyms:

The query below after execution results in a list based on the CUIs. It is used to retrieve synonyms based on that concept unique identifier.

SELECT DISTINCT `LAT`, `NSTR`, `CUI`, `LUI`, `SUI` from `cld_database`.`mrxns_eng` WHERE CUI=' C0010398'

4.5.3 Mapping corresponding to definition query:

```
mappingId   MAPID-9f06b5ce747f424ea0bf8a818541b13b
target      :{CUI} a :Cruveilhier-Baumgarten_Syndrome ; :hasDefinition :{DEF} .
source      select `CUI`, `DEF` from `cld_database`.`mrdef` where CUI='C0010398'
```

4.5.4 Mapping corresponding to synonyms query:

```
mappingId   MAPID-c77fcce459e046b093fc39257876ba2d
target      :{CUI} a :Cruveilhier-Baumgarten_Syndrome ; :hasSynonyms :{NSTR} .
source      select distinct `NSTR`, `CUI` from `cld_database`.`mrxns_eng` where
CUI='C0010398'
```

4.5.5 Mapping corresponding to relation query:

```
mappingId   MAPID-06d614dc1d08401cb86a91673ad0a9b2
target      :{CUI1} a :Cruveilhier-Baumgarten_Syndrome ; :hasRelationType
:{REL}{CUI2} .
source      select `CUI1`, `REL`, `CUI2` from `cld_database`.`mrrel` where
CUI1='C0010398'
```

4.5.6 Mapping for Relation type Broad:

```
mappingId   MAPID-362092171bb34577babf154f4b3b7d4c
target      :{CUI1} a :Cruveilhier-Baumgarten_Syndrome ; :has_Relation_Broad
:{REL}{CUI2} .
source      select `CUI1`, `AUI1`, `STYPE1`, `REL`, `CUI2`, `AUI2`, `STYPE2`,
`RELA`, `RUI`, `SRUI`, `SAB`, `SL`, `RG`, `DIR`, `SUPPRESS`, `CVF` from
`cld_database`.`mrrel` where CUI1='C0010398' and REL='RB'
```

4.5.7 Mapping for Relation type Narrow:

```
mappingId   MAPID-efe58a2bbbbe45de91b3550c813cb221
target      :{CUI1} a :Cruveilhier-Baumgarten_Syndrome ; :has_Relation_Narrow
:{REL}{CUI2} .
source      select `CUI1`, `AUI1`, `STYPE1`, `REL`, `CUI2`, `AUI2`, `STYPE2`,
`RELA`, `RUI`, `SRUI`, `SAB`, `SL`, `RG`, `DIR`, `SUPPRESS`, `CVF` from
`cld_database`.`mrrel` where CUI1='C0010398' and REL='RN'
```

4.6 OWL Files

4.6.1 Owl file class for Cruveilhier-Baumgarten Syndrome:

```
<!-- http://www.semanticweb.org/kanwalwahab/
ontologies/2019/6/untitled-ontology-31#Cruveilhier-Baumgarten_Syndrome -->

<owl:Class rdf:about="http://www.semanticweb.org/kanwalwahab/
ontologies/2019/6/untitled-ontology-31#Cruveilhier-Baumgarten_Syndrome">
    <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Liver_Cirrhosis"/>
    <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Liver_diseases"/>
    <has_Symptoms rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Abdominal_wall_bruit"/>
    <has_Symptoms rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Portal_Hypertension"/>
    <has_Symptoms rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#distended_paraumbilical_veins"/>
</owl:Class>
```

4.6.2 Owl file class for Alagille Syndrome:

```
<!-- http://www.semanticweb.org/kanwalwahab/
ontologies/2019/6/untitled-ontology-31#Alagille_Syndrome -->

<owl:Class rdf:about="http://www.semanticweb.org/kanwalwahab/
ontologies/2019/6/untitled-ontology-31#Alagille_Syndrome">
    <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Intrahepatic_Cholestasis"/>
    <has_Symptoms rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Cardiac_defects"/>
    <has_Symptoms rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Cholestasis"/>
    <has_Symptoms rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Deformity_of_Spine"/>
    <has_Symptoms rdf:resource="http://www.semanticweb.org/kanwalwahab/
    ontologies/2019/6/untitled-ontology-31#Xanthoma"/>
</owl:Class>
```

4.7 Discussion

The developed ontology can be used as a backend data source for clinical decision support systems. Moreover, a rule based expert system on liver diseases can be developed from this ontology. In addition to this, a complete machine-readable encyclopedia can be developed which can be queried in a very accurate and easy-to-read way. In consideration, of the availability of a few reliable sources other than UMLS we could have gone a bit further and would have extended the coverage of our ontology, but effective knowledge modelling of our ontology was preferred over increasing the sample-space and incorporating the data from these other sources into our well modelled base is now left as a future work.

Concept Name	Concept Identifier	Concept Relations
Liver Disease	C0023895	47
Alcoholic Liver Disease	C0023896	20
Alcoholic Hepatitis	C0019187	16
Fatty Liver	C0015695	18
Liver Cirrhosis Alcoholic	C0023891	13
Budd-Chiari Syndrome	C0856761	11
Chronic Liver Disease	C0341439	13
Cholemic Nephrosis	C0699858	10
Chronic Nonalcoholic Liver Disease	C0221388	4
Hepatorenal Syndrome	C0019212	7
Hepatic Encephalopathy	C0019151	14
Hepatocellular Carcinoma	C2239176	15

Other Chronic Nonalcoholic Diseases	C0029546	9
Reye Syndrome	C0035400	9
Wilson Disease	C0019202	11
Cruveilhier-Baumgarten syndrome	C0010398	2
Focal Nodular Hyperplasia	C0333980	2
Hepatic Coma	C0019147	10
Hepatic Necrosis	C0151798	8
Acute and Subacute Liver Necrosis	C0001308	10
Anicteric Type B Viral Hepatitis	C0276615	3
Diffuse Hepatic Necrosis	C0267794	1
Focal Hepatic Necrosis	C0267800	2
Hepatic Infarction	C0151731	8
Necrosis Of Liver In Pregnancy	C0269676	2
Peripheral Hepatic Necrosis	C0267802	3
Periportal Hepatic Necrosis	C0546389	1
Hepatic Porphyria	C0162533	4
Hepatic Tuberculoses	C0041313	2
Hepatic Veno-Occlusive Disease	C0019156	5
Hepatitis	C0019158	55
Acute Hepatitis	C0267797	7
Fulminant Hepatitis	C0302809	3
Alcoholic Hepatitis	C0019187	16
Cholestatic Hepatitis	C0149904	8
Chronic Hepatitis	C0019189	19
Cytomegalovirus Hepatitis	C0276252	2
Drug Induced Hepatitis	C1262760	1
Granulomatous Hepatitis	C0235369	2
Hepatitis Due To Acquired Toxoplasmosis	C0276802	1
Hepatitis Due To Infection	C0400894	4
Hepatitis E Without Mention Of Hepatic Coma	C0153088	3
Hepatitis Toxic Obstructive	C0235373	3
Lobular Dissecting Hepatitis	C0267828	1
Lupus Hepatitis	C0267807	1
Malarial Hepatitis	C0276835	2
Mumps Hepatitis	C0153096	1
Neonatal Hepatitis	C0027613	4
Noninfectious hepatitis	C0679412	4
Oriental Cholangiohepatitis	C0267921	2
Phlebitis Of Portal Vein	C0034192	2
Portal Triaditis	C0040860	1
Radiation Hepatitis	C0267793	2

Toxic Hepatitis	C0019193	5
Viral Hepatitis	C0042721	26
Epstein-Barr Virus Hepatitis	C0554114	1
Hepatosplenomegaly	C0019214	6
Osteopetrosis	C0029454	20
Intrahepatic Cholestasis	C0008372	8
Alagille Syndrome	C0085280	4
Biliary Cirrhosis	C0023892	7
Cholangiolitis	C0008308	1
Liver Abscess	C0023885	3
Liver Failure	C0085605	21
Liver Neoplasms	C0023903	16
Neonatal Jaundice	C0022353	11
Maroteaux-Lamy Syndrome	C0268218	3
Trichohepatoenteric Syndrome	C1857276	1
Peliosis Hepatis	C0030781	4
Zellweger Syndrome	C0043459	9
Total count of related concepts		**570**

Table 2 "Liver Diseases, their Unique Identifiers and number of related concepts"

The created ontology comprises facts on relations and semantic types for 88 categories of liver diseases including 31 main types and 57 subtypes. The number of related concepts to these 88 categories are 570.

4.7.1 Comparison to existing knowledge base

KB	Domain	Source	Size	Semantic Relations	Causative factors of diseases
Proposed Ontology	Liver Diseases, Definitions, Relations, Synonyms and Symptoms	UMLS	88 types of liver diseases and 570 related diseases concepts	Included	Included
OMIM	genes, genetic disorders, including phenotype description and body parts	Manually generated by scientists and physicians	~18,597 genes	Contains relations of specialized context, e.g. relations between genotype and drug response phenotype	Contains causes of specified perspective, e.g. contains genetic factors which can

					contribute to a disease
MeSH	General medical subjects for indexing articles for PubMed database	PubMed publications	~25,186 entities	Not included	Included as a separate entry term
Diseasome	Human disease network	OMIM	>4,213 diseases, >91,182 genes	Contains relations of specialized context	Not included
FMA	Human anatomy	UMLS	120,000 concepts	Included	Not included

Table 3 "Comparison between existing Knowledge Bases"

The above table lists different ontologies in biomedical domain. Their domains are specified and a detail of their coverage is mentioned. The basis on which they were created are listed under the data source category. The semantic coverage along with causative factors are the key factors to determine their quality apart from coverage.

We have also carried out a comparison between existing liver disease ontologies. The results are shown in the table below.

Ontology	Data Source	Definition	Synonyms	Relations	Symptoms
Our Ontology	UMLS	✓	✓	✓	✓
Ontology of liver for radiology	UMLS	×	×	✓	×
Disease Ontology	MeSH	×	✓	✓	×
Symptoms Ontology	NCBI	×	×	×	✓
Human Disease ontology	NCBI	✓	✓	✓	×

Table 4 "Comparison between our ontology and other existing liver ontologies"

The table above shows a comparison between existing liver ontologies and our liver disease ontology. The stated ontologies for comparison lack disease symptom relation. The focus is entirely on diseases or entirely on symptoms. We as our ontology not only

encompasses the relation between symptoms and diseases it also states proper relation of liver diseases to other diseases.

Chapter 6

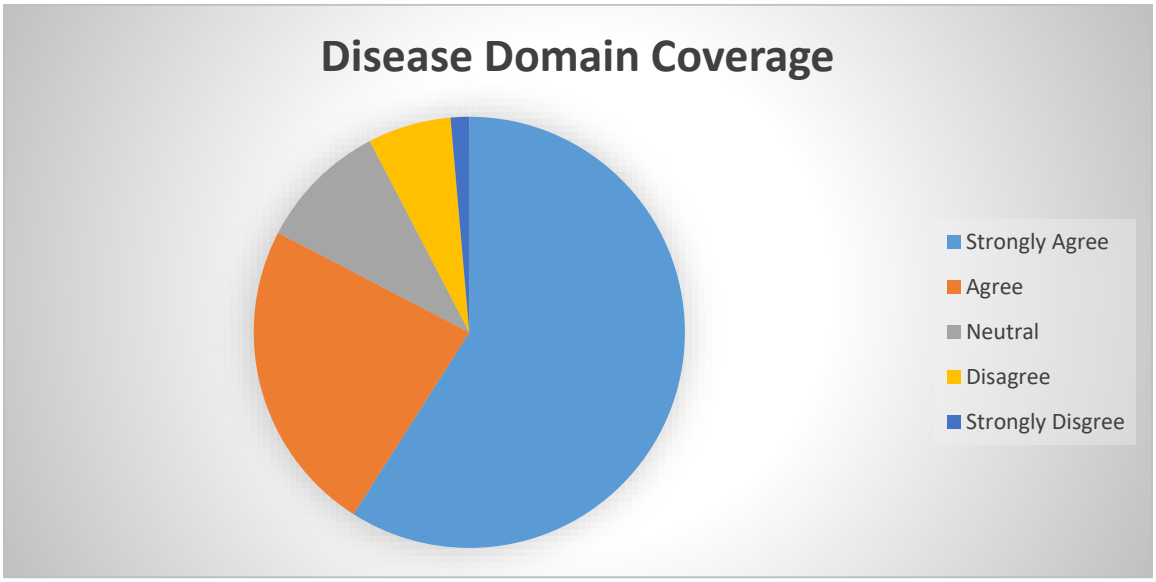
Ontology Evaluation

CHAPTER 5: ONTOLOGY EVALUATION

5.1 Expert evaluation:

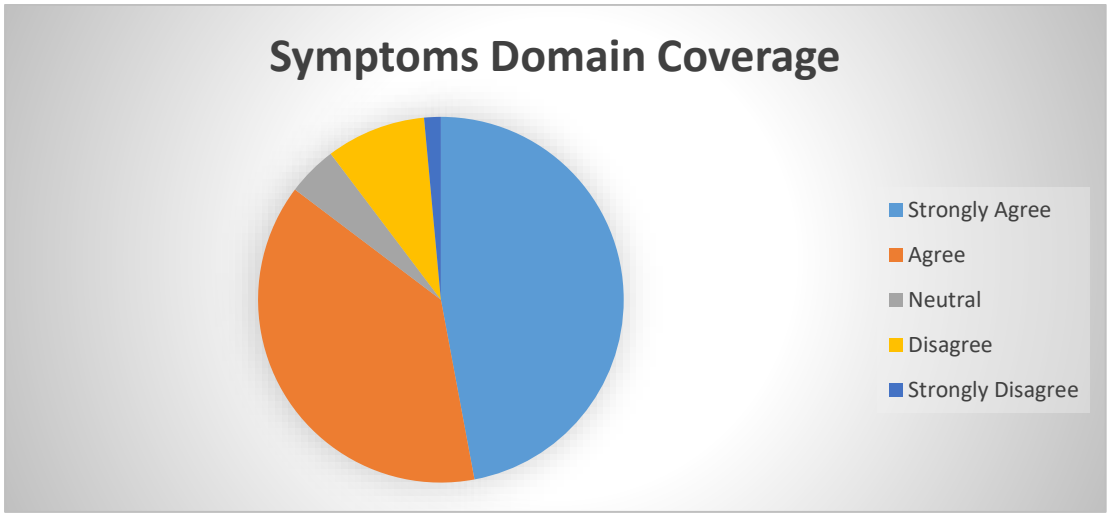
The 5-point like scale is used to evaluated expert opinion about the newly developed ontology. A group of 10 general physicians and 5 internal medicine specialists evaluated the ontology. This survey covered the domain coverage as well as the correctness of the ontology.

5.1.1 Disease Domain Coverage:



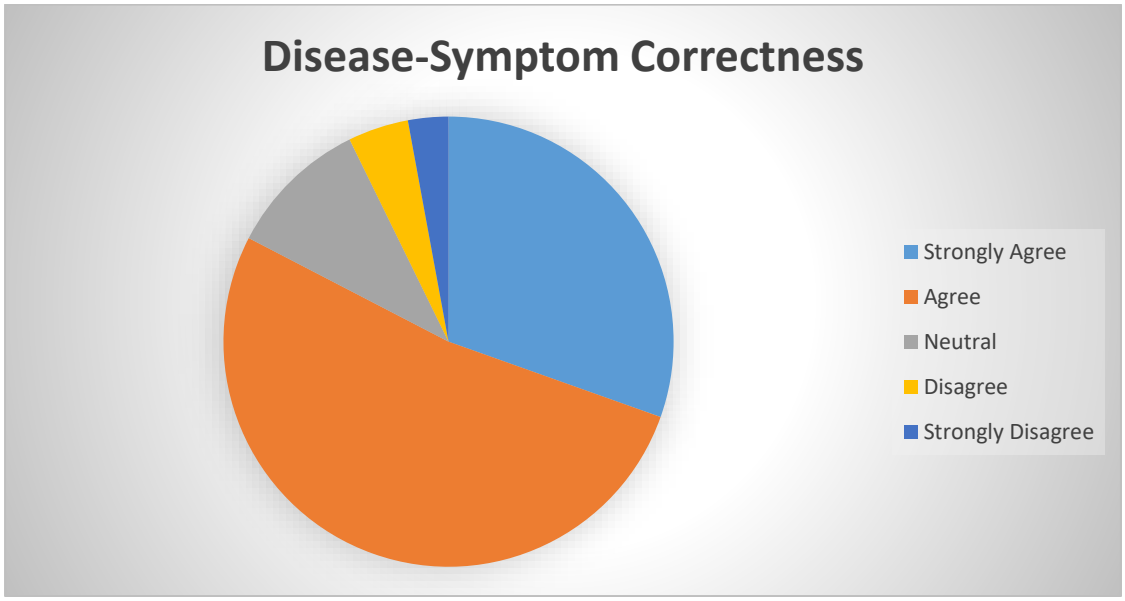
This question emphasizes on the coverage of liver disease domain defined in the proposed ontology. Among the domain experts 59% strongly agreed, 24% agree, 10% remained neutral while the remaining 7% suggested improvement to make the domain wider.

5.1.2 Symptoms Domain Coverage:



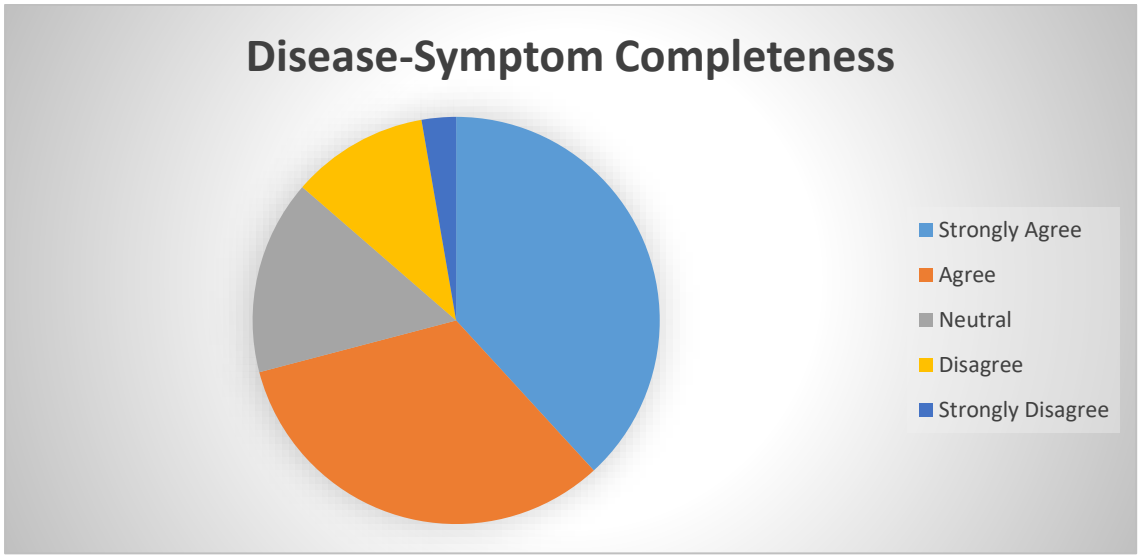
This question emphasizes on the coverage of symptoms domain defined for liver diseases in the proposed ontology. Among the domain experts 47% strongly agreed, 38% agree, 4% remained neutral while the remaining 11% suggested improvement to add symptoms not only associated with liver and general but all symptoms related to the human body.

5.1.3 Disease-Symptom Correctness:



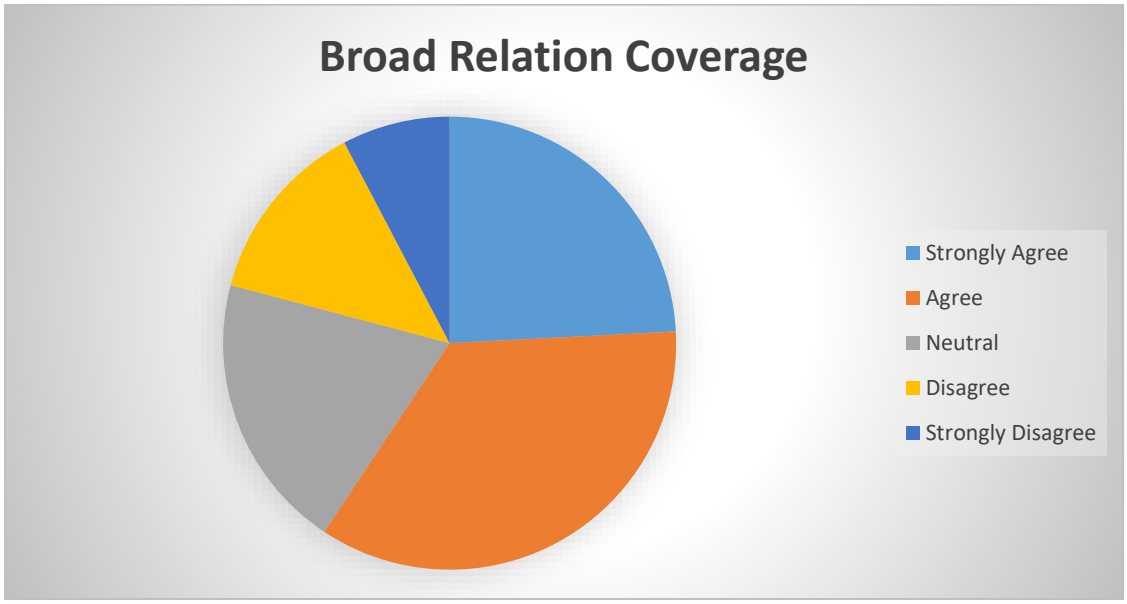
This question emphasizes on the correctness of symptoms domain defined for liver diseases in the proposed ontology. Among the domain experts 31% strongly agreed, 52% agree, 10% remained neutral while the remaining 7% suggested improvement to the existing symptoms by introducing sub classes of symptoms.

5.1.4 Disease-Symptom Completeness:



This question emphasizes on the completeness of symptoms with respect to liver disease domain in the proposed ontology. Among the domain experts 38% strongly agreed, 33% agree, 15% remained neutral while the remaining 13% suggested improvement to the existing symptoms. The suggested to add detailed symptoms and explore maximum symptoms for each disease be they generic or linked to other organs as well.

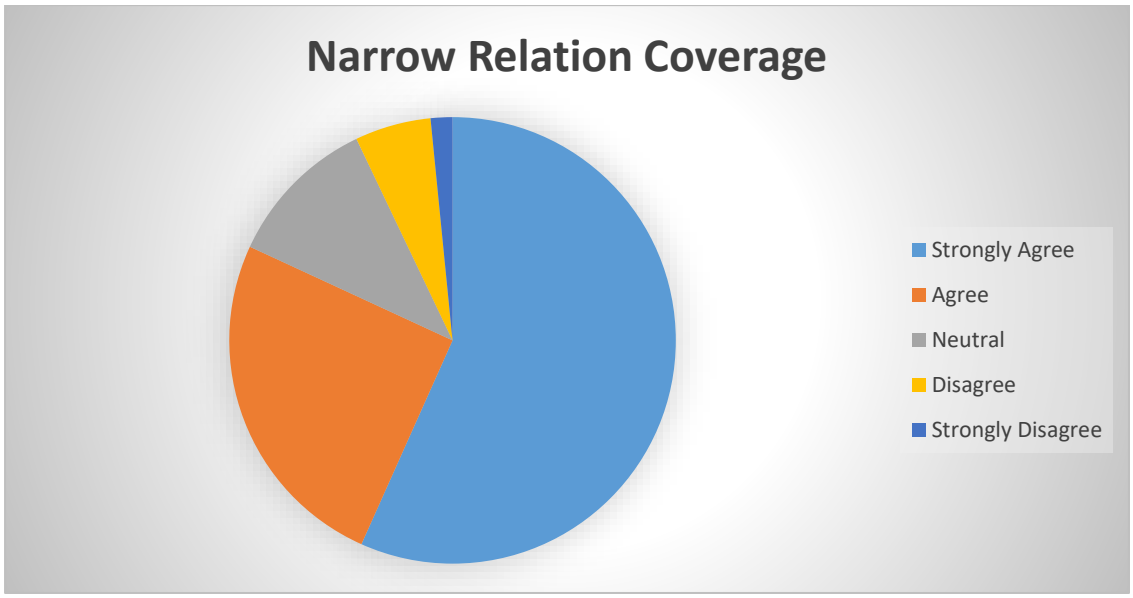
5.1.5 Broad Relation Coverage:



This question emphasizes on the coverage of broader relations defined in the proposed ontology. Among the domain experts 24% strongly agreed, 35% agree, 20% remained neutral

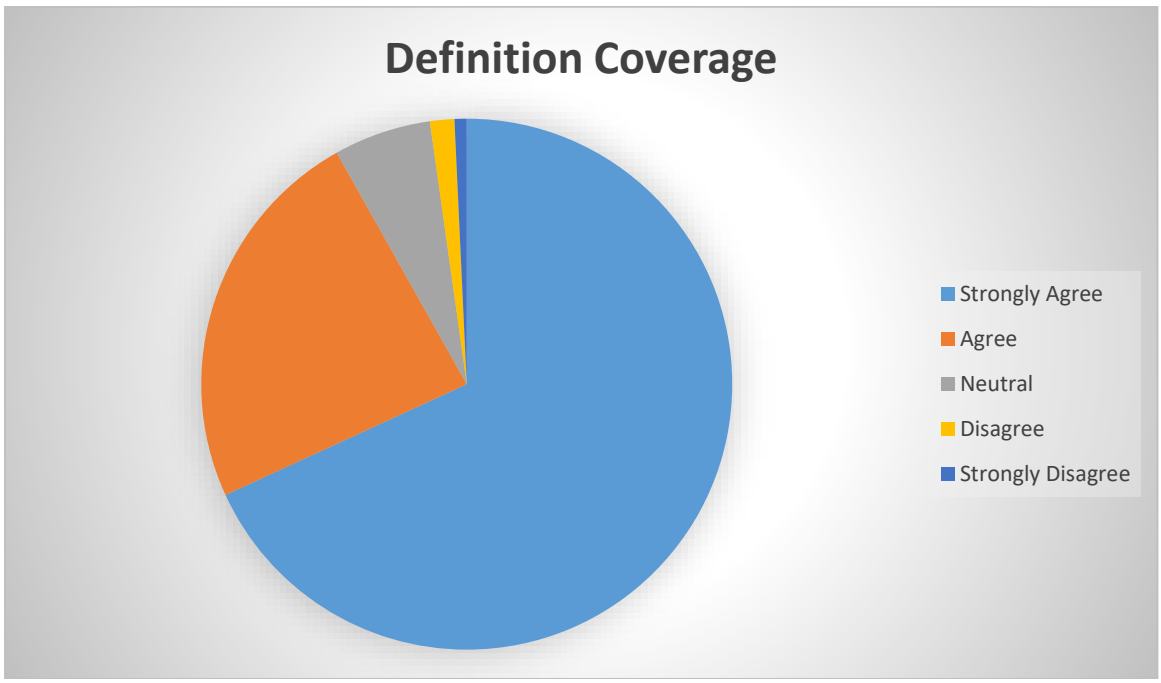
while the remaining 21% suggested improvement to the broader relations. The existing relations do no cover the relation of liver diseases to the entire human body.

5.1.6 Narrow Relation Coverage:



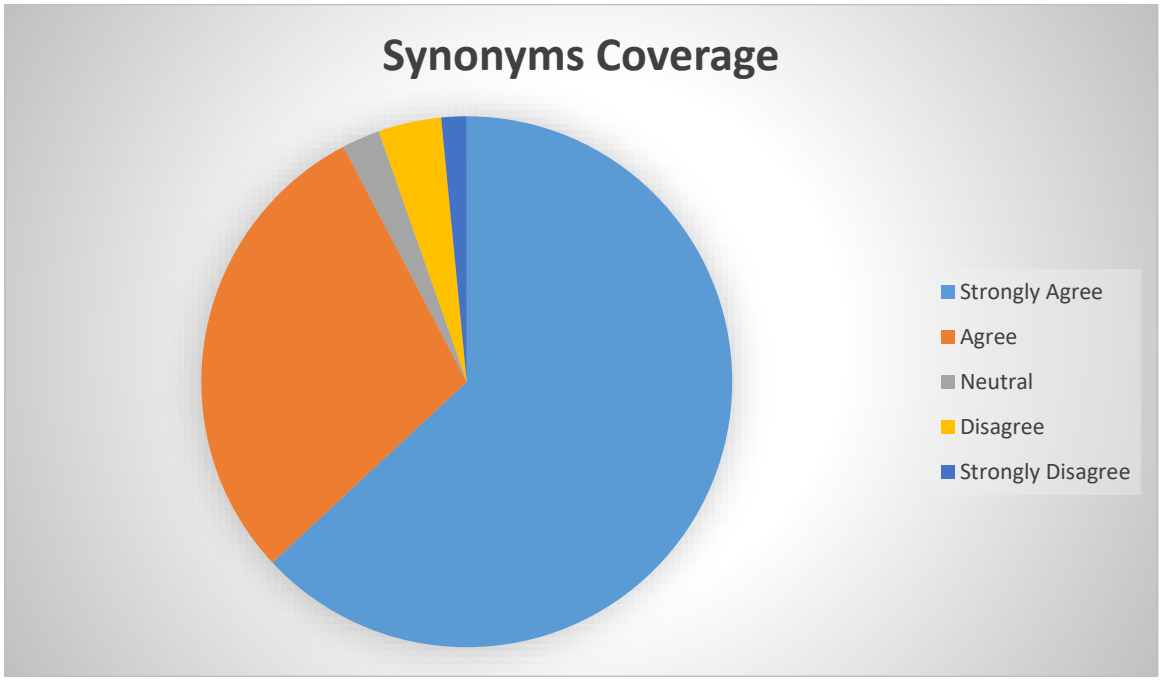
This question emphasizes on the coverage of narrow relations defined in the proposed ontology. Among the domain experts 57% strongly agreed, 25% agree, 11% remained neutral while the remaining 7% suggested improvement to the narrow relations.

5.1.7 Definition Coverage:



This question emphasizes on the coverage of definition defined in the proposed ontology. Among the domain experts 68% strongly agreed, 24% agree, 6% remained neutral while the remaining 2% suggested improvement through adding multiple languages other than English.

5.1.8 Synonyms Coverage:



This question emphasizes on the coverage of synonyms defined in the proposed ontology. Among the domain experts 63% strongly agreed, 29% agree, 2% remained neutral while the remaining 6% suggested improvement through adding more languages.

Chapter 6

Conclusion & Future Work

CHAPTER 6: CONCLUSION & FUTURE WORK

This chapter deals with an overview of research, conclusion covered in Section 5.1 whereas future work in section 5.2

6.1 Conclusion

In this research, I presented the complete procedure to develop biomedical ontology which contains complete coverage of domain knowledge with respect to Liver Disease. Inside the domain of biomedicine previously, there were abundantly present knowledge resources. After going through an overview of the contemporary art of existing knowledge collections of general as well as biomedical domain, we noticed that none of them was capable of providing certain level of details regarding different causative factors of diseases at one place through one core KB, without focusing on a specialized context. So, this gave birth to this idea of creating a core KB, that can provide integrated knowledge regarding, diseases, symptoms.

The core KB was achieved by extracting data from UMLS and sorting out through it on the basis of CUIs. These CUI formed a list of diseases related to liver and formed the basis of our ontology creating classes for the diseases stated above. Their definitions, synonyms and relationship type with other diseases are explored creating triplets from the data to form machine-readable context of the vast data. Liver Disease ontology unlike other ontologies not only focus on disease's definition, synonyms and relationship type but also focuses on disease-symptom relation.

A disease symptom relation is created using existing symptoms ontology namely, Symptom Ontology (SYMP). The ontology developed by us namely Liver Disease Ontology and Symptom Ontology by the Genome Sciences are merged and through experts opinion the link from diseases was created to its respective symptoms. Resulting into a new ontology having disease-symptom relation along with definition, synonyms and relationship types triplets in ".obda file", classes, subclasses, annotation properties, object properties and datatype properties in ".owl file".

6.2 Future Work:

In future, I would create a web application where the ontology is made available to clinical personnel. As well as graphical representation of the data which explores relations of a single disease to multiple other diseases. Also linking symptoms and diseases from on organ of the body to another. The ontology would be accessed by the web application through Java

API namely OWLAPI for this purpose. Furthermore, the ontology can be extended to multiple languages through the UMLS multilingual platform that gives access to biomedical data in multiple languages.

```
package com.kanwalontologyproject.liverontology.owlapi;

import java.io.File;
import org.semanticweb.owlapi.apibinding.OWLManager;
import org.semanticweb.owlapi.model.IRI;
import org.semanticweb.owlapi.model.OWLOntology;
import org.semanticweb.owlapi.model.OWLOntologyCreationException;
import org.semanticweb.owlapi.model.OWLOntologyManager;


/**
 *
 * @author Kanwal Wahab
 */
public class OWLAPIFirst {

    public static final IRI liver_ontology= IRI.create("C://Users/Kanwal Wahab/Desktop/ontology/testcld.owl");
    public static void main(String[] args) throws OWLOntologyCreationException
    {
        OWLOntologyManager ontologyManager = OWLManager.createOWLOntologyManager();
        File file = new File("C://Users/Kanwal Wahab/Desktop/ontology/testcld.owl");
        OWLOntology manLOntology = ontologyManager.loadOntologyFromOntologyDocument(file);
        System.out.print(manLOntology);

}
    }
```

Figure 25 "Reading OWL file through OWL API"

References

REFERENCES

1. Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, Hans-Peter Kriegel, “Extraction of semantic biomedical relations from text using conditional random fields” (2008)
2. Konopka, B. (2015). Biomedical ontologies—A review. Biocybernetics and Biomedical Engineering, 35(2), pp.75-86.
3. Nicola Guarino, “Formal Ontology and Information Systems”, Formal Ontology in Information Systems, Proceedings of FOIS’98, Trento, Italy, 6-8 June 1998
4. Feldman R, Regev Y, Hurvitz E, Finkelstein-Landau M: “Mining the biomedical literature using semantic analysis and natural language processing techniques” (2003)
5. Gruber, T. R, “Toward principles for the design of ontologies used for knowledge sharing”, International Journal Human- Computer Studies Vol. 43, Issues 5-6,Novemer 1995, Pages: 907-928
6. Ahn, H. S., & Kim, H. J. (2012). Development and implementation of clinical practice guidelines: current status in Korea. Journal of Korean medical science, 27(Suppl), S55-S60.
7. De Backere, F., Moens, H., Steurbaut, K., Colpaert, K., Decruyenaere, J., & De Turck, F. (2012). Towards automated generation and execution of clinical guidelines: Engine design and implementation through the ICU Modified Schofield use case. Computers in biology and medicine, vol. 42, pp. 793-805.
8. Heckerman, D., & Horvitz, E. J. (2013). The myth of modularity in rulebased systems [Online]. Available at: http://arxiv.org/abs/1304.3090, [Acessed on: 20 October 2013].
9. Berner, E. S. (2007). Clinical Decision Support Systems (pp. 3-22). New York: Springer Science+ Business Media, LLC.
10. Chen, R. C., & Bau, C. T. (2013). An Ontological Approach for Guideline-based Decision Support System. International Journal of Computer, Consumer and Control (IJ3C), vol. 2, no. 3.
11. James Mayfield, Tim Finin. Information retrieval on the Semantic Web: Integrating inference and retrieval. SIGIR 2003 Semantic Web Workshop, 1 August 2003
12. Natalya F. Noy, Deborah L. McGuinness,“Ontology Development 101: A Guide to Creating Your First Ontology”, Stanford University, 2005.
13. Bodenreider O, Stevens R: Bio-ontologies: Current Trends and Future Directions. Briefings in Bioinformatics 2006,7(3):256-274
14. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009; 37:W170–173.
15. Noy N, McGuinness D. Ontology development 101: a guide to creating your first ontology, Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880; 2001.
16. Yu AC. Methods in biomedical ontology. J Biomed Inform 2006; 39:252–66.
17. Tiffani J. Bright, E. Yoko Furuya, Gilad J. Kuperman, James J. Cimino. Suzanne Bakken “Development and evaluation of an ontology for guiding appropriate antibiotic prescribing Journal of Biomedical Informatics 45 (2012) 120–128
18. Weikum, Gerhard, and Martin Theobald. "From information to knowledge: harvesting entities and relationships from web sources." Proceedings of the twenty-ninth ACM SIGMOD-SIGACTSIGART symposium on Principles of database systems. ACM,(2010).

19. Alobaidi, M., Malik, K. and Hussain, M. (2018). Automated ontology generation framework powered by linked biomedical ontologies for disease-drug domain. Computer Methods and Programs in Biomedicine, 165, pp.117-128.
20. UMLS: Unified Medical Language System. http://www.nlm.nih.gov/ research/umls/
21. Sabir, B., Qamar, U. and Muzzafar, A. (2014). Ontology development and evaluation for urinal tract infection. 2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM).
22. Ciobanu, G., Horne, R. and Sassone, V. (2016). A descriptive type foundation for RDF Schema. Journal of Logical and Algebraic Methods in Programming, 85(5), pp.681-706.
23. Iqtidar, A., Muzaffar, A., Qamar, U. and Rehman, S. (2019). A biomedical ontology on genetic disease.
24. Protégé, "Protégé: A open source ontology editor and knowledge-base framework," http://protégé.stanford.edu
25. Shen, Y., Yuan, K., Chen, D., Colloc, J., Yang, M., Li, Y. and Lei, K. (2018). An ontology-driven clinical decision support system (IDDAP) for infectious disease diagnosis and antibiotic prescription. Artificial Intelligence in Medicine, 86, pp.20-32.
26. Hanna, J., Joseph, E., Brochhausen, M. and Hogan, W. (2013). Building a drug ontology based on RxNorm and other sources. Journal of Biomedical Semantics, 4(1), p.44.
27. RxNorm: https://www.nlm.nih.gov/research/umls/rxnorm (Last access on 19th January, 2019)
28. Kibbe, W., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C., Binder, J., Malone, J., Vasant, D., Parkinson, H. and Schriml, L. (2014). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Research, 43(D1), pp.D1071-D1078.
29. Horridge, M., Tudorache, T., Nuylas, C., Vendetti, J., Noy, N. and Musen, M. (2014). WebProtégé: a collaborative Web-based platform for editing biomedical ontologies. Bioinformatics, 30(16), pp.23842385.
30. Zenuni, X., Raufi, B., Ismaili, F. and Ajdari, J. (2015). State of the Art of Semantic Web for Healthcare. Procedia - Social and Behavioral Sciences, 195, pp.1990-1998.
31. Rehman Faisal Alharbi, Jawad Berri, Samir ElMasri. "Ontology based clinical decision support system for diabetes diagnostic", 2015 Science and Information Conference (SAI), 2015
32. Natalya F. Noy, Deborah L. McGuinness,"Ontology Development 101: A Guide to Creating Your First Ontology", Stanford University, 2005
33. Olivier Bodenreider1 and Anita Burgun2 "BIOMEDICAL ONTOLOGIES" MEDICAL INFORMATICS U.S. National Library of Medicine, Bethesda, Maryland 20894 ; USA; 2Laboratoire d'Informatique Médicale, Université de Rennes I, 35043 Rennes Cedex, France.
34. Simon Walk Masters Dissertation "Pragmatic analysis of collaborative ontology engineering processes" Knowledge Management Institute Graz, 19.April 2012.
35. Yu, A. C. (2006). Methods in biomedical ontology. Journal of Biomedical Informatics, 39, 252-266.
36. Chen, Y., Sabri, S., Rajabifard, A. and Agunbiade, M. (2018). An ontology-based spatial data harmonisation for urban analytics. Computers, Environment and Urban Systems, 72, pp.177-190.
37. L. Zhao and R. Ichise. Ontology Integration for Linked Data. Journal on Data Semantics, 3(4):237–254, May 2014.

38. L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease Ontology: a backbone for disease semantic integration. Nucleic acids research, 40(Database issue):D940–6, Jan. 2012
39. Corbridge C, Rugg G, Major NP, Shadbolt NR, Burton AM. Laddering: technique and tool use in knowledge acquisition. Knowledge Acquis 1994;6:315–41
40. Butte AJ, Chen R: Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. American Medical Informatics Association Annual Symposium, AMIA'06 2006, 106-110.
41. Yu AC. Methods in biomedical ontology. J Biomed Inform 2006;39:252–66.
42. Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering:Principles and methods. Data & Knowledge Engineering, 25(1-2):161–197, mar1998.
43. Gruber, T. R, "Toward principles for the design of ontologies used for knowledge sharing", International Journal Human- Computer Studies Vol. 43, Issues 5-6, Novemer 1995, Pages: 907-928.
44. James Mayfield, Tim Finin. Information retrieval on the Semantic Web: Integrating inference and retrieval. SIGIR 2003 Semantic Web Workshop
45. Guarino, N. (1997). Understanding, building and using ontologies. International Journal of Human and Computer Studies, 46, 293-310
46. Butte AJ, Chen R: Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. American Medical Informatics Association Annual Symposium, AMIA'06 2006, 106-110.
47. Pub med health (http://www.ncbi.nlm.nih.gov/pubmedhealth)
48. https://uts.nlm.nih.gov/home.html
49. https://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/help.html
50. https://www.springer.com/in/book/9783642387203
51. https://www.springer.com/in/book/9780387698991
52. https://www.springer.com/in/book/9783319332444
53. http://disease-ontology.org/
54. http://disease-ontology.org/resources/related-resources
55. https://github.com/DiseaseOntology/SymptomOntology
56. https://bioportal.bioontology.org/ontologies/FMA
57. https://omictools.com/diseasome-tool
58. https://www.ncbi.nlm.nih.gov/omim
59. https://www.ncbi.nlm.nih.gov/mesh
60. http://www.snomed.org/
61. http://www.genome.jp/