BUILDING A BIOMEDICAL ONTOLOGY ON COVID-19

Author

Sehar Shafique

00000319660

MS-19 (CSE)

Supervisor

Dr. Usman Qamar

DEPARTMENT OF COMPUTER AND SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

DECEMBER 2021

Building A Biomedical Ontology on COVID-19

Author

Sehar Shafique

00000319660

MS-19 (CSE)

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

MS Computer and Software Engineering

Thesis Supervisor

Dr. Usman Qamar

Thesis Supervisor's Signature: ____

DEPARTMENT OF COMPUTER AND SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

DECEMBER, 2021

Declaration

I certify that this research work titled “*Building a Biomedical Ontology on COVID-19”* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student
Sehar Shafique
00000319660
MS-19 (CSE)

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student
Sehar Shafique
00000319660
MS-19 (CSE)

Signature of Supervisor
Dr. Usman Qamar

Copyright Statement

Acknowledgements

In the name of Allah, the most beneficent, the merciful. All praises to Him, who bestowed me with His countless blessings, one of which include the ability to conduct and complete this research to the fullest of my capabilities. Without the knowledge, strength and perseverance provided by Him, I would not have been able to achieve anything.

The amount of effort and knowledge invested in me by my supervisor Dr. Usman Qamar, who is also designated as the Head of Department, is also limitless and invaluable. He guided me through every step of my thesis research and played a very crucial role in concluding this research with me.

Furthermore, I would like to extend my deepest gratitude to my entire thesis committee including Dr. Wasi Haider Butt and Mr. Jahan Zeb. Their invaluable support and guidance lead me towards the completeness of this research work.

Last but not the least, my family has been the epitome of support and courage throughout my journey. My Parents, who were my constant source of inspiration throughout my life, and my sisters, who have been my motivation, played a vital role in making me who I am today.

To My Father: The most courageous man and my biggest cheerleader

And

To My Mother: The Iron lady, and my support system

Abstract

The purpose of this thesis research work is to cater the requirements, techniques, and findings of implementing an ontology for the domain of COVID-19. This task is achieved by using a data source named UMLS in order to ease the access of biomedical data for the end users and clinical researchers. The literature available in the form of research articles is unorganized and unstructured therefore it is difficult for the seeker to find the right data. COVID-19 being a deadly viral disease has affected almost the entire population of the world in one way or another. The data related to COVID-19 pandemic is so vast and much research are being conducted on a daily basis in order to find the causes and cure of this viral infection. Keeping in mind this information, the accessibility of the data while keeping its semantic meanings intact is a tricky job. Therefore, a knowledge base must be designed in order to depict information and data related to the desired domain in such a manner that its semantic meaning is intact and the data is available in machine readable format. Hence, the major purpose of this research is to create an ontology on the primary concepts and their relationships as well as the knowledge related to the domain of COVID-19 or Corona Virus. This research is conducted by conducting a thorough literature survey on the existing ontologies and their implementation methodologies. The interfaces for the facility of end users and researchers are also reviewed. The ontology for the domain of COVID-19 is developed using UMLS as a data source. Furthermore, the data extraction techniques along with the ontology development are also catered.

The major concepts, attributes and relationships among them are mapped on a live portal (https://ontologycovid.000webhostapp.com/) consisting of other domain related knowledge for the facilitation of end users and clinical researchers.

Keywords: ***COVID-19 ontology, knowledgebase on COVID-19, UMLS COVID-19 Knowledgebase, COVID-19 portal, Corona Virus***

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

Chapter 1 Introduction

1. Introduction

The use of internet since its advent has increased drastically as almost majority of the world’s population has access to it. The users of internet frequently visit the webpages for their day to day queries and research. The data available online belongs to almost every aspect of our lives. Ranging from healthcare to education, science to arts, sports to music, internet has data relevant to all the fields of our daily lives. Amongst all these domains, healthcare and medical sciences are of great value along with other sub domains. Health care has garnered a lot of interests of internet users be it researchers, scholars or even a regular person, everyone is interested in finding out the secrets of medicine and health.

Despite all these facts, it is also true that majority of the data available online is jumbled and is not interpretable by everyone. In many circumstances, the data available is not refined and it must be passed through various normalization procedures in order for it to make sense. Many a times the data is available in the form of scientific articles or journals which is also unstructured and requires an expert for that particular domain to be understandable. In other words, the data available online is useless unless it is structured and is make available in a standard format understandable by everyone.

The rise in biomedical research is the result of the number of discoveries and the data gathered in this domain. In result, the research experts and scholars are studying this domain and publishing valuable findings at a very fast pace. Therefore, new tools and techniques are being introduced by the researchers rapidly. But the published data and techniques are unstructured and is not compatible with the search engines in terms of size and structure hence resulting in pulling unrelated and irrelevant search results. As the number of data and publications rise, researchers, after extensive research, decided to introduce semantic web in order to focus on the semantics of the data by storing the relational data into machine readable format.

1.1Background and Overview

The literature and the published data in biomedical field has increased rapidly since the previous score and the main reason behind this escalation is the advent of innovative tools and techniques [1]. This is apparent that the huge number of written sources in biomedical field can be adeptly processed with the help of automatic text-extraction processes[2]. The information available online in terms of healthcare is gargantuan and huge but due to its unstructured format, it is highly challenging to extract the relevant results from web which is why many researchers and other internet users are unable to benefit from this data [3].

The data related to biomedical research available in research articles and journal publications is unstructured and those of web or search engines, it is available in semi-structured or unstructured formats [4]. Google, which is the most common search engine and is accessed by millions of users on a daily basis, is not able to resolve this obstacle. Whereas the general directories like Yahoo and gratis text-based search engines are also facing the same issue [5]. The major reason behind this inability of pulling the relevant results through search engines is the lack of semantic structure of the web pages. The vast knowledge of biomedical research is very useful if structured properly and can be of help if it is made available for the internet users. The knowledge can be helpful in thwarting different diseases by diagnosing at an early stage and treating early and correctly. These methods can even help prevent the epidemics and pandemics and resolving the issues related to them.

The biomedical research is increasing at a very fast pace with every passing day and the researchers and scholars working in the fields of clinical medicine, biology and biomedical research are overwhelmed by the number of discoveries prompted by the various factors [6]. The data and information regarding this domain are huge and is increasing rapidly. The existing and mostly used popular search engines are unable to fetch the relevant results due to the unstructured format and different size of available data. In order to resolve this issue, the researchers, after extensive research and analysis, figured out the solution to be the use of semantic web [3], [4], [7], [8].

When it comes to sematic web, there are a lot of misconceptions about its exact meaning and definitions. A net of information woven in a way that is understandable and processable by all the machines trouble-free [9],[10]. We can also define sematic web as an approach that defines a language which is machine readable [11]. Keeping in mind these facts, it is easily deduced that adding semantics to a web page is not an easy task and requires extreme concentration and authentic data hence, leading us towards ontology creation.

The word ontology was first used by a renowned philosopher Aristotle in the field of Metaphysics[12]. He portrayed ontology as the analysis of aspects that are related to one another on the basis of their indistinguishable traits. Ontology, as a term is defined by various researchers and scholars all over the world in many different ways, but the most authentic and worldwide accepted definition of ontology is proposed by a famous scholar Studer et al. 1998. According to Studer, the definition of Ontology is “unambiguous formal specifications of some shared conceptualization where conceptualization is used as a simplified representation or view of the real world that is to be represented by ontology”[13]. In other words, the symbolization of various concepts of a specified domain, their sub-concepts and the relationships between them is known as Ontology [14], [15] The ontology act as the building block of semantic web as the app y exchange the information available related to a specified domain among the applications and groups [7].

The information available nowadays on web and other available sources is very vast and to dig the data relevant to the specified domain is not an easy task. Gathering the well-structured and authentic knowledge base is the most important task in accessing knowledge and the processing tasks associated with it. For this purpose, RDF or Resource Description Framework is used which is a model solely established on the visual representation of data or graphs. The data is displayed in the form of first-order logic also called object-predicate-subject format. Such a type of data representation is termed as RDF triple. For instance, if we talk about the statement “COVID-19 has symptom of fever”, then in the format of RDF triple, it can be depicted as COVID19-symptom-fever. In this statement, “COVID-19” refers to the subject, “symptom” refers to the predicate and “fever” refers to the object of a statement or sentence. In the same manner, if the data is represented in the form of graphs or triples, then it can be easily understandable by the end users and researchers. In order to make an ontology related to the biomedical domain effective, it is important that it caters all the concepts, relations, attributes etc related to a particular disease and covers all the definitions, synonyms, semantic types of that disease effectively. Such a type of ontology can only be created if the data source used is authentic and reliable. The data sources such as UMLS, MeSH, NLM etc are authentic and reliable data sources as they are updated on a quarterly basis.

1.2Domain Selection

The selected domain for this research is COVID-19, a novel and fast growing viral infection. COVID19 is an ongoing pandemic of a respiratory disease named Corona Virus. It is caused due to the severe acute respiratory syndrome also called SARS-CoV-2. This outbreak was first reported in Wuhan, China in December 2019 where local people were identified with a respiratory infection. This infection was later on identified to be a severe viral disease which was spreading rapidly. COVID-19 Virus spread throughout the world within a matter of months and infected many people. The number of deaths caused by this viral disease also increased within months. According to WHO, up until now around two hundred million COVID-19 cases are reported along with around 4 million deaths all over the world [16].

This work intends to create an ontology mainly based on UMLS as a data source for the domain of COVID-19. The UMLS data comprises of definitions, concepts, synonyms, relationships, and their semantic types. Furthermore, the developed ontology is evaluated by domain experts. As we are aware of the fact that COVID-19 is spreading rapidly since its first case was reported in 2019 resulting in a global pandemic infecting hundreds of millions with million being dead. The world is facing the challenge of corona virus outbreak and many people lost their lives due to this infection being novel and no research being done in this field. The number of literature being published related to COVID-19 on a daily basis is very vast and a huge amount of data is being gathered in the result of these research works. To acquire the specific data from such a huge collection of information is practically an impossible task hence, giving purpose to this research to create an ontology.

1.3Outline

The thesis outline depicts the flow of the research in this document. The literature of already done work is discussed in the chapter 2. All the previous research in the field of ontology creation and development are discussed in chapter 2. Chapter 3 caters all the techniques and methodologies used in order to extract the data from UMLS resources. The data extraction and data transformation processes are discussed in that chapter. The storage of collected data and creation of knowledge base is catered in chapter 4. The creation of MySQL database along with storing the data in that database is also covered in forth chapter. The chapter 5 discussed the ontology creation and development along with the tools used to create a live portal for the facilitation of end users. Chapter 6 shows all the results obtained from the COVID-19 ontology as well as the results of Ontology validation by the domain experts. In the chapter 7, all the findings of this research are discussed. Furthermore, the activities for future along with the research applications are also discussed in the seventh chapter.

CHAPTER 2

LITERATURE REVIEW

2. **Literature review**

This chapter discusses the various ontologies designed and implemented to date including both the general and the biomedical ontologies. The overview of already developed biomedical repositories along with their domain specific features are presented. The literature review and the related work done in the field of biomedical ontologies with the help of various datasets are also discussed in this chapter. Moreover, the shortcomings and limitations in the previous studies are also described.

2.1Overview

Proving the existence of anything is defined as Ontology which itself has various semantic meanings in various contexts. The origin of this word Ontology is from the field of knowledge known as Philosophy. In the nineteenth century, the foremost ontologists in the history namely Bolzano, Husserl and Frege presented principles for distinguishing various objects along with their correlation among each other. By the twentieth century, the use of ontology was distributed towards the field of Artificial Intelligence which gave way to the usage of ontology in terms of sharing data and presenting the concepts in a descriptive manner [1]. Keeping that in mind, an ontology can be defined as "the representation of concepts of a particular domain in a hierarchical manner in correspondence to their capabilities"[1]. Ontology can be defined in another way as "an explanatory model of a particular domain which articulates and illustrates the concepts related to that particular domain, the attributes and the relationship among them" [1], [2].An Ontology consists of concepts and the attributes along with the relationships between them. The concept in a particular domain is of great significance as it helps in understanding and sharing the knowledge and data.

The idea of ontology was first given by Aristotle in the field of research in Metaphysics [17]. He proposed that the depiction of elements that are related to each other on the basis of their identity is defined as an ontology. A researcher named Studer defined ontology as "The comprehensive details or a specific description of a concept and that particular concept is represented in a streamlined manner or a real time vision that needs to be characterized by the ontology"[13]. Ontology can also be defined in another explanation as the illustration of a particular area of research which discusses and displays the data related to that area of research, the elements, along with the correlations between them [14], [15]. In semantic web, the ontologies act as the building blocks as it enables the exchange of data among the classes and systems[7].

RDF or Resource Description Framework consists of a framework based on data which includes visual representations, for instance graphs etc, as well as generates the statements related to the concepts in the given domain. Those statements are demonstrated in the form of a structure which follows the “subject-predicate-object” form. This particular format is known to be the RDF triple. To explain it with the help of an example, the sentence “COVID-19 has the symptom of fever” can be portrayed as “COVID-19-symptom-fever” in terms of an RDF triplet. The above mentioned statement portrays “COVID-19” as a subject, “has Symptom” as predicate and “fever” as an object. Keeping these things in mind, we can deduce that RDF structure is more suitable and easier to understand as compared to that of relational data in terms of portraying information. RDF represents the data in the form of statements which are further illustrated in the form of graphs. If the information or data is contained and portrayed in the format which is easy to understand and is semantically correct, only then is it considered valuable, which is how a biomedical ontology should be. A biomedical ontology must consist of all the attributes along with the subdomains concerning a particular area of research including the domain disease knowledge, symptoms, factors causing the disease etc. Furthermore, an ontology created for these purposes should be implemented precisely mindful of the data sources e.g MeSH, UMLS, SNOMED etc. An ontology also comprises of concepts related to the selected domain, features as well as the associations among them. Moreover, an ontology must be protractible, platform independent as well as reusable. This application must be accessible to everyone so that efficient and valuable research can be performed in the biomedical domain. As we know that the COVID-19 pandemic has taken the whole world by storm and must be cured through vaccinations in order to prevent further loss of lives. These research must be performed by the clinical researchers to help produce the effective vaccinations for COVID-19. The ontologies are logical models designed to specify the concepts of an area of interest. In order to gain maximum benefit from the ontologies, they must be developed in a formal representation.

2.2Ontology

An ontology is described as an information model which is accurate and provides the complete and comprehensive knowledge of a particular domain along with its concepts, sub-concepts and the relationship among them [18], [19], [20].

As the ontologies provide the collective and distinguished knowledge of a domain therefore, it is not understood by the humans and is only comprehended by machines. Subsequently, this information which is comprehended by machines only could be further classified into various modified systems along with the structuring of many other applications comprising data interpretations [20], [18], [10], implementing expert systems[18], extracting information along with NLP also known as Natural Language Processing [10], [21], [20]and is also used as a necessary concept in Semantic Web. Furthermore, ontologies can also be transformed into Clinical Decision Support System or CDSS for enhancing knowledge and data reusability [22].

2.3Applications of Ontology

The ontologies designed for a particular domain act as the main ingredient in designing and implementing the intelligent computer systems [23]. They deliver a practical solution to the problems or challenges being faced and are considered as the major elements for the applications. The design and implementation of an ontology can be used as the data for building another ontology. In other terms, in order to implement an ontology, a previously implemented ontology can be re-used. The resultant of an ontology which is linked to another ontology lies under the "Open World Assumption" which means that anyone can "say anything about anything", which is considered to be the primary notion of Sematic Web. As far as databases are concerned, they follow the "closed world assumption", according to which the databases are only limited to the concepts of a particular concept or in other words, the data is limited to a particular source. Ontologies, as the flexible blueprint can be employed throughout the various systems in terms of sharing and using data, knowledge, and information in compliance with W3C standards [24], [25].

Ontology is best described to be the classified representation of the concepts along with their relations between each other. These concepts are designed and implemented in order to provide the extensive information associated with a specific area of interest. The main function of an ontology is to share, represent, search and swap data or information regarding a specified domain between systems [26]. The applications of ontologies are the data generalization domain along with classification, terminology composition [27], reuse and evaluation. Moreover, the creation of vocabularies in terms of data representation, development of domain knowledge to retrieve data for clinical decision support system [28], [22], [29].

2.3.1 Ontology Development

Mainly, there is no specified or standardized technique of creating an ontology[30]. As ontology itself is domain-specific, therefore, every ontology is unique and represents its own domain very vastly. The two most common methods are iterative development trajectory and collaborative technique [30], [31],[32]. A few guidelines or general steps are available that can be followed in order to develop an ontology for a specified domain. The four most commonly used guideline principles are mentioned as under [33].

i. The knowledge of a specified domain and create the specifications for requirements.
ii. Create the concepts of that particular information in a set of Intermediate Representations (IR).
iii. Develop the model of conceptualized concepts in the form of a formal language.
iv. Validate the implemented ontology throughout all steps and stages involved in the life cycle by a defined platform or domain experts

The events involving these corresponding steps are mentioned as under[34]:

i. Select and describe the area of research clearly and unambiguously.
ii. Gather all the required information for that particular domain. Adjust and figure out all the concepts and the relationships among those concepts within the domain.
iii. Understand the information or data in detail along with the manuals available in the printed format.
iv. Normalize the information gathered and available in that particular domain in order to ensure the following:
    - Logical reasoning
    - Compatibility and reasoning with the relevant ontologies
    - Understandable by humans
v. Formulate the gathered information in the machine understandable vocabulary.

In the development of an ontology, the Intermediate Representation or IR model is defined in the first four steps mentioned above. It includes obtaining the data or information related to the domain in the form of tables or records or other data management techniques. This method, using IR for mapping concepts in a particular domain, works as the bridge between the necessary attributes of the field and the ensuing ontologies. It can be utilized in order to gather, scrutinize, and convey the data in the knowledge base.

As ontology is a concept of philosophy hence the limitations or constraints are also based on philosophical proverbs for example “realism”, “adequatilism”, “fallibilism” and “perspectivism”[35]. To make it easy to understand we can say that the manner in which they are group together is crucial as it affects their services and ability to perform. It also contains the insight of the selected domain, purpose of implementing an ontology, as well as the manuals for creating the ontology[36].

2.3.2 Components of Ontology

The ontology works on a methodology named as first-order logic process where it is broken down into different features. The features of an ontology include an object, its relationships with other objects, along with its purpose or function [20]. In order to have a better understanding of a first order logic, the most common daily life example can be of help which states that “All birds have feathers. Robin is a bird. Therefore, Robin has feathers.” In this way, a triple is formed which gives a better understanding of relations and objects in a human understandable form. In terms of classes, it can also be explained through the following example:

“classStudent isPartof classDepartment”

“classDepartment isPartof classUniversity”

Hence,

“classStudent isPartof classUniversity”

The identification and depiction of a relationship among objects can be explained in a much easier manner if ontologies follow the above mentioned procedure. In terms of computer processing technique, the Acyclic systems generate a diagrammatic composition which makes the interpretation and understanding simpler as compared to the tree like formation of cyclic expressions[37].

2.3.3 Classes

The classes define and depict the concepts and object in the particular domain[38]. The explanation of a classes can be the object that exists in the real world. These classes are a part of an ontology and are identified by a unique identifier. The relations of these classes with other classes and attributes are also present in an ontology. However, the class of an ontology is different from the conventional database class. The most distinct and important difference is the accessibility of an object from several classes. For instance, if we consider an ontology related to "University", an object can belong to the class "Person", class "Student", class "Teacher" as well as the class "Thing" at the same time. The definition of a class can be the collection of data within a particular concept in such a manner that the concepts are related to every concept in the class furthermore, the description and other details regarding that concept are also part of that data [23].

2.4Ontology Classification

The use of ontology is becoming very common among the scholars as discussed in the previous chapter and these ontologies are of various distinct kinds and types. Therefore, keeping in mind the diversity of ontologies with respect to the disciple of their studies, these ontologies are classified into different categories which are discussed as under:

2.4.1 Upper Ontology

Upper ontologies are the ontologies which provides the basic information of a specific discipline with lesser details regarding a particular area of research [29],[18]. This type of ontologies is helpful in order to understand the highest level information within the domains[39]. For the domain of biomedical sciences, an ontology named BFO, or Basic Formal Ontology is implemented as the upper level ontology which depicts the high level or basic information. This ontology is also recommended and acknowledged throughout the biomedical domain as the specification for interoperability between ontologies[40]. Also, for the domain of biomedicine, the Disease

Ontology is considered as the upper ontology which caters all the basic information related to the biomedical domain.

2.4.2 General Ontology

The general ontologies are said to be those ontologies where the interpretation of data is available at an average level. Furthermore, these type of ontologies are independent of the task. Time/space theory etc are examples of general ontologies.

2.4.3 Domain Ontology

These ontologies are associated with a particular domain and provide the complete and comprehensive information or data related to that particular domain. For instance, the Gene ontology represents knowledge related to one particular domain. In the same way many other biomedical diseases are represented in the form of an ontology.

2.4.4 Application Ontology

The ontologies developed for the specified task falls under the category of application ontologies. These ontologies are only developed in order to carry out a specific task or project.

2.4.5 Reference Ontology

The ontologies which are not dependent on any outcome and are designed specifically to assist the common components among the various domains are called as reference ontologies.

2.5Reasons to Develop Ontology

The purpose of creating ontologies is distributing the data with the researchers as well as scholars that work on related field of information. The principal intentions for developing ontologies include [18],[41].

- To share or distribute the data or knowledge related to the same domain
- To reuse the ontologies implemented priorly
- Differentiating the domain and functional data

2.6 Biomedical Ontologies

Earlier on the way to the beginning of computer systems, the immense information of natural science and biomedicine was saved in vague or ambiguous form. However, it is astounding that researchers who study the domain of biomedicine worked and initiated discovering methods to determine complications and also began to characterize data in organized manner. There is a section of efforts in this domain and experts arrange the biomedical data as well as described the basic concepts. The structure and standardization of terms in biomedicine have reduced in the categorization of infections, repository management, terminologies as well as vocabularies or ontologies.

The use of ontologies in the area of biomedicine is developing with the passage of time [42],[23]. For instance, a language of the biomedical ontologies, includes more than 300 ontologies along with detailed sources in BioPortal2[42].

2.6.1 GALEN

It is an acronym for Generalized Architecture for Languages, Encyclopedia and Nomenclature in Medicine. GALEN is a European venture built for reusing the repositories in context with the clinical applications[43]. It is progressed from KB terms developed by Alan Rector’s Pen and Pad EMRS (Electronic Medical Record System). Many of the customary repositories are synchronized inside the knowledge assembly nonetheless GALEN delivers the repository for basic chunks of recounting the terminologies. For instance, the terms Adenocyte as well as Thyroid glands are available inside GALEN. Nevertheless, rather than provision of an unambiguous demonstration for Adenocyte thyroid gland GALEN designates description of these concepts through amalgamation of terms as indicated, Adenocyte that is a fundamental element of Thyroid gland. An open informant foundation known as Open GALEN was developed in year 2000 for distribution of orientation prototypical and operation with software hawkers and repository builders for supporting its allowance and usage. Subsequently Galen prototypical is expended for studying nursing repositories, CDSS Clicinal Decision Support Systems, surgical practises as well as human anatomy

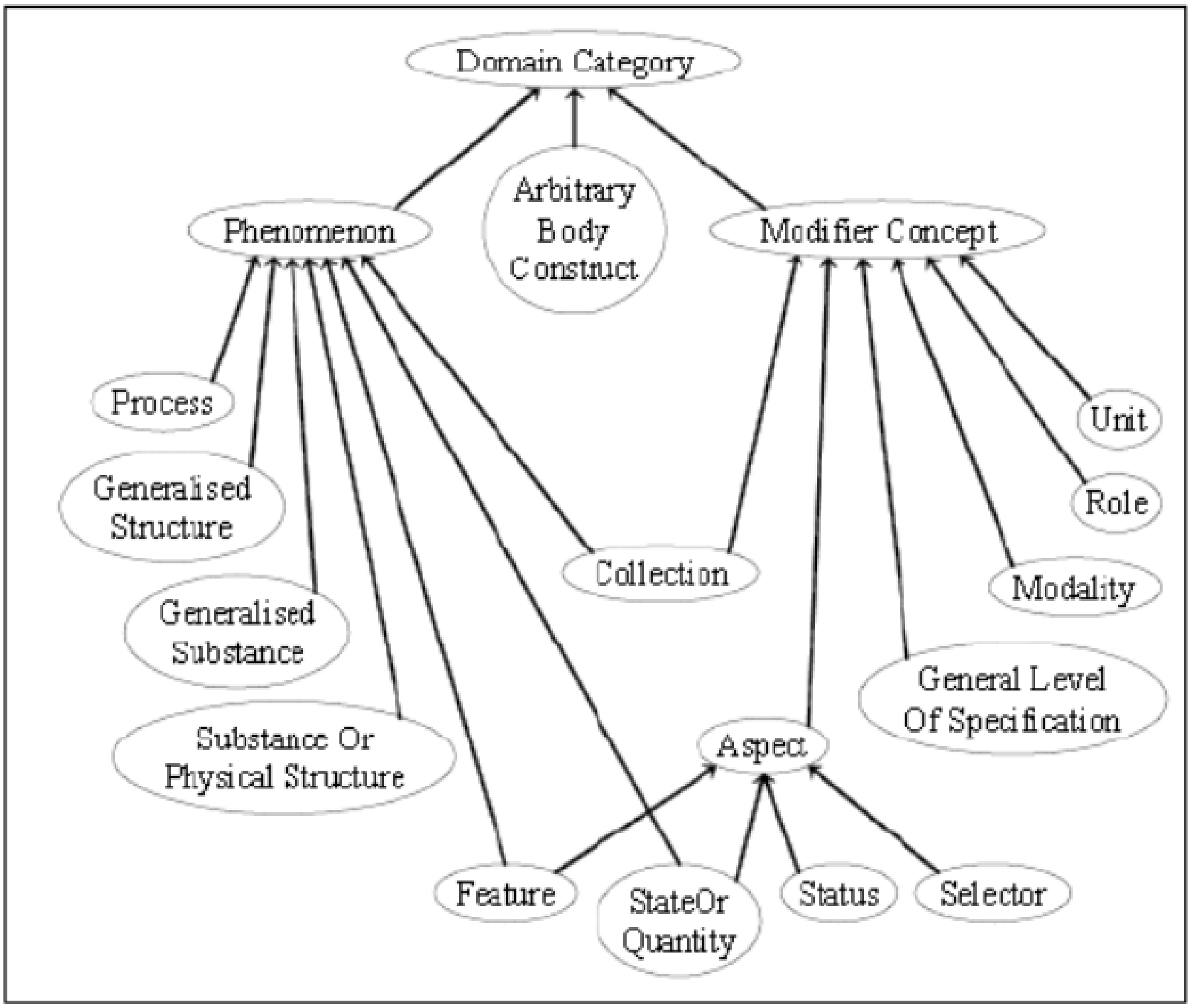


Figure 2.1: GALEN Ontology [15]

2.6.2 SNOMED

It is an abbreviation of Systematized Nomenclature of Medicine or SNOMED CT abbreviated as 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 number of bio-medical terms and concepts. The latest version of SNOMED contains 269,864 terms as well as 407,510 names[23].

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.
To access any term inside SNOMED-CT every term which is present in SNOMED CT can be elucidated with the help of different rudiments. For instance, class “viral meningitis” contains a individual identifier <58170007>. SNOMED-CT contains 18 autonomous pyramids by the help of which categorization of the term organization can be seen in earlier varieties of SNOMED-CT

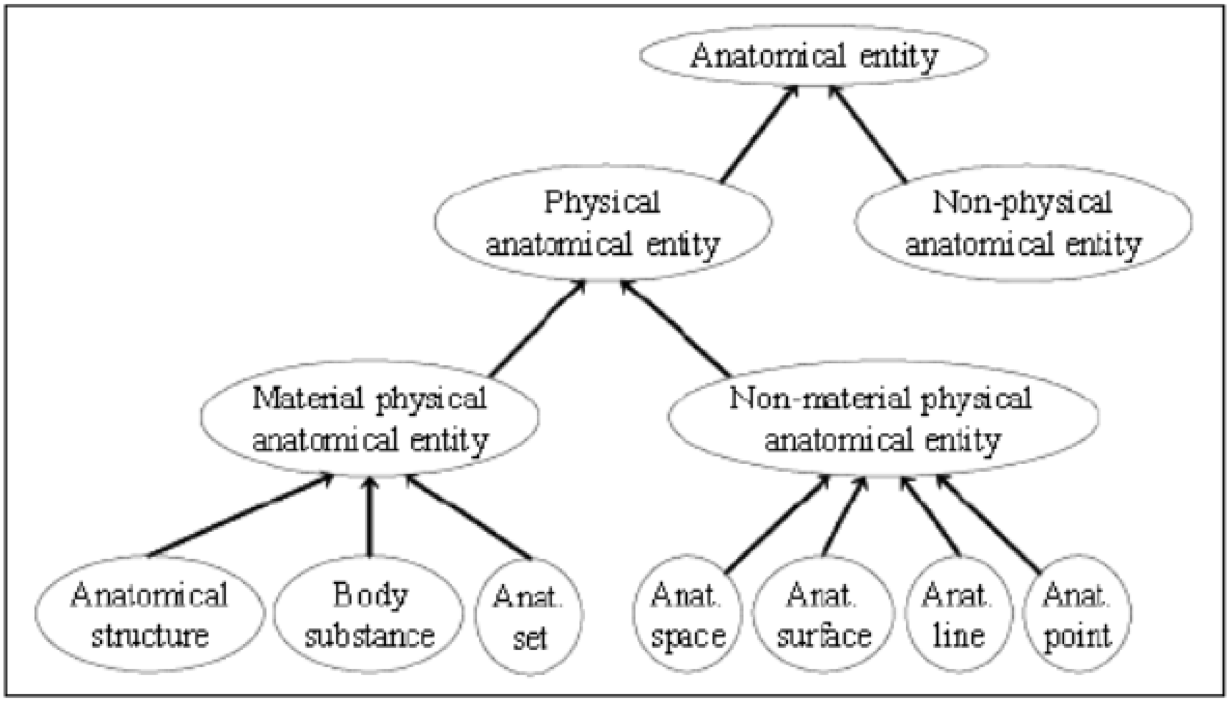

Anatomical entity
Physical anatomical entity
Non-physical anatomical entity
Material physical anatomical entity
Non-material physical anatomical entity
Anatomical structure
Body substance
Anat. set
Anat. space
Anat. surface
Anat. line
Anat. point

Figure 2.2: SNOMED-CT Ontology [44]

2.6.3 Open Cyc

It is a conventional ontology developed by the Cycorp Inc. The elaboration of this ontology was started on this endeavor and it is developed about a staple of approx. greater than one million hand- implied announcements. It is articulated CycL formal language which captures “ordinary intellect” KB as well as it permits diversity of knowledge-Based systems. This ontology contains about of six thousand terms as well as sixty thousand proclamations on these terms [38].

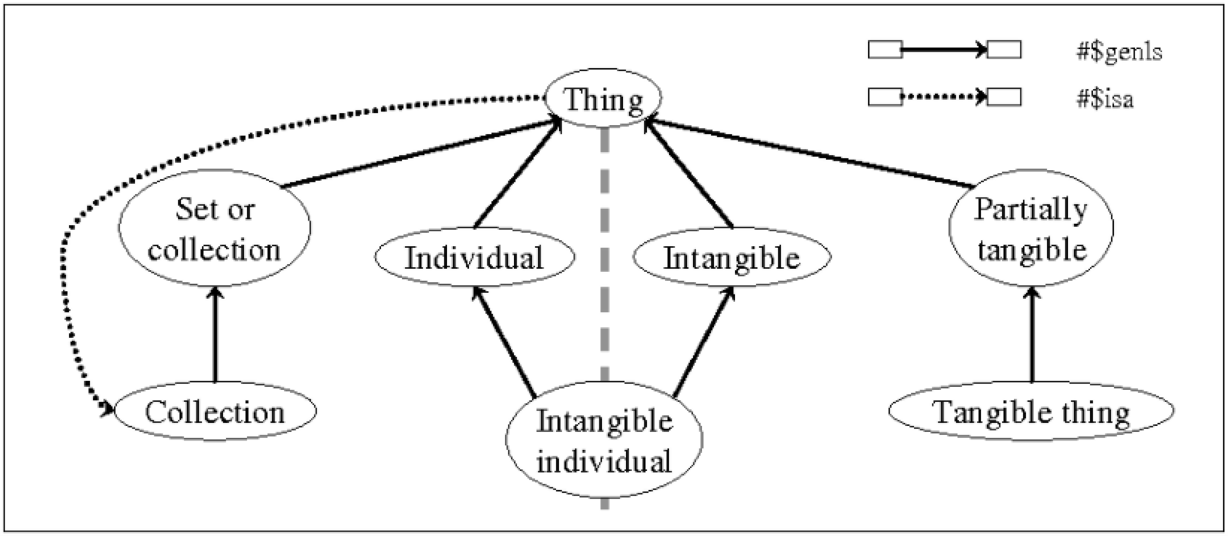


Figure 2.3: OpenCyc Ontology

2.6.4 WordNet

It is an automated etymological repository. Princeton University which roles out as a source for systems developed in NLP (Natural Language Processing) as well as information retrieval developed WordNet.

For one base term inward organization in WordNet contains a clique of synset also known as synonyms. Synset conformation is based on synonyms which introverted connotation is articulated by several concepts as well as polysemy that designates that 1 concept entails many differentiating connotations.

The version 2.0 of WordNet contains greater than 114,000 synonyms which are nouns and are alienated into 9 pyramids. Every hierarchy begins with a distinct novice. Each synonym inside the hierarchy of nouns turns into 1 is-a tree structure as well as possibly or goes to few part-of tree hierarchies[45].

Inside WordNet notches of terms which resembles to disorders of health in the medical expressions are categorized suitably for instance Leukemia is a hyponym of Cancer[46],[47]. Alternately inside few of the artworks a medical emblem or indication originates into prospect simply as a hyponym of a concept which is non-medical. This viewpoint acmes corporeal appliance to a convinced amount than pathology, as a consequence there exists no recognized connotation amongst Vasoconstriction as well as the bio-medical domain in the WordNet.

2.6.5 FMA

FMA is an acronym for Foundation Model of Anatomy. The elaboration of Foundational Model of Anatomy (FMA) was at University of Washington and it rosette out of previous exertion for editing to supplement the structural substance of the UMLS. Emphasizing absolutely on the interpretation of structural organization, FMA gaze onward and appears as citation ontology, for example for permitting diverse ontological systems of what an anatomy can be constituted to be concomitant along[38].
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[14]. Protégé provides a framework for editing and building ontology The classifications of the corporeal organizational concepts included in FMA are articulated by specifying manacles [48] developed on the longitudinal factors; amass, as well as intrinsic 3-dimension silhouette, concluded as well as over the organizational components that mount the physique. Relations in contrast to this are restricted towards the organizational anatomy of corporeal functional concepts.

2.6.6 MENELAS

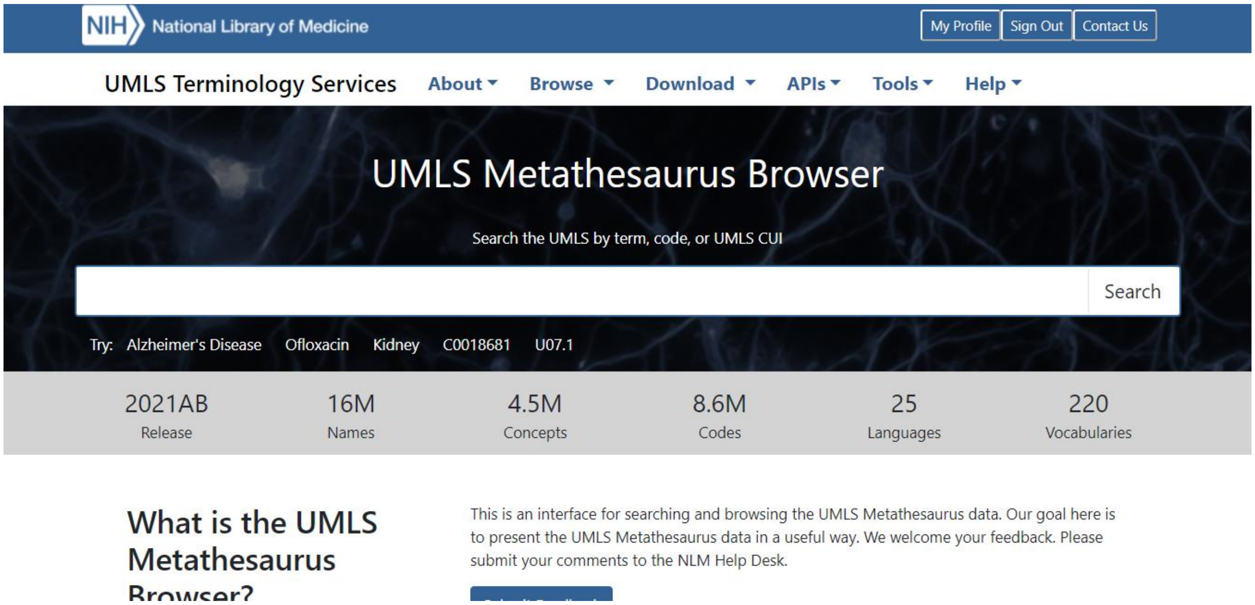
It is a European Union research endeavor developed for the purpose of retrieving bio medical chronicles inside diverse European vernaculars[35]. MENELAS clutches a KB methodology for understanding the natural language. An undeviating system all-encompassing coronary artery syndrome is established, as well as possessions that are characterized as theoretical grids, view justification of domain of specific semantic and syntactic dictionaries in addition to this it serves as ontology of coronary artery syndromes enhanced with designed comprehensive acquaintance for each term.
Currently the developed MENELAS ontology encompasses about eighteen hundred terms as well as three hundred association categories to become from an amount of foundations that

comprises consultations with domain experts, recycle of the reachable repository of possessions, as well as a quantity investigation. It was initially industrialized as a framework [49], [30] nonetheless, to escape indistinctness owed to assorted heirloom, the criterions of adversary of relations as well as exceptional semantic alliance were far along espoused accompanying a sapling assembly[30].

Besides this, some research work has also been carried out regarding relations amongst the diseases as well as their causative influences. We conducted this research in accordance with diverse frameworks, inclusive of relationships amongst the genotype as well as drug rejoinder phenotype [50], the genetic associations with a disease, as well as the etiological factors and disease[18].

2.6.7 UMLS

It is an abbreviation of Unified Medical Language System. It was established by NLM for facilitating experts in the health care domain as well as the scholars for accessing the biomedicine knowledge after different foundations[51]. The repository entails a massive collection of the terms, as well as the Semantic System, an enclosed relationship of about one hundred and thirty five semantic categories, assimilated above 100000 terms encompassing about one hundred terminologies as well as repositories. The construction of each foundation is endangered in constructing the repository; correspondent concepts are congregated inside a semantically special term. Interrelate connotations are whichever consented from fundamental repositories or deliberately fashioned.
UMLS recognizes any entrance into 3 collections: filament – demonstrating a concept as repository. The Etymological group – filaments of identical construction could be plotted. Term - sequence of alike connotation. The information is characterized by definite occurrences of relationships amongst definite concepts



UMLS Metathesaurus Browser

Search the UMLS by term, code, or UMLS CUI

2021AB	16M	4.5M	8.6M	25	220
Release	Names	Concepts	Codes	Languages	Vocabularies

What is the UMLS Metathesaurus Browser?

This is an interface for searching and browsing the UMLS Metathesaurus data. Our goal here is to present the UMLS Metathesaurus data in a useful way. We welcome your feedback. Please submit your comments to the NLM Help Desk.

Figure 2.4: UMLS Meta-thesaurus Browser

There is no compulsion imposed by the Meta-thesaurus on the resources i.e. it does not bequeath among the type of association projected from the ontology. In contrast to this the Semantic Network provides identity- overriding of terminologies incorporated inside the Meta-thesaurus as well as it hands out as a rudimentary, upper level ontology restrained for the biomedicine domain[52].

The classification of opinions by semantic category is dependent on the bargain rule corresponding towards the interpretation of stinginess established [53],[54]. The consequences of frugality wellspring for demonstrating information inside UMLS are contended elsewhere[37].

2.7COVID-19/Corona Virus

2.7.1 COVID-19 Overview

The viral disease named COVID-19 after the ongoing pandemic also known as Corona Virus is an infection caused by SARS-CoV-2 also known as severe acute respiratory syndrome coronavirus 2. It is a disease which spreads from person to person and falls under the viral infection category. The first ever reported case was found in the late December of 2019 in a province of China named

Wuhan[55]. The virus has spread throughout the world causing a major world pandemic affecting more than half the population of the world[56].

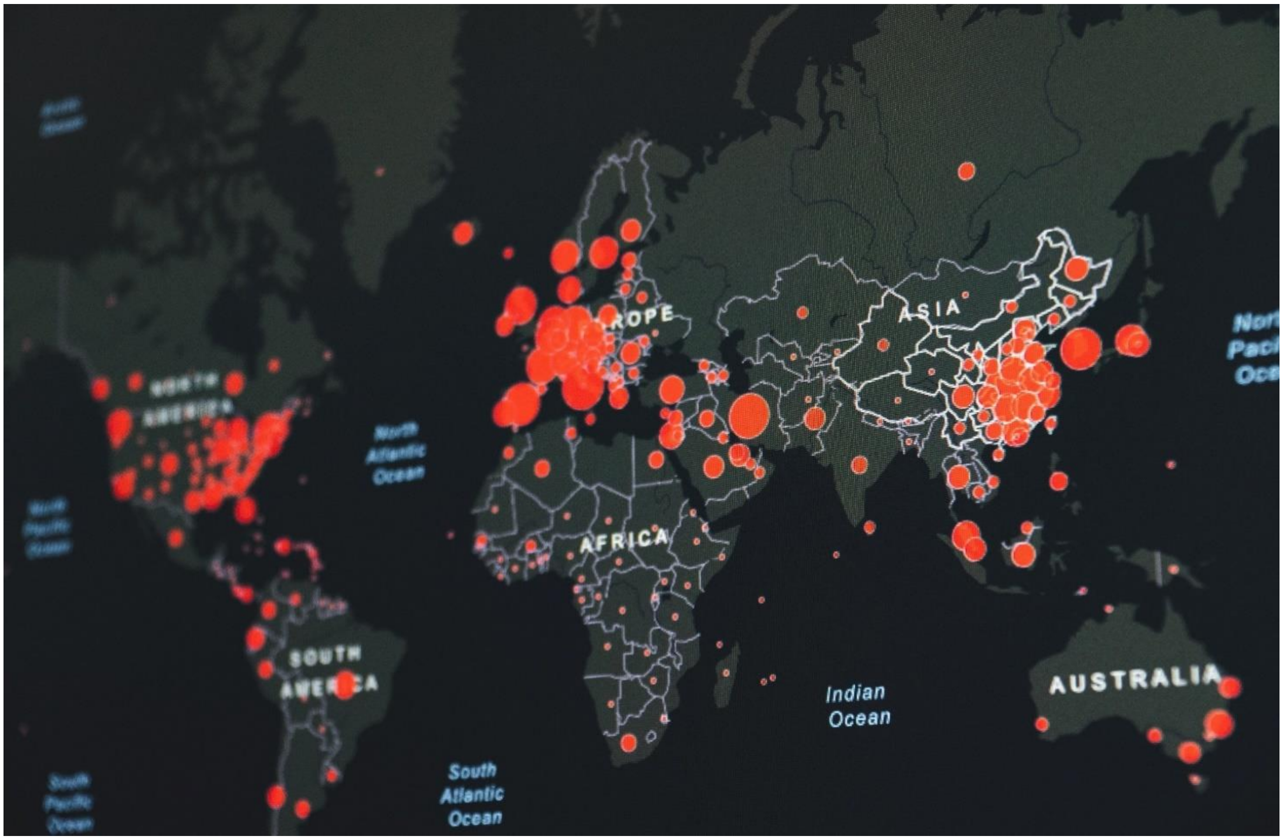


Figure 2.5: COVID-19 Stats according to WHO

2.7.2 Symptoms of COVID-19

This viral infection has reported the symptoms of headache[57], fever[58], fatigue, cough, difficulty in breathing, loss of smell and taste[59], [60], [61]. The symptoms of this viral infection are initiated after two to fourteen days of virus exposure and in one third cases, the symptoms are not even detected [62]. Out of the patients contaminated by the virus and develop symptoms, around 80% show the slight to normal symptoms which includes pneumonia, whereas around 15% suffer from acute symptoms which include hypoxia, lung issues etc. Moreover, around 5% patients show critical symptoms which include shock, organ dysfunction, respiratory dysfunction [63]. The COVID-19 poses a bigger threat to the people of older ages as they have other diseases like blood pressure, diabetes etc. Furthermore, in many cases after the patient recovers form COVID-19, they experience the long-term symptoms of COVID-19 which may result in organ damage [64]. In this regard, research is still in progress, and the long-term effects of COVID-19 are being explored further[64].

2.7.3 COVID-19 Transmission

The transmission of COVID-19 occurs through the breathing of infected air in the form of particles present in the air consisting of virus. It usually spreads from people to people and in most cases happens in the closed spaces. The infected person can transmit the virus to those present in the close vicinity by breathing, touching etc. The virus can transmit through the eyes, mouth or nose and in some cases, through infected proximities. The most crucial aspect about this virus transmission is that many a times, people carry this infection and do not develop the symptoms and spread the virus unknowingly as it can be carried in the human body for up to twenty days [64], [65].

2.7.4 Diagnostic Techniques

The virus can be diagnosed with the help of tests created specifically to detect the viral infections. The most commonly used method for COVID-19 testing is the Nucleic Acid Testing by RT-PCR also called as reverse transcription polymerase chain reaction, using a nasal swab. In many cases, a technique called TMA or transcription-meditated amplification can be used in order to detect the viral infection. Also, RT-LAMP technique which is also called reverse transcription loop-mediated isothermal amplification, using a nasopharyngeal dab.

2.7.5 COVID-19 Preventive Measures

The research regarding the cure of this viral infection was started immediately after the spread of COVID-19 all over the globe. The researchers were working to find the cure of this infection. Up till now, many vaccines are developed by many researchers throughout the world and the World Health Organization has approved a few vaccines after thorough and rigorous testing and results. Those approved vaccines are supplied in multiple regions of the world in order to cure the COVID-

19 viral infection. The governments of those countries have initiated the mass vaccination programs in order to make sure that their population is vaccinated and is able to fight the pandemic. Many other techniques are also used in order to help prevent the spread of this virus which include imposing lockdowns, social distancing, restricting the people gatherings to the outdoor spaces only, washing hands properly and frequently, quarantining those who show the symptoms of COVID-19 and many others. The most commonly used preventive method is the use of face masks in the public, as they reduce the risk of COVID-19 to around 80%.

2.7.6 Vaccines

The vaccines approved by WHO are the ones which are tested properly and thoroughly and are effective in curing the corona virus. These vaccines are developed by the researchers of different countries. The most effective vaccines include Pfizer BioNTech, AstraZeneca is closely followed by Sinopharm, Jansen, Sinovac and Moderna. These vaccines have been distributed among various regions of the world and are being produced in a bulk in order to cater the whole population of the world.

2.7.7 Variants

The virus is mutating with its spread in various region of the world and therefore, there are numerous variants of this virus which are grouped into the classes [66], [67]. These variants are denoted with the Greek alphabetical letters as Alpha, Beta, Delta, Gamma etc. In this way, the virus variants can be distinguished easily [68].

The different forms and variants of viruses have started emerging in the year 2020. The variant named cluster 5 was first found in Denmark among the mink farmers [69], [70]. By the mid of 2020, the most spreading variant of corona virus among the humans of different regions of the world belonged to the categories named alpha, beta, delta, and gamma. The alpha variant, also described as B.1.1.7 is originated from the London and Kent region and is named as the UK variant. The beta variant was first found in the regions of South Africa as is called as B.1.35.1 officially. Another variant originated from Brazil and is referred as P.1 falls under the category of gamma variants. Furthermore, the deadliest and the most spread variant of corona virus was discovered originally in India and is referred as B.1.617.2 is called the Delta variant [71].

2.8Related Work

[1] created an ontology for the diabetic patients using rule based and decision support system. The editor used to implement the ontology is Protégé and JENA structure is used to translate all the rules created in order to develop an ontology.

[15] created an ontology for the identification of liver diseases caused using drugs. They mapped the identified drugs form their samples using UMLS Meta-map and the resulting analysis is mapped to SNOMED CT.

Another author created an ontology in the domain of breast cancer. The data source used is UMLS and all the concepts, relations found in the breast cancers and its types are covered in this ontology. Both broad and narrow relations among the various types of breast cancer are catered. The resulting ontology is further portrayed in the form of a lie portal for the ease of clinical researchers and end users.

[72] created an ontology for obesity treatment. The author designed an intelligent e-therapy for the treatment of obesity in order to improve the efficiency of treatment for obese patients.

[73], developed an ontology for urinal tract infection. The ontology consists of symptoms, definitions, and properties among the related concepts and is developed using protégé, an open source editor for ontology. The data is collected form UMLS, and the developed ontology is further validated by domain experts. The resulting ontology is available online and is available for users to gather information regarding the specified domain.

[74] developed an ontology for chronic liver disease. This is a biomedical ontology with UMLS as a data source. It is designed in Protégé and is further evaluated by domain experts through a questionnaire.

[75], developed an ontology for genetic diseases. The ontology caters major concepts and their properties. The data is collected from UMLS, and the ontology is developed using Protégé with a live portal available. The ontology is then validated by domain experts through a questionnaire.

[76], the author developed an ontology for correct drug information. The data is collected through RxNorm which is a standard vocabulary developed by National Library of Medicine and is further mapped to the entities required for creating this ontology. The developed ontology is available on a web portal for easy access to the users and is validated through use cases where students and researchers identify the drugs from the information available by ontology.

[77]worked on the development of an ontology for human diseases. Their work included various human diseases using different biomedical ontologies related to them. The language used is web ontology language (OWL) through protégé as a toolkit. Protégé is an open source editor available for developing ontologies and other intelligent systems [78]. And OWL is a language used for the development of ontologies [79].

CHAPTER 3

DATA GATHERING

3. Data Gathering

3.1Problem Statement

Data collection and gathering is considered to be the most important and foremost part of any research. Similarly, in this research the most necessary and leading is step is to gather data for the efficient development of ontology. This chapter discusses the data source of our ontology as well as the methods and techniques used to extract the data from said sources. Moreover, the reliability of data is considered as the most frequent issue which is also discussed in this chapter.

As we are well aware that in order to achieve good and accurate results, the data used for analysis should be reliable and clean. If the data gathered is not clean or in other words it is not pre-processed, then the generated results will be inaccurate and false. In order to create an ontology, it is important that the data source should be well-known, and its trustworthiness or reliability is confirmed. If the data is not reliable and accurate, the resulting ontology will not produce the efficient and accurate results. Therefore, the data which is gathered from reliable sources and is accurate will work as a backbone of the ontology implemented. The purpose of using this data is to create an ontology and display it in a such a way that the semantic meaning of concepts belonging to the domain are clear and unambiguous and it is able to initiate the process of learning which defines the concept so that it is interpretable and could be used practically. Keeping this in mind, we considered numerous publicly accessible biomedical data repositories. After rigorous research, we finally used a data source called UMLS managed by NLM. UMLS also known as Unified Medical Language System is a corpus of biomedical vocabularies managed by NLM or National Library of Medicine. This data repository is selected as the data source for our ontology because it is open source, free for use for everyone, documented in detail, updated and maintained on a regular basis hence making it more reliable and trustworthy. Moreover, this data repository contains a vast set of concept attributes for a huge number of diseases. After the thorough analysis of various vocabularies, UMLS is selected as it is the most authentic and suitable repository and is its comparison is shown in the table below.

Ontology Name	Number of Concepts	Scope
SNOMED CT	352,567	Clinical medicine
FMA	73000	Human Anatomical structure
Gene Ontology	44,945	Functional Annotation of Gene products
MeSH	27,000	Biomedicine (Descriptors for indexing literature)
RxNorm	3.5 million apporx	Standard names for prescription drugs
LOINC	88,192	Clinical observation and laboratory tests
UMLS	6 million approx.	Terminology integration in life sciences

Table 3.1: Comparison of Biomedical repositories

When it comes to data, the data is dirty and should be pre-processed and handled correctly. As this issue is a bit complicated so we divided it into two parts. In the first part, data retrieval is tackled and in the second stage the extracted data is stored into the relational database or RDBMS. This part is managed by making sure that the semantic constraints including the relationships among the related concepts are preserved. In this chapter, we will only discuss about the first part that is data retrieval form the data source named UMLS. The management and storage of this data is discussed in the next chapter. The data retrieval stage is crucial and is discussed in detail in this chapter.

The most necessary and crucial requirement of development of this ontology is the gathering of precise and correct data. This chapter is going to focus on the techniques and methodologies applied in order to collect the data for ontology development.

The methods and techniques for collection of data from the data sources along with the comprehensive methodology for implementing and designing an ontology for the domain of COVID-19 is discussed further in this chapter.

3.2Research Methodology

As we have mentioned earlier that the data source for this ontology is UMLS. The UMLS itself consists of three main crucial parts which are discussed s under:

- The first part is called the Meta-thesaurus. It consists of information and data related to the biomedical concepts along with the terms and vocabularies used from various standards. The meta-thesaurus also contains the updated terms and vocabularies along with the classified applications as they are the building blocks of the biomedical knowledge base.
- The second part of UMLS consists of a semantic network. It caters the steadfast classification of every concepts available in the UMLS meta-thesaurus.
- The third part comprises of the Lexicon Specialist. This section offers the original and derivative information regarding the concepts in the UMLS meta-thesaurus.

The whole process involving the ontology learning is shown in the figure below. For the learning strategy of our ontology, we have collected summarized versions of literature available related to the biomedical research from the sources like Medline and Pub-Med repositories [34]. This technique was carried out through the search query which generated results on the basis of our search string which is for instance “COVID-19”. This search string produced results in the form of important concepts related to covid-19. After this stage, the next step involved mapping of these concepts to the UMLS terms and vocabularies related to biomedicine. Furthermore, the association rule system [3] were also used in order to find the related terms and vocabularies from the term pairs list.

In the next step, the collected term pairs from the previous stage and the selected relations are mined. The UMLS semantic network section consists of data related to the fundamental semantic classes which can be mapped with terms with the help of meta-thesaurus. Furthermore, the details regarding the relations available between the various semantic concepts are also available in the semantic network of UMLS. Around one hundred and twenty seven semantic types along with fifty four semantic relations are available in the UMLS semantic network 2021AA version. The upper level semantic concepts relationships are accessible and are mapped using is-a relationship which depicts the child-parent relation in a hierarchy.

3.3Data Collection

The data collection is the most crucial and important stage in any field of research. In order to perform the analysis of data and generating the accurate results, the quality of data and data source plays a very vital role. The use of high quality, accurate and clean data increases the accuracy of results and the methodology used. If the data is collected through a reliable source and is pre-processed thoroughly, the likelihood of an error occurring decreases immensely. Whereas in case of unreliable data, many consequences can be faced such as:

- Failure to produce the query results accurately
- Invalid and irrelevant query results which leads to inaccurate research findings.

As we know that the process of gathering the data, converting it and then storing the data into RDBMS is long and crucial therefore, we have discussed it in two chapters descriptively. The focus of the first chapter is the methods and techniques used to gather the data from the data source which is UMLS and in the next chapter we have discussed the process of data transformation and loading the extracted data into the RDBMS using query language mysql.

We have mentioned it earlier that UMLS also known as Unified Medical Language System is the biggest and the most distributed data source in the biomedical domain. It contains a large amount of data and information related to the biomedical domain hence making it the backbone of many application being developed in the field of medical sciences. UMLS itself is considered most authentic and reliable source of data because it is updated on a regular basis, every quarter to be precise, contains a vast number of sub-domains and their hierarchies and their terminologies and vocabularies. There are around hundred and seventy biomedical terminologies and vocabularies which are updated and maintained on a regular basis. Moreover, there are around three million labels and titles for 900,551 concepts and classes existing in meta-thesaurus browser. In addition to that, there are around twelve million relations both broad and narrow between the various concepts. The concepts and classes covered in the UMLS are tremendous and their number is increasing with the increase in literature and are represented in the figure below [80].

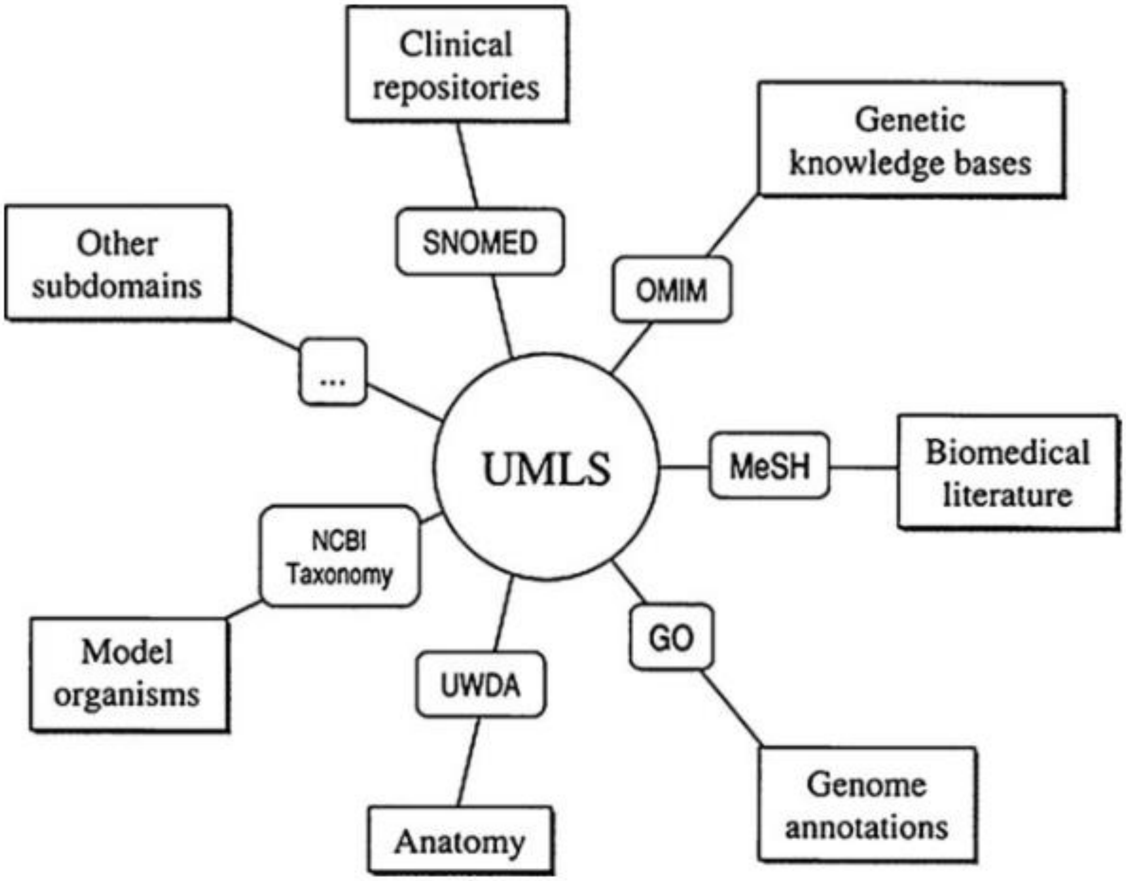


Figure 3.1: UMLS domains and sub-domains

In order to create an ontology using information from UMLS, a concept plays a very vital role. To develop a concept, all the identical concepts are gathered and arranged in a particular format. After that, these concepts are connected to the other concepts through relationships, hence forming a graph. In the previous paragraph it is mentioned that there are around twelve million relations in the UMLS meta-thesaurus. These relations are both inter and intra relations among the concepts.

3.4Data Gathering from UMLS

This step involves the methods and techniques used to fetch the data from data repository. The data is collected from the resource called UMLS. The UMLS data is accessed by signing up and creating an account for the purpose of getting license. After creating an account, the request is sent for approval of accessing data. The approval is only available for the research purposes and this UMLS data cannot be used for personal use. The licensing process is completed, and no fees is charged for this service. The data is then available to be downloaded on the machine or computer, with a size of around 58 GB. This data, after downloading, is then installed on the system to make it accessible for creating and developing our ontology. The data extraction process is carried out through a java based interface named Metamorphosys which is shown in the figure below. Metamorphosys is a UMLS tool used to extract the concepts and atoms related to any area of research. The metamorphosys is then configured in order to be used and to extract the data.

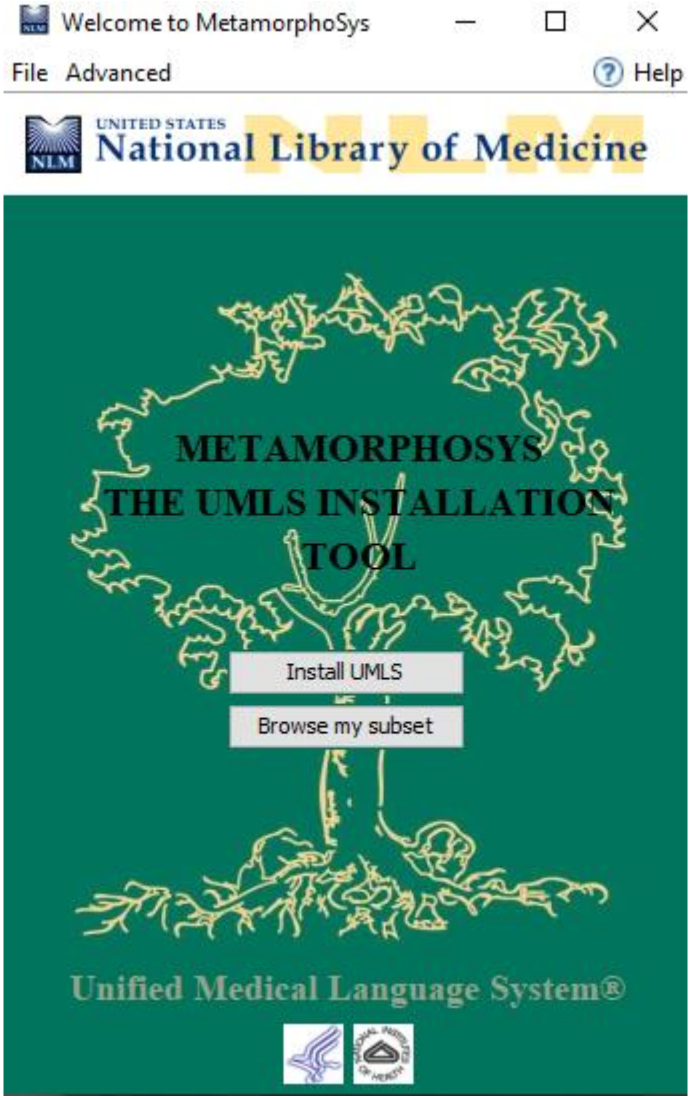
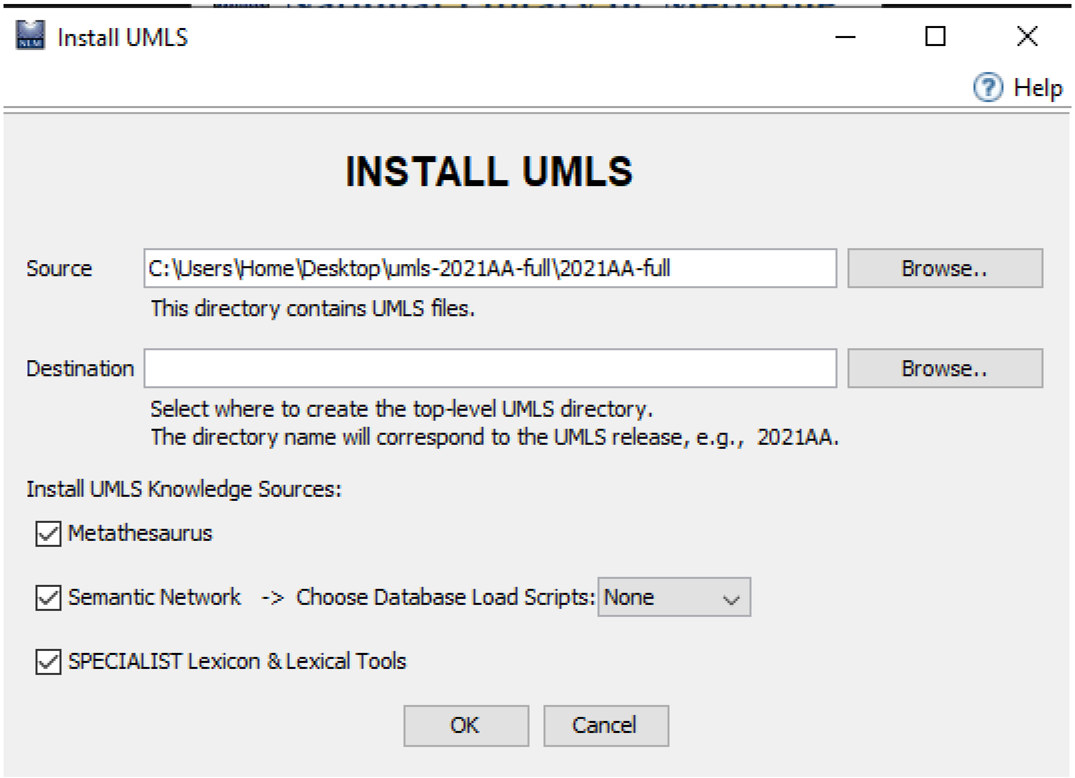


Figure 3.2: UMLS Metamorphosys Application

3.4.1 Extraction Technique

When we are performing data extraction from the data repository, we have to keep in mind the technique or method we are going to use in order to gather all the data. We have the liberty to choose whether we want to collect data in a single step or in the form of a number of steps. This type of data extraction technique is used in those cases where the data in repository is changing and updating constantly. In this scenario the data changes frequently and the database also needs to be updated on a constant basis in order to achieve maximum and precise results. As the data in UMLS changes quarterly therefore we have collected all the data in a single go from UMLS.



INSTALL UMLS

Source: C:\Users\Home\Desktop\umls-2021AA-full\2021AA-full

This directory contains UMLS files.

Destination:

Select where to create the top-level UMLS directory.
The directory name will correspond to the UMLS release, e.g., 2021AA.

Install UMLS Knowledge Sources:

☑ Metathesaurus

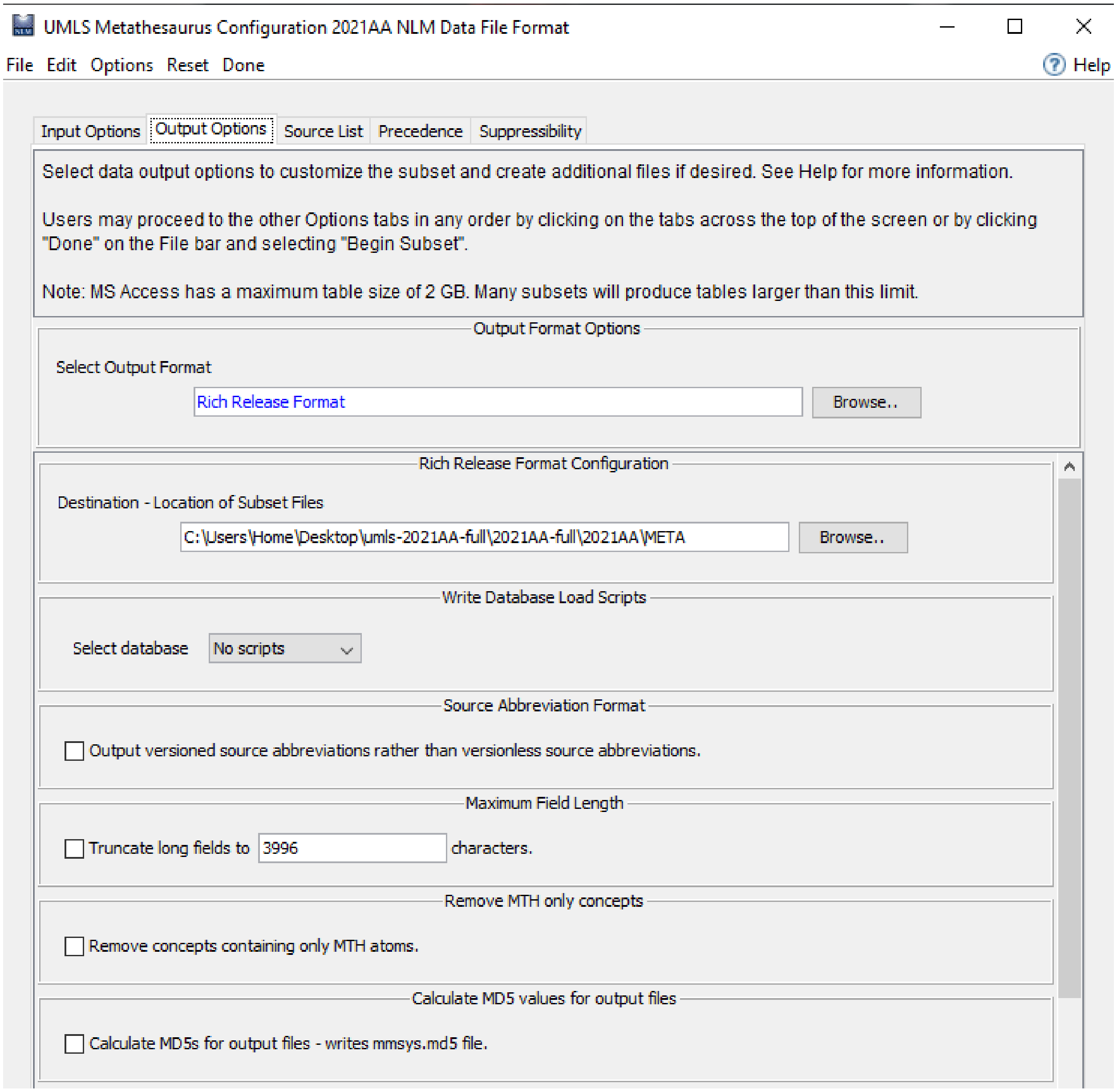
☑ Semantic Network -> Choose Database Load Scripts: None

☑ SPECIALIST Lexicon & Lexical Tools

Figure 3.3: UMLS Installation

3.4.2 Configuration

The figure shows the configuration process of meta-thesaurus. In order to extract data from met-thesaurus, we have to select the input and output format which can be set up in the configuration process.


UMLS Metathesaurus Configuration 2021AA NLM Data File Format
File Edit Options Reset Done
Help
Input Options
Output Options
Source List
Precedence
Suppressibility
Browse..
Browse..

Select data output options to customize the subset and create additional files if desired. See Help for more information.

Users may proceed to the other Options tabs in any order by clicking on the tabs across the top of the screen or by clicking "Done" on the File bar and selecting "Begin Subset".

Note: MS Access has a maximum table size of 2 GB. Many subsets will produce tables larger than this limit.

Output Format Options

Select Output Format: Rich Release Format

Rich Release Format Configuration

Destination - Location of Subset Files: C:\Users\Home\Desktop\umls-2021AA-full\2021AA-full\2021AA\META

Write Database Load Scripts

Select database: No scripts

Source Abbreviation Format

□ Output versioned source abbreviations rather than versionless source abbreviations.

Maximum Field Length

□ Truncate long fields to 3996 characters.

Remove MTH only concepts

□ Remove concepts containing only MTH atoms.

Calculate MD5 values for output files

□ Calculate MD5s for output files - writes mmsys.md5 file.

Figure 3.4: UMLS Configuration

3.5UMLS Meta-thesaurus

3.5.1 Overview

The UMLS Meta-thesaurus is enormous in size, and it is used for various aims. As it includes information and data in a number of different languages therefore there are multiple terms of a single concept in different languages. Moreover, the UMLS Meta-thesaurus consists of various terminologies, hierarchies, coding classes along with the concept, their names, their relations among each other. The meta-thesaurus comprises of almost hundred and seventy biomedical dictionaries which are also called the source vocabularies.

3.5.2 Scope of Meta-thesaurus

The concepts and their names/semantics are the core foundations of meta-thesaurus. To map the concepts with the similar semantics and domain in one concept, keeping in mind the expediency of various concepts and the relationship among them.

The range or scope of meta-thesaurus is determined on the basis of collective scope of all the concepts, names, attributes and relations in a meta-thesaurus. The concept names, relationships, definitions, synonyms are all associated with NLM but mainly all concepts and classes are originated from the source dictionaries. In case of a concepts not being presented in a source vocabulary, then it will not be displayed in meta-thesaurus even if it is a concept of medicine and health.

3.5.3 Content Preservation

The UMLS meta-thesaurus stores all the data related to the concepts in the source vocabularies. If such a situation arises where there is a same concept name in two different hierarchies, meta-thesaurus shows both and exhibits with their vocabulary name. In case a similar concepts appears in a different hierarchy; it shows both of them. Furthermore, if two concepts emerge in two separate hierarchies and an ambiguity occurs in their relationships, meta-thesaurus combines both opinions.

3.5.4 Need to Customize the meta-thesaurus

UMLS Meta-thesaurus is effective as it comprises of concepts in several languages and versatile UMLS resource from various terminologies. In order to retrieve the data competently, the information should be customized as it includes comprehensive data.

3.5.5 Release Formats of Meta-thesaurus

There are two types of release formats provided by UMLS and they can be selected at the time of application installation. These formats are the relational layouts and are named RRF and ORF respectively. RRF, also known as Rich Release Format, is developed in 2004 whereas ORF is also known as Original Release Format. One of these formats can be used to display the output in UMLS. In case of RRF (Rich Release Format) the file extension is .rrf whereas in case of ORF, it is.orf. When working with the decision systems, RRF layout is preferred layout for the output data.

3.5.5.1 Rich Release Format

Rich release format or RRF was first introduced in 2004. It presents excellent benefits in terms of clarity in source vocabulary. It implies it has the potential to offer suitable interpretations of concept name, categorical and source context data, characterize the widespread semantics of all source vocabularies with precision and has the capability of producing improved collections amongst forms of the meta-thesaurus with enhanced accuracy.

There are no normalized formats of meta-thesaurus. There is repetition of information between various files in a consequence. Hence, it is essential to build a conclusion so that the redundancy in information should be decreased or preserved relying upon the usage and its requirement in a specific application.

Meta-thesaurus consists of replicas, so we have discovered and decreased the redundancy in our expert system. In that way, our ontology will produce the accurate results.

3.5.6 Terminologies used in meta-thesaurus

In order to grasp the basic understanding of Metathesaurus and its composition, it is important to comprehend the terms and vocabularies used in UMLS.

Concepts, Concept Names, and Their Identifiers

According to the UMLS data, the meta-thesaurus is based on concepts. The concepts available in the meta-thesaurus consists of unique identifiers which are used for the classification of concepts. Furthermore, in order to link the various namespaces, present in the related concept, to the same concepts present in the terminologies, the unique identifiers are used. It is the main functionality of the identifier. Moreover, the meta-thesaurus provides distribution of several classes and their attributes according to the concepts available.

Concepts and Concept Identifiers

In terms of semantics and understanding, the word concept in UMLS is a term that has more than one meaning. The basic role of meta-thesaurus is to understand the meaning of all the concepts available in the specific vocabulary and to combine those concepts that have the same meaning. Every concept in meta-thesaurus has its own unique identifier also known as CUI. The value of CUI itself has no interpretation and cannot be used to deduce any semantics related to that particular concept. In other words, it is safe to say that any change in a label or term of a concept does not affects its CUI, which in each case remains the same. The table MRCONSO.RRF in the database shows the concept composition and how the data related to a concept can be visualized is shown in Figure 2. In many cases, several files contain the Atom Unique Identifier also known as AUI as well as Concept Unique Identifier or CUI. If searched at the outset, the table MRCONSO.RRF displays information regarding all the atoms present in a particular concept. Furthermore, all the attributes, relationships among the concepts and its attributes and hierarchies are also yielded if searched thoroughly. Moreover, in case of MRREL.RRF, if we conduct search according to CUI, another field named CUI2 will be discovered which is also available in MRCONSO.RRF. The other concept which is associated with the particular concept is CUI2 and is identified using this search technique.

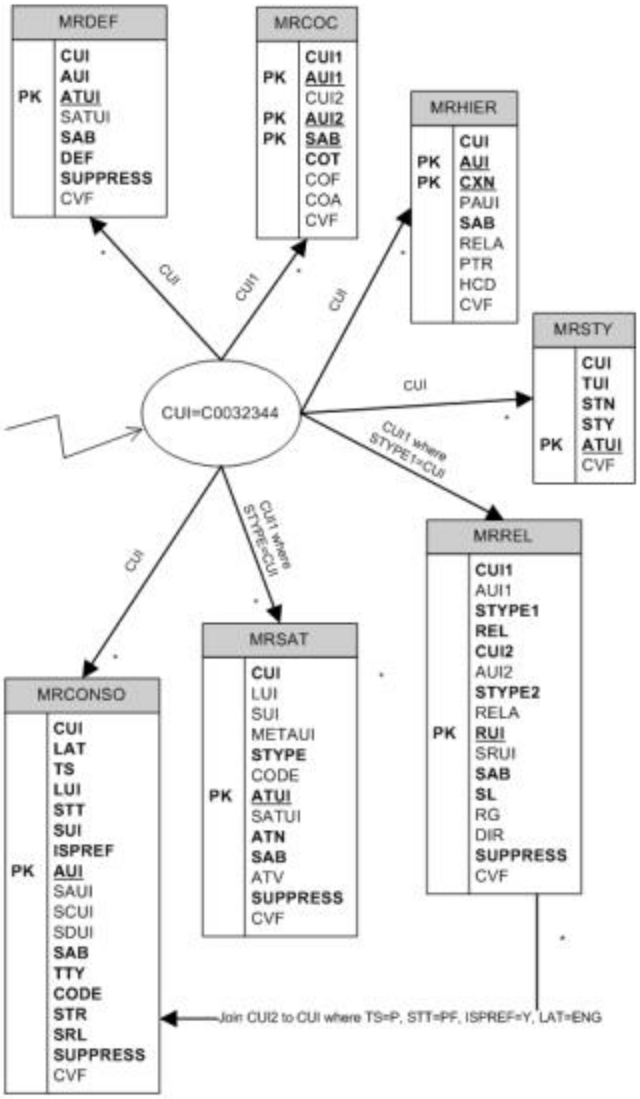


Figure 3.5: Concept Unique Identifier Associations

Concept Names and String Identifiers

In meta-thesaurus, every concept name or string, no matter from what language, consists of a distinctive and constant attribute named String Unique Identifier or SUI. If it has a different character set or there is a variation in its punctuation or even if it has a difference in upper-lower case values, then it is considered as a different and independent string containing a different and independent SUI. It is usually identified with a string containing S in the beginning subsequently followed by seven numerals.

Concept name indexes

The following figure 3 shows the three concept names and indexes

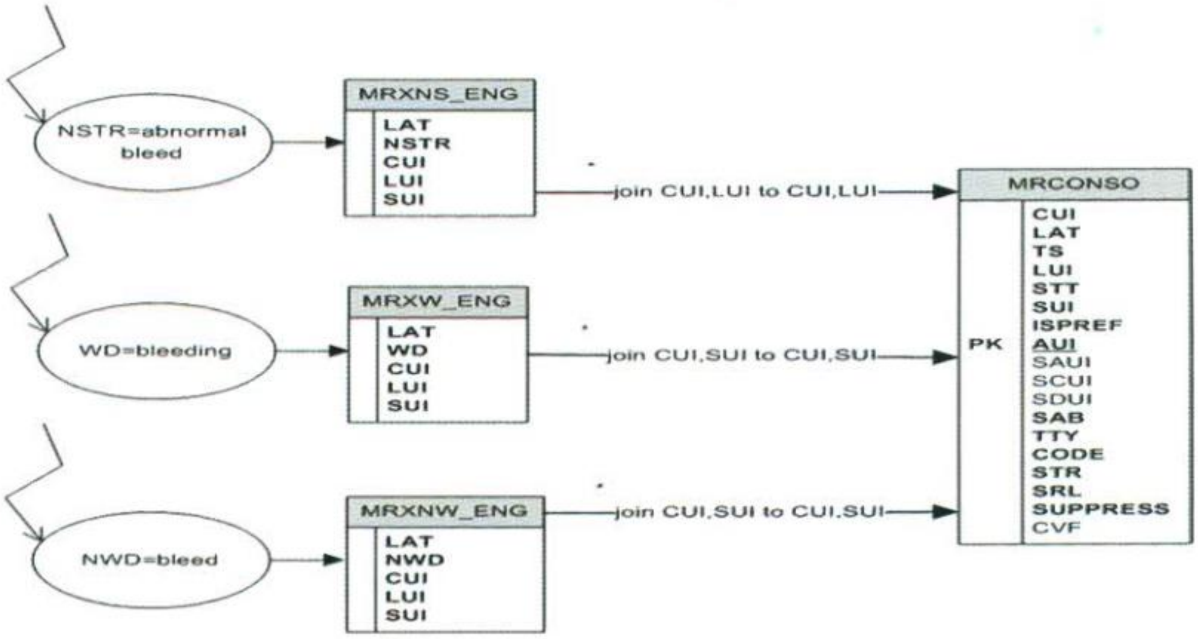


Figure 3.6: Concept Names and Indexes

Atoms and Atom Identifier

The structure of meta-thesaurus is built on the basis of basic and primary component termed as atom. Atoms are described as the names or labels along with their strings present in all the vocabularies available. A unique identifier is assigned to each string present in the vocabulary named Atom Unique Identifier or AUI. As a particular string in the vocabulary can have only one semantic meaning, therefore, every concept identifier or CUI has a relation with a unique atom identifier or AUI. Furthermore, the occurrence of AUIs is only limited to .RRRF files and are not available in .ORF files.

Terms and Lexical Identifiers

The strings present in the meta-thesaurus are interconnected to each of the lexical variants termed as LUI or Lexical Unique Identifier. The LUI is available only for the terms of English language, therefore, if an entry is described in English language, then it is a collection of all the strings which are each other's lexical variants. Moreover, for every AUI or a string there exists only a single LUI. The method to identify the lexical variants is using LVG program or Lexical Variant Generator program, which is considered to be one of the lexical tools of UMLS. The value of LUI in the database consists of the letter L subsequently followed by seven numerals.

Uses of Concept, String, Atom and Term Identifiers

In meta-thesaurus, every CUI or Concept Identifier is combined with one or more SUI or String Identifier, AUI or Atom Identifier, along with LUI or Lexical Identifier. Every AUI has a relationship with one CUI, a single SUI along with a distinct LUI. An SUI can have more than one CUIs and AUIs but when it comes to LUI, there exists only one distinct LUI. Moreover, every LUI can be connected to multiple SUIs, AUIs and CUIs.

A truncated instance is stated in the table below. Atrial Fibrillation is assigned a unique AUI for each circumstance and is described as an atom in numerous source vocabularies. As each of these atoms have either an indistinguishable concept name or string therefore, they are linked to a particular SUI.

Concept (CUI)	**Terms (LUIs)**	**Strings (SUIs)**	**Atoms (AUIs)** *** RRF Only**
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation	**L0004238** Atrial Fibrillation (preferred) Atrial Fibrillations	**S0016668** Atrial Fibrillation (preferred)	**A0027665** Atrial Fibrillation (from MSH) **A0027667** Atrial Fibrillation (from PSY)

Auricular Fibrillations		**S0016669** (plural variant) Atrial Fibrillations	**A0027668** Atrial Fibrillations (from MSH)
	C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation (preferred)	**A0027930** Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	**A0027932** Auricular Fibrillations (from MSH)

Table 3.2: An Instance of CUIs, LUIs, SUIs and AUIs in UMLS

The concept Atrial Fibrillations has a distinct string identifier because it is actually a different concept as it is a plural of atrial fibrillation. But it has one LUI. In case of Auricular Fibrillation and its plural Auricular Fibrillations, there exists separate LUI and several AUIs and SUIs. For instance, knowledge of every comprehensive concept is included in CUIs, or it can be said that in order to access all the concept names, relationships and attributes of a particular concept, a CUI is used.

In addition to all these points, a CUI must be used in order to retrieve the concepts related to medicine and health case, that are further interlinked with numerous source vocabularies.

3.5.7 Relationships and Relationship Identifiers

There is a concept of ‘synonymous relationships’ in terms of structuring of meta-thesaurus. A lot of these are mined from well-defined source vocabularies. Though NLM incorporates a few of the relationships of concepts during the formation of Meta-thesaurus. The consumers of Meta-thesaurus platform provide many of these relationships in order to maintain specific kinds of applications.

The relationships are denoted in the form of CUIs in case of RRF and ORF, however, in case of RRF, only AUIs are used. The format of every relationship is identified by Meta-thesaurus largely, including an issue of Meta-thesaurus or another provider, source vocabularies. A few of the relationships were enhanced in the course of early phases of Meta-thesaurus creation. These are recognized by the Meta-thesaurus, although they belong to different source vocabularies.

3.5.7.1 Basic Categories of Non-synonymous Relationships

The relationships established by the source vocabularies associated closely related concepts like the relationships which allocate any common attribute or maybe they are connected in terms of definition. For example, amoxicillin belongs to the class of drugs and will be related to the class named antibiotics. In the same way, any viral disease will be associated with the class of virus which causes it.

3.5.7.2 Intra-source relationships

As it can be seen from the names or labels of the concepts that these relations belong to the same vocabulary. The relations are available in the MRREL table in the database and consists of both the statistical and hierarchical relationships for instance in case of the concepts that occurred in the form of main topic are placed in the category of co-occurrence relationships.

3.5.7.3 Inter-source relationships

The association of atoms and concepts from a vocabulary to the atom and concept relation of another vocabulary are termed as the inter-source relationships. These types of relationships are of great value when it comes to the concepts which are not associated to any concept. In other words, they have no ancestral or child relations in the vocabulary and are termed as Orphan concepts. These orphan concepts are associated with the relations in other source vocabularies, hence making it an effective feature of Meta-thesaurus. MRREL table consists of the inter-source relationships.

Moreover, the UMLS meta-thesaurus also consists of mappings which are located in the tables MRMAP and MRSMA. Mappings are required in order to form a connection between different source vocabularies. They are usually formed under the management of NLM itself by a third party.

3.5.7.4 Relationship labels

The relationships that are not a part of any fundamental concept structure are described as REL. This defines and identifies the kind of relationships for instance, broad, narrow, parent of, successor of etc. In most of the cases these relations are either presented directly or are identified indirectly by the source vocabulary. The details of labels regarding the common relationship are available in a UMLS Meta-thesaurus table named MRDOC.RRF. In order to identify the relationships of the nature “is a”, “component of”, “branch of”, another additional label named RELA is also available in Meta-thesaurus. It caters around 25% of the relationships in UMLS Meta-thesaurus.

3.5.7.5 Relationship Identifiers

In Meta-thesaurus, every relationship consists of distinct identifier named as RUI. This identifier keeps the record of all, or any changes panned out in the relationships within the different versions of Meta-thesaurus. If an RUI is present or absent in the Meta-thesaurus relationship, it depicts a change has occurred in the relationships. A few of the source vocabularies comprises of their own identifiers or in this case, RUIs.

3.5.8 Attribute and Attribute Identifier

All the information or data related to a concept is stored in the attributes and attributes identifiers. In Meta-thesaurus, they are denoted by AUI also known as Atom Unique Identifiers.

3.5.8.1 Attribute kinds

In UMLS Meta-thesaurus, there exists three kinds of attributes

Atom Attributes (AUI)

Concept Attribute (CUI)

Relationship Attribute (RUI), further divided into RELA and REL

Following are the details of these attributes

3.5.8.1.1 Concept Attribute

The concept attributes are included in the meta-thesaurus at the time of its formation. It contains all the names of a particular concept for instance, “Atrial Fibrillation” is the name of a concept, and it has the semantic types “Genetic Function” and “Carbohydrate Sequence”. Every atom related to this concept has the same sematic types.

3.5.8.1.2 Atom Attribute

An atom is an essential component of a concept in Meta-thesaurus. Every existing concept of Meta-thesaurus is related to an atom or a number of atoms and their usage changes application wise.

3.5.8.1.3 Relationship Attribute

These are the attributes that belongs to a specific source vocabulary and depict unique qualities of that specific relationship in the source vocabulary.

3.5.8.1.4 Attribute Identifiers

The appearance of every attribute in UMLS Meta-thesaurus is denoted with a distinct identifier named ATUI or Attribute Unique Identifier. If an ATUI is present or absent in the Meat-thesaurus, then it depicts that the changes are occurred in the Meta-thesaurus. Furthermore, ATUIs are only available in the RRF format and is not found in ORF format.

3.5.9 Meta-thesaurus Metadata

The data about Meta-thesaurus is termed as Metadata and it refers to the following points:

- Qualities of the existing edition of the Meta-thesaurus
- Modifications between the existing version and the prior version
- The complete record of CUIs from the year 1991 to the present.

3.5.10 Data Files

Meta-thesaurus contains a huge amount of data and can only be articulated through numerous files or relation. This information is divided into four classes and indexes in correspondence to these files and are mentioned below:

- Concepts, their names and source attributes
- Relationships
- Metadata/data about Meta-thesaurus
- Indexes

3.6 Summary

Clean and pre-processed always produces valid and precise results whereas dirty and unprocessed data does not produce accurate results. In order to create an ontology, the quality of data matters a lot as the efficiency and trustworthiness of an ontology depends on this data. The input data must be of good quality on order for an ontology to provide reasonable coverage. This data acts as the building block of an ontology. The data source we have selected is UMLS which is both reliable and enough as well. All the concepts related available in Meta-thesaurus are gathered and the procedure of loading the data into database and development of an ontology are discussed in the upcoming chapters.

CHAPTER 4

DATA LOADING AND TRANSFORMATION

4. Data Loading and Transformation

In the previous chapter, we discussed the process of collection of data and extraction technique used for the data gathering in our ontology building. This chapter will discuss the process of loading that extracted data into the database and transforming that loaded data in the database to our requirement. The RDBMS data storage and the transformation of data in order to make it standardized to produce accurate results is discussed in detail in this chapter

4.1Data Loading

The data gathered from UMLS resources is available in the textual format. In order to make it effective and easy to use, we have to understand the data and then load it into the relational database or in other words RDBMS. The RDBMS used in this research is MySQL as it is free to use, open-source as well as used widely among all the database users.

4.2Lexical Data Loading into Mysql Database

The UMLS lexical tools of NLM are used in many clinical NLP methodologies in order to conduct some part of processing. For instance, if we talk about MRXNS_ENG, which is an index file of Meta-thesaurus, this file can only be utilized when converted into the right format.

A script is developed in order to load all the data into the database. Following figure displays the process of loading the data into the database.

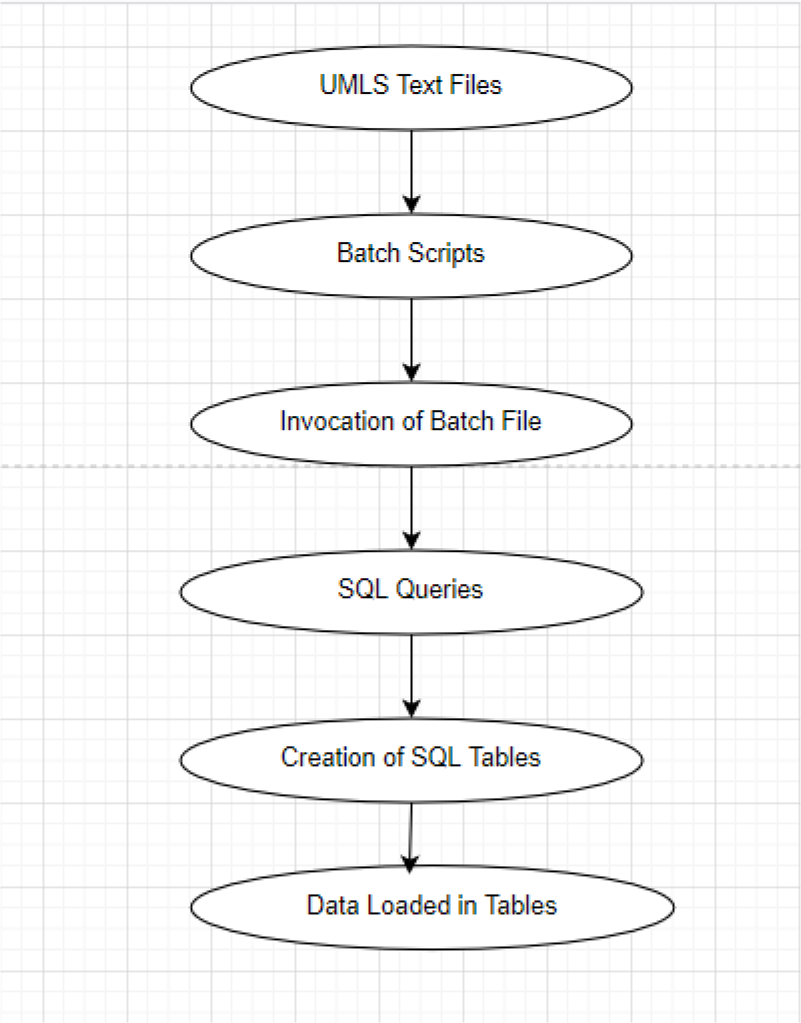


Figure 4.1: Data loading into the Database

The steps used in order to load the data into the database available locally are mentioned as under:

1. Download all the UMLS resources required for our ontology from the official website of UMLS[81], by singing up and registering for a license as discussed above. The downloaded file consists of the following eight files:
    - 2021AA.CHK
    - 2021AA.MD5
    - 2021AA-1-meta.nlm
    - 2021AA-2-meta.nlm
    - 2021AA-otherks.nlm
    - mmsys.zip

- Copyright_Notice.txt
- README.txt

2. The zipped file named mmsys.zip is unzipped into the main directory and Metamorphosys is executed. At the start of the Metamorphosys, we have set up input and output destinations. The Metamorphosys configuration results in the creation of three subfolders namely LEX, NET and META.

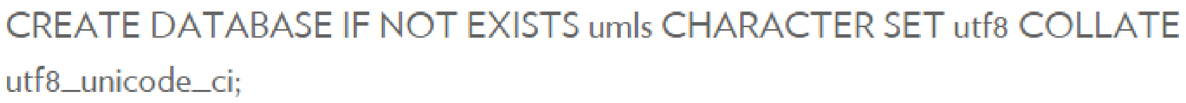
4.3Rich Release Format Mysql Load Script

The database is created using the batch scripts in mysql. The rich release format also known as RRF is loaded into the UMLS meta-thesaurus. In order to generate the necessary files, a subset is developed, and Write Database Load Scripts is checked in Output Options tab. These files are then created by selecting the subclass. In the welcome screen of the Metamorphosys, advanced options are chosen.

4.4Creating the Database

In order to load the UMLS data into the mysql database, an existing database can be used, or a new database can be created. For this research, we have created a new database in order to avoid redundancy and data duplication. To create a mysql database, there are two main observations in this regard. Setting the character-set to default along with the scenarios related to classification. The settings UTF-8 and utf8_unicode_ci are recommended.

- Metamorphosys is configured in order to generate the lexical tools, create the load scripts for MySQL. furthermore, English language is selected as the language for terms from the data source.
- In the next step, we navigated to the folder named META and run mysql file with admin privileges. In this way, a UMLS database can be created easily by querying the create command.



```
CREATE DATABASE IF NOT EXISTS umls CHARACTER SET utf8 COLLATE
utf8_unicode_ci;
```

Figure 4.2: Database Creation Query

4.5Performance Parameters

The configuration file available in mysql server installation folder is modified according to the requirement of our desired database. This file is named as “my.ini” or “my.cnf”. the parameters like buffer size, table cache etc are changed and set to the required setting in order to optimize the mysql server memory usage and increasing efficiency.

- Adjusting the key-buffer is the foremost step in the configuration settings of mysql. the key-buffer stores the indexes in the memory. In order to store large amount of data, the size of buffer should be increased. It can be done with the statement key_buffer = 600M in case of setting the buffer size to 600 MB.
- The table_cache is set to 300 in order to make sure the number of tables in a database can be opened simultaneously, hence improving the efficiency by completing multiple requests at a time. The higher the size of table_cache, the higher the reliability and query completion time.
- Next parameter that needs to be adjusted is sort_buffer_size. This parameter monitors the size of buffer which is generated when mysql is supposed to sort rows in the records. In order to sort the UMLS data, which is huge in size, the size of sort_buffer_size is increased up to 500MB using the statement sort_buffer_size = 500MB.
- Another parameter setting which is important to adjust is read_buffer_size. Its purpose is to adjust the size of memory in order to distribute it to all the sequential assessment of data within mysql records. The size of read_buffer_size is set to 200MB using the statement read_buffer_size = 200M.
- The next parameter is query_cache_limit which decides that the results produced from the query that can be cached will be of what size. 1 MB is the default size of query_cache_limit which is changed to 3MB in order to cater the query results.
- Query_cache_size is the next parameter that needs to be adjusted. This parameter modifies the memory size generally accessible for keeping the query cache. It can be said that the size of database and this parameter’s value are dependent on each other. In case of a large sized database, this parameter must also have a large size. The default value of query_cache_size is 0M which is adjusted to 100M. the size 0 M indicates that this parameter was disabled by default.

- The parameter termed as myisam_sort_buffer_size is the next parameter to be adjusted. The function of this parameter is to retain the buffer size that generated the indexes and catalogs over myisam records. This parameter is quite useful while loading of tables. The statement used to adjust the value of this parameter is myisam_sort_buffer_size = 200M.
- Bulk_insert_buffer_size is the next parameter to be adjusted. It is used to store the resultant data from running insert query. The size of this parameter is adjusted to 100M through the statement bulk_insert_buffer_size = 100M.
- The parameter named join_buffer_size helps in querying the non-indexed joins in the tables. Increasing the value of this parameter results in high efficiency and fast querying results. The value of this parameter is set to 100M.

4.6Loading Data in Database

The data loading process is carried out using queries of create database and create tables. After the creation of databases and tables, textual data files are loaded into those tables. The queries and SQL commands used to perform this task are mentioned as under. Furthermore, the SQL commands are executed using the main batch script and are stored in a single file.

4.6.1 Data Table Creation

The SQL command used to create a table is discussed as under.

```
DROP TABLE IF EXISTS MRCONSO;
```

This above mentioned command is used to make sure that no other table with the same name exists already in the database. In order to create a table, for instance MRCONSO, following create command can be used. The first word is the name of column, for instance “CUI”, the next word is the type of data to be load, for instance “char”. The numeral in the brackets represent the length of entered data whereas the clauses like NOT NULL indicates that the particular field is not supposed to be vacant. In other words, the fields marked as NOT NULL must not be left empty.

```
CREATE TABLE MRCONSO (
   CUI char(8) NOT NULL,
    LAT	char(3) NOT NULL,
    TS	char(1) NOT NULL,
    LUI	varchar(10) NOT NULL,
    STT	varchar(3) NOT NULL,
    SUI	varchar(10) NOT NULL,
    ISPREF char(1) NOT NULL,
    AUI	varchar(9) NOT NULL,
    SAUI varchar(50),
    SCUI varchar(100),
    SDUI varchar(100),
    SAB	varchar(40) NOT NULL,
    TTY	varchar(40) NOT NULL,
    CODE varchar(100) NOT NULL,
    STR	text NOT NULL,
    SRL	int unsigned NOT NULL,
    SUPPRESS char(1) NOT NULL,
    CVF	int unsigned
)
```

Using this technique, we have created 47 tables by using an SQL command batch script named as mysql_tables.sql. This file, consisting of 47 table creation commands, is invoked and all the tables are created in a sequence.

4.6.2 Loading Data from Textual Files

For loading the data into the table in a database from textual files, following command is used. By using this command, all the data will be loaded into the MRCONSO.RRF table and on the basis of '|', the data is analyzed by this command after which the data can be loaded into the MRCONSO.RRF table.

```
CHARACTER SET utf8;


load data local infile 'MRCONSO.RRF' into table MRCONSO fields terminated by '|' ESCAPED BY " lines terminated by '\r\n'

(@cui,@lat,@ts,@lui,@stt,@sui,@ispref,@aui,@saui,@scui,@sdui,@sab,@tty,@code,@str,@srl,@suppress,@cvf)

SET CUI = @cui,

LAT = @lat,

TS = @ts,

LUI = @lui,

STT = @stt,

SUI = @sui,

ISPREF = @ispref,

AUI = @aui,

SAUI = NULLIF(@saui,"),

SCUI = NULLIF(@scui,"),

SDUI = NULLIF(@sdui,"),

SAB = @sab,

TTY = @tty,
CODE = @code,

STR = @str,

SRL = @srl,

SUPPRESS = @suppress,

CVF = NULLIF(@cvf,");
```

This process helps enables us to create all 47 tables in the database automatically by invoking the file containing all the SQL commands, named as mysql_tabes.sql.

4.7Data Indexing

Data indexing is a task which is required while dealing with a huge bulk of data including millions of records. This task involves the writing of extra lines of code along with additional memory space and results in fast execution of queries hence making the system more efficient to use. The process of indexing does not search every table or row from the database and locates the desired field or record very fast. In order to create an index, one or more columns in the database record are used. In our research, we have developed a script that creates all the indexes on the tables in a database automatically when invoked.

```
CREATE INDEX X_MRCONSO_CUI ON MRCONSO(CUI);


ALTER TABLE MRCONSO ADD CONSTRAINT X_MRCONSO_PK  PRIMARY KEY
BTREE (AUI);


CREATE INDEX X_MRCONSO_SUI ON MRCONSO(SUI);


CREATE INDEX X_MRCONSO_LUI ON MRCONSO(LUI);


CREATE INDEX X_MRCONSO_CODE ON MRCONSO(CODE);
```

CREATE INDEX X_MRCONSO_SAB_TTY ON MRCONSO (SAB, TTY);

CREATE INDEX X_MRCONSO_SCUI ON MRCONSO(SCUI);

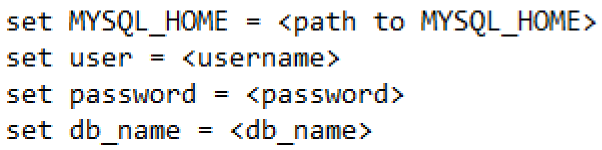
CREATE INDEX X_MRCONSO_SDUI ON MRCONSO(SDUI);

CREATE INDEX X_MRCONSO_STR ON MRCONSO(STR(255));

4.8Database Batch Script

As we have already discussed in the above heading that we have developed two files containing SQL commands named “mysql_tables.sql” and “mysql_indexes.sql”. The file named “mysql_tables.sql” consists of all the SQL commands required to create the tables as well as loading the data into the database. Whereas the file called “mysql_indexes.sql” comprises of the SQL commands which creates all the indexes on the records available in a database. These two files need to be executed in order to run those SQL commands and create tables and indexes in a sequence. For this purpose, a batch script is prepared that invokes these SQL files and runs all the SQL commands in the files in order to create the tables and indexes chronologically. The parameters and other relevant details regarding the said batch script are discussed below.

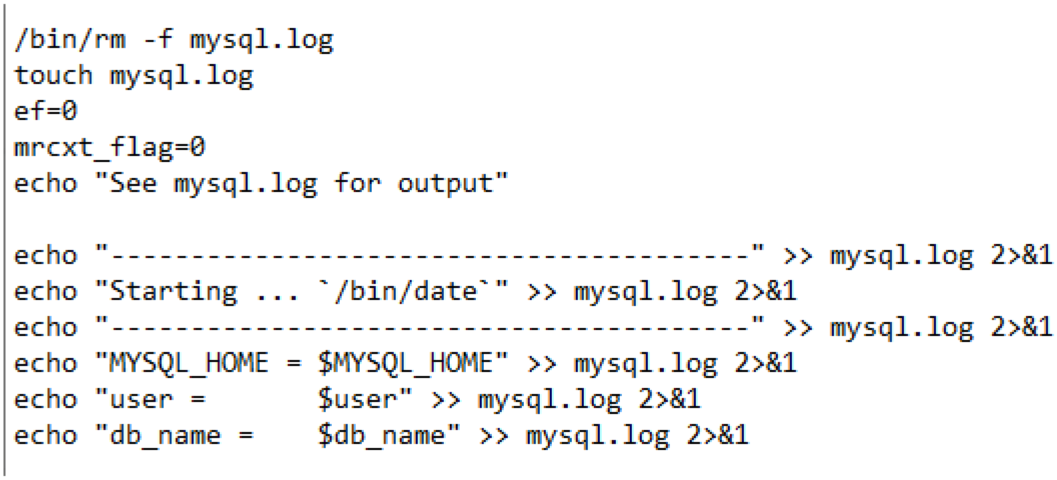
To connect the database, following variables must be defined



```
set MYSQL_HOME = <path to MYSQL_HOME>
set user = <username>
set password = <password>
set db_name = <db_name>
```

Figure 4.3: Database Parameters

Log files in the mysql directory must be initialized along with the joining of primary information

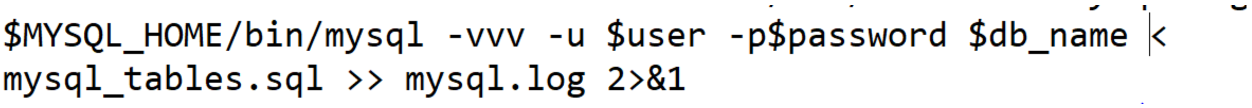


```
/bin/rm -f mysql.log
touch mysql.log
ef=0
mrcxt_flag=0
echo "See mysql.log for output"

echo "----------------------------------------" >> mysql.log 2>&1
echo "Starting ... `/bin/date`" >> mysql.log 2>&1
echo "----------------------------------------" >> mysql.log 2>&1
echo "MYSQL_HOME = $MYSQL_HOME" >> mysql.log 2>&1
echo "user =       $user" >> mysql.log 2>&1
echo "db_name =    $db_name" >> mysql.log 2>&1
```

Figure 4.4: Initializing log files

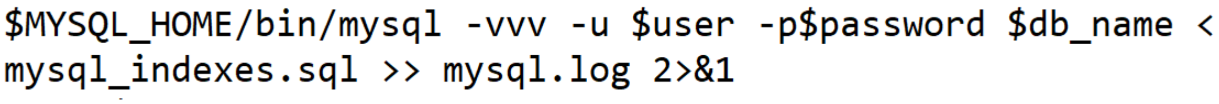
The following command will execute the file called “mysql_tables.sql” which in turn creates all the tables in the database as well as loading the data into the database. The SQL commands available in the file named “mysql_tables.sql” will be executed in order for the tables to be created and data to be loaded.



```
$MYSQL_HOME/bin/mysql -vvv -u $user -p$password $db_name <
mysql_tables.sql >> mysql.log 2>&1
```

Figure 4.5 Invoking mysql_tables.sql

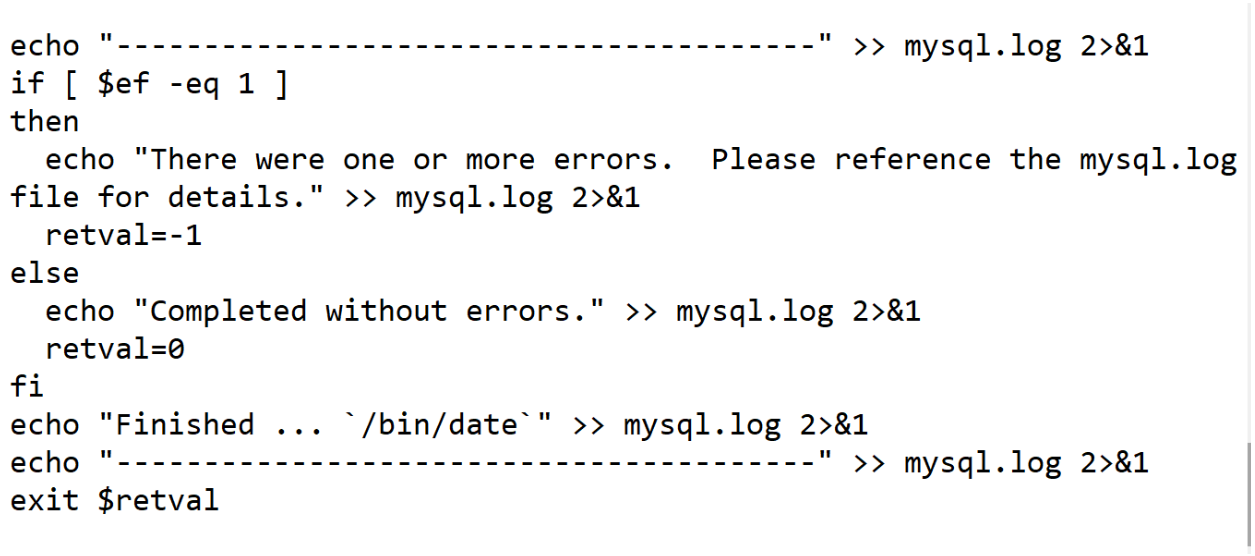
The following command invokes the file named “mysql_indexes.sql” which creates all the indexes on the tables in the database. The SQL commands present in this file are executed by running the following script.



```
$MYSQL_HOME/bin/mysql -vvv -u $user -p$password $db_name <
mysql_indexes.sql >> mysql.log 2>&1
```

Figure 4.6 Invoking mysql_indexes.sql

The following command displays and records errors



```
echo "-----------------------------------------" >> mysql.log 2>&1
if [ $ef -eq 1 ]
then
  echo "There were one or more errors.  Please reference the mysql.log
file for details." >> mysql.log 2>&1
  retval=-1
else
  echo "Completed without errors." >> mysql.log 2>&1
  retval=0
fi
echo "Finished ... `/bin/date`" >> mysql.log 2>&1
echo "-----------------------------------------" >> mysql.log 2>&1
exit $retval
```

Figure 4.7: Command to display errors

4.9Data Transformation

The data cleaning and removal of duplicates was accompanied by the process of data gathering in the database. The transformation of data in the tables was not a hectic task as UMLS has a completely transformed knowledge base and the data available is structured properly. In many cases, the transformation process was required only in terms of relationships. At some points, relations were stored in the form of abbreviations for instance, “PAR” is a stored relation in the database which means “is parent of”. These types of transformations were carried out in the same format.

4.10 Table Structure

The tables created in the database by running the batch script files are discussed as under:

4.10.1MRCONSO.RRF

This table caters all the features and explanation regarding concepts in the meta-thesaurus. Furthermore, it also elaborates how each concept in relation or associated with the CUIs in the UMLS meta-thesaurus as well as the source vocabularies and languages.

Following are the attributes available in MRCONSO.RRF

CUI	Concept Unique identifier
LAT	Term Language
TS	Term status
LUI	Term Unique identifier
STT	String type
SUI	String Unique identifier
ISPREF	Atom status
AUI	Atom Unique identifier
SAUI	Source asserted atom identifier
SCUI	Source asserted concept identifier

SDUI	Source asserted descriptor identifier
SAB	Source name abbreviation
TTY	term type abbreviation
CODE	source asserted identifier
STR	String
SRL	Source restriction level
SUPPRESS	Suppressible flag
CVF	Content View Flag.

Table 4.1: MRCONSO.RFF

4.10.2MRDEF.RRF

The definitions of all the concepts available in the UMLS meta-thesaurus are available in this table. Following is the structure of the table MRDEF.RRF

CUI	Concept Unique identifier
AUI	Atom Unique identifier
ATUI	Attribute Unique identifier
SATUI	Source asserted attribute identifier

SAB	Source name Abbreviation
DEF	Definition
SUPPRESS	Suppressible flag.
CVF	Content View Flag.

Table 4.2: MRDEF.RFF

4.10.3MRFILES.RRF

All the files available in meta-thesaurus along with its details are present in this table. This table covers the details such as explanation, column names, total number of rows and columns along with their sizes etc. The structure of this table is as under:

FIL	File name
DES	Descriptive Name
FMT	Comma separated list of column names
CLS	number of COLUMNS
RWS	Number of ROWS

BTS	Size in bytes

Table 4.3: MRFILES.RFF

4.10.4MRCOLS.RRF

The data related to the columns are available in this table. All the information regarding the columns in the UMLS meta-thesaurus tables are present in this table. Following is the structure of this table:

COL	Column name
DES	Descriptive Name
REF	Documentation Section Number
MIN	Minimum Length
AV	Average Length
MAX	Maximum Length
FIL	File name

DTY	data type

Table 4.4: MRCOLS.RFF

4.10.5MRSAT.RRF

Following is the table structure of MRSAT.RRF

CUI	Concept Unique identifier
LUI	Term Unique identifier
SUI	String Unique identifier
METAUI	Meta-thesaurus atom identifier
STYPE	String type
CODE	source asserted identifier
ATUI	Attribute Unique identifier

SATUI	Source asserted attribute identifier
ATN	Attribute name.
SAB	Abbreviated source name
ATV	Attribute value.
SUPPRESS	Suppressible flag
CVF	Content View Flag.

Table 4.5: MRSAT.RFF

4.10.6MRSTY.RRF

All the semantic types present in the UMLS meta-thesaurus with respect to available concepts are present in this table. Following is the structure of this table

CUI	Concept Unique identifier
TUI	Semantic type Unique identifier
STN	Semantic Type tree number
STY	Semantic Type.

ATUI	Attribute Unique identifier
CVF	Content View Flag.

Table 4.6: MRSTY.RFF

4.10.7MRHIST.RRF

The previous records and information related to history are present in this table. Following structure is followed in this table

CUI	Concept Unique identifier
SOURCEUI	Source asserted unique identifier
SAB	Abbreviated source name (SAB).
SVER	version number of a source
CHANGETYPE	Source asserted code
CHANGEKEY	Concept status
CHANGEVAL	Concept status after change
REASON	Explanation of change if present
CVF	Content View Flag.

Table 4.7: MRHIST.RFF

4.10.8MRREL.RRF

The relations between the concepts are stored in this table. Both narrower and broader relations are covered in this table with attributes like REL and RELA. The concepts between which relation are present are distinguished by the attributes like CUI1 and CUI2. Following is the structure of this table.

CUI1	Concept Unique identifier 1
AUI1	Atom Unique identifier 1
STYPE1	The name of the column in MRCONSO.RRF that contains the identifier used for the first element in the relationship
REL	Relationship of second concept
CUI2	Concept Unique identifier 2
AUI2	Atom Unique identifier 2
STYPE2	The name of the column in MRCONSO.RRF that contains the identifier used for the second element in the relationship
RELA	relationship label

RUI	Relationship Unique identifier
SRUI	Source asserted relationship identifier
SAB	Abbreviated source name
SL	Source of relationship labels
RG	Relationship group.
DIR	Source asserted directionality flag.
SUPPRESS	Suppressible flag.
CVF	Content View Flag.

Table 4.8: MRREL.RFF

4.10.9MRCOC.RRF

This table contains the compilation of connotations that co-appears in the external sources. Following is the structure of this table.

CUI1	Concept Unique identifier 1
AUI1	Atom Unique identifier 1
CUI2	Concept Unique identifier 2

AUI2	Atom Unique identifier 2
SAB	Abbreviation of the source
COT	Type of co-occurrence
COF	Frequency of co-occurrence
COA	Attributes of co-occurrence
CVF	Content View Flag

Table 4.9: MRCOC.RFF

4.10.10 MRHIER.RRF

The atoms and concepts that exists in the UMLS meta-thesaurus follow some hierarchical formation. Those hierarchies are catered in this table. Following is the table structure of MRHIER.RRF

CUI	Concept Unique identifier
AUI	Atom Unique identifier

CXN	Context number
PAUI	Unique identifier of atom's immediate parent
SAB	Abbreviated source name
RELA	Relationship of atom to its immediate parent
PTR	Path
HCD	Source asserted hierarchical number
CVF	Content View Flag.

Table 4.10: MRHIER.RFF

4.10.11 MRCXT.RRF

This table also contains the hierarchies of atoms and concepts and can be used as a replacement to the table MRHIER.RRF. It has the following structure.

CUI	Concept Unique identifier
SUI	String Unique identifier
AUI	Atom Unique identifier

SAB	Abbreviated source name
CODE	Unique identifier or code for string
CXN	The context number
CXL	Context member label
RNK	rank of the ancestors
CXS	String or concept name for context member
CUI2	Concept identifier 2
AUI2	Atom identifier 2
HCD	Source hierarchical number
RELA	Additional relationship label.
XC	indicates children
CVF	Content View Flag.

Table 4.11: MRCXT.RFF

4.10.12 MRMAP.RRF

The mapping information between the various vocabularies is stored in this table. In most of the cases, the mappings are drawn between the unique identifiers and codes from various vocabularies. Following structure is available in the table MRMAP.RRF

MAPSETCUI	map set Unique identifier.
MAPSETSAB	map set Source abbreviation

MAPSUBSETID	Map subset identifier
MAPRANK	map set Order
MAPID	Unique identifier
MAPSID	Mapping Source asserted identifier
FROMID	Identifier for the entity being mapped from.
FROMSID	Mapped Source asserted identifier
FROMEXPR	Entity being mapped from
FROMTYPE	Type of entity being mapped from.
FROMRULE	Machine processable rule
FROMRES	Restriction

REL	Relationship
RELA	Additional relationship label
TOID	Identifier for the entity being mapped to.
TOSID	Source asserted identifier
TOEXPR	Entity being mapped to
TOTYPE	Type of entity being mapped to.
TORULE	Machine processable rule
TORES	Restriction
MAPRULE	Machine processable rule
MAPRES	Restriction

MAPTYPE	Type of mapping
MAPATN	Mapping name of the attribute
MAPATV	Mapping value of the attribute
CVF	Content View Flag

Table 4.12: MRMAP.RFF

4.10.13 MRSMAP.RRF

The mapping information is available in this table but is considered straightforward and uncomplicated as compared to that of represented in table MRMAP.RRF. This table has the following column structure.

MAPSETCUI	map set Unique identifier
MAPSETSAB	map set Source abbreviation
MAPID	Mapping Unique identifier
MAPSID	Source asserted identifier

FROMEXPR	Entity being mapped from
FROMTYPE	Type of entity being mapped from.
REL	Relationship
RELA	Additional relationship label.
TOEXPR	Entity being mapped to
TOTYPE	Type of entity
CVF	The Content View Flag

Table 4.13: MRSMAP.RFF

4.10.14 MRSAB.RRF

The source abbreviations also known as SAB available in data files are represented in this table but in a root format. Following are the columns in the table MRSAB.RRF

VCUI	CUI

RCUI	Root CUI
VSAB	Versioned Source Abbreviation
RSAB	Root Source Abbreviation
SON	Official Name
SF	Source Family
SVER	Version
VSTART	Meta Start Date
VEND	Meta End Date
IMETA	Meta Insert Version
RMETA	Meta Remove Version

SLC	Source License Contact	
SCC	Source Content Contact	
SRL	Source Restriction Level	
TFR	Term Frequency	
CFR	CUI Frequency	
CXTY	Context Type	
TTYL	Term Type List	
ATNL	Attribute Name List	
LAT	Language	
CENC	Character Encoding	

CURVER	Current Version	
SABIN	Source in Subset	
SSN	Source Short Name	
SCIT	Source Citation	

Table 4.14: MRSAB.RFF

4.10.15 MRRANK.RRF

In order to create a new customized application, this table is useful as it contains dataset and is useful for developers. Following are the columns in this table.

RANK	Numeric order of precedence
SAB	Abbreviated source name
TTY	term type
SUPPRESS	Source and Term Type

Table 4.15: MRRANK.RFF

4.6.15 AMBIGSUI.RRF

All the pairs of SUI-CUI present in the UMLS meta-thesaurus are represented in this table. It has the following structure.

SUI	String Unique Identifier
CUI	Concept Unique Identifier

4.11 Meta-thesaurus Change Files

The meta-thesaurus versions are revised on a quarterly basis. These change records are stored in the six tables present in Meta-thesaurus. The purpose of this data is to clarify the variations among the prior and current record. The differences seen and made throughout the version releases is stored in these tables along with the information like what CUIs are still being used and which are being withdrawn and are very useful for the developers to develop the applications and other systems. Following are the details of these tables.

4.11.1DELETEDCUI.RRF

All the concepts that are deleted from the meta-thesaurus versions are stored in this table. All the related information about the concepts and its unique identifier can be found in this table. For instance, if a CUI was available in the previous version but is deleted in the current one, then its record can be found in the form of a row and if a concept was deleted in the previous version but is unified with a CUI in the current version then it can be found in a table named MERGEDCUI.RRF. This table has the following column structure.

PCUI	Concept Unique Identifier in the previous Metathesaurus
PSTR	Preferred name of this concept in the previous Metathesaurus

4.11.2MERGEDCUI.RRF

In a case where a particular concept is related to another concept in UMLS meta-thesaurus, and those two concepts, named CUI1 and CUI2, are combined together by UMLS, then one of them is withdrawn and only one remains in the meta-thesaurus. The complete merging history is available in this table and has the following column structure.

PCUI1 Concept Unique Identifier in the previous Metathesaurus

CUI Concept Unique Identifier in this Metathesaurus in format C#######

4.11.3DELETEDLUI.RRF

This table contains all the deleted Lexical Unique Identifiers also known as LUIs. The LUIs in meta-thesaurus which are deleted or removed from the data source vocabulary is stored in this table. Following is the column structure of this table.

PLUILexical Unique Identifier in the previous Metathesaurus

PSTRPreferred Name of Term in the previous Metathesaurus

4.11.4MERGEDLUI.RRF

LUIs which are related to another LUI are combined into a single LUI in meta-thesaurus. Those two merged LUIs may have a different meaning when they were not merged together but after merging, they represent the same semantic meaning.

4.11.5DELETEDSUI.RRF

All the string unique identifiers which are either removed from the meta-thesaurus or are deleted are stored in this table. Their history regarding their string names and other features are all available in this table.

4.11.6MRCUI.RRF

The history of the CUI appeared in the prior and current versions are available in this table in the form of one or more rows. The synonym mappings along with the mapping reasons are also available in this table.

4.11.7MRAUI.RRF

The change or modify history of AUI from both the previous and upcoming versions is stored in this table. Following column structure eis followed in this table.

AUI1	Atom unique identifier
CUI1	Concept unique identifier
VER	Version
REL	Relationship
RELA	Relationship attribute
MAPREASON	Reason for mapping
AUI2	Atom Unique identifier 2
CUI2	Concept Unique identifier 2
MAPIN	Mapping in current subset.

Table 4.16: MRAUI.RFF

4.11.8MRXW_ENG.RRF

There are ae indexes and strings available for every concept in a database table. These types of explanations are available in the table named MRXW_ENG.RRF. following are the columns in this table.

LAT	Abbreviation of language of the string
NWD	Normalized word
CUI	Concept identifier
LUI	Term identifier
SUI	String identifier

Table 4.17: MRXW_ENG.RFF

4.7.9 MRXNW_ENG.RRF

The words available in the normalized format are available in this table. As the scope of meta-thesaurus is very vast and it caters words and concepts from many languages, but normalized words are available for the English language only. It has following columns.

LAT	Abbreviation of language of the string
NWD	Normalized word
CUI	Concept identifier
LUI	Term identifier
SUI	String identifier

Table 4.18: MRXNW_ENG.RFF

4.11.9MRXNS_ENG.RRF

In the previous table, words in the normalized format were catered whereas in this table the normalized strings are available. Just like the previous table, this table also consists of normalized strings in the English language only. It has following columns.

LAT	Abbreviation of language)
NSTR	Normalized string)
CUI	Concept identifier

LUI	Term identifier
SUI	String identifier

Table 4.19: MRXNS_ENG.RFF

CHAPTER 5

PROPOSED ONTOLOGY

5. Proposed Ontology

This chapter discusses the implementation and development of our proposed ontology. The domain ontology can now be prepared as the prerequisite requirements, like data gathering, data extracting along with the loading transformation of data into the database, are fulfilled. This chapter will shed light on the techniques and software tools used in order to implement the ontology for the domain of COVID-19. Furthermore, a live portal is developed in order to make the COVID-19 ontology available and accessible to all the end users, medical researchers as well the healthcare professionals. All the queries applied for the retrieval of required and concerned data are also discussed in this chapter.

5.1Tools and Software Used

- ### MySQL

SQL is an acronym of Structured Query Language and is considered as the most widely used relational database management system. MySQL is mostly used in projects where database is required. It is free and open source; hence, it is used in both the larger and smaller scale projects. It is compatible with a lot of platforms as it is constantly maintained by Oracle as well as reliable and fast in terms of execution.

- ### PHP

PHP is a scripting language, and it stands for Hypertext Preprocessor. It is the most widely used scripting language for the development of web pages on the server. The developed websites and webpages are both interactive and dynamic. PHP is nowadays considered as the competitor of Microsoft ASP.net which is also used in the development of websites and other web services.

- ### APACHE Webserver

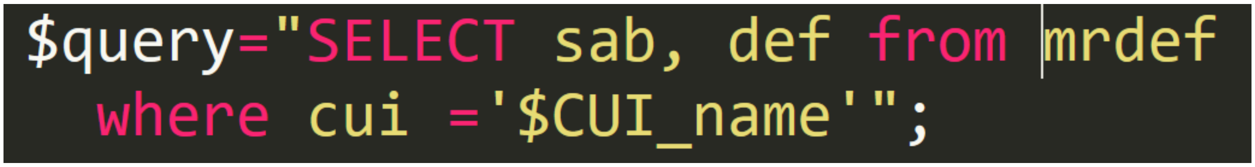
Apache webserver is a tool developed in order to make communication possible over World Wide Web. The HTTP requests are accepted from the users over internet and in turn the required data is transferred to the user in the form of webpages and other web files. It is an open source tool which is available for use without any cost.

- **Protégé**

It is an open source and free tool for building ontologies along with other applications. It is an editor who uses various languages and frameworks in order to develop and define ontologies. Protégé is supported and used by a group of academic or business consumers. It also consists of a large number of plugins and frameworks required for various purposes and building new applications and systems on the basis of ontology being one of them.

5.2SQL Queries used

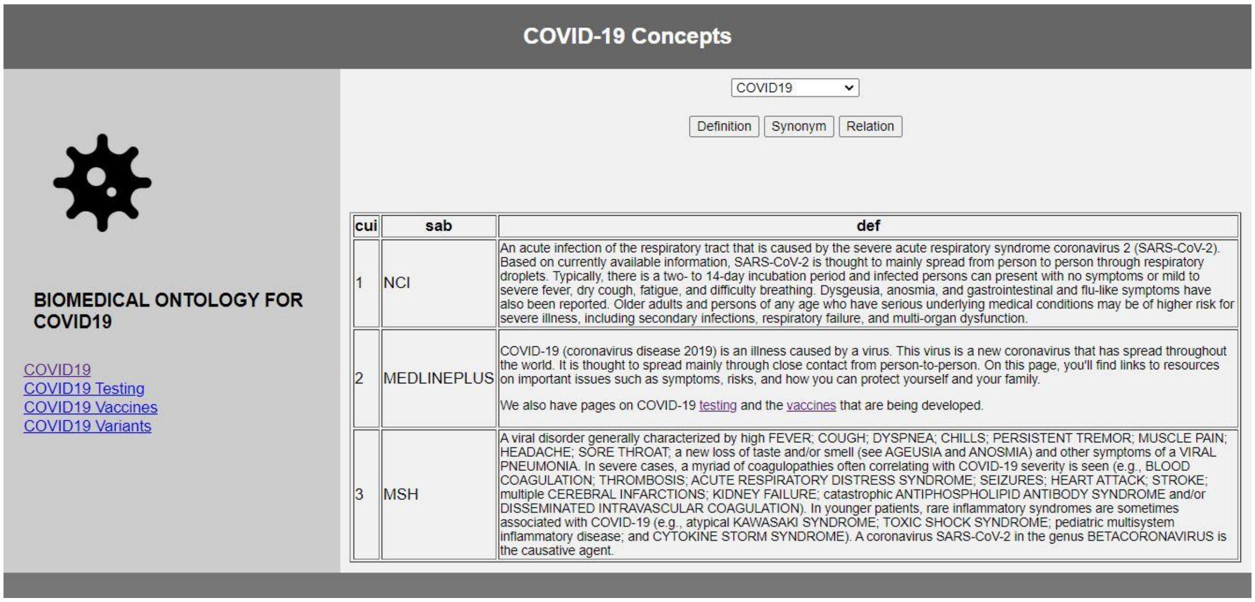
5.2.1 Display Definition



```
$query="SELECT sab, def from mrdef
  where cui ='$CUI_name'";
```

Figure 5.1: Query to display definitions of a concept

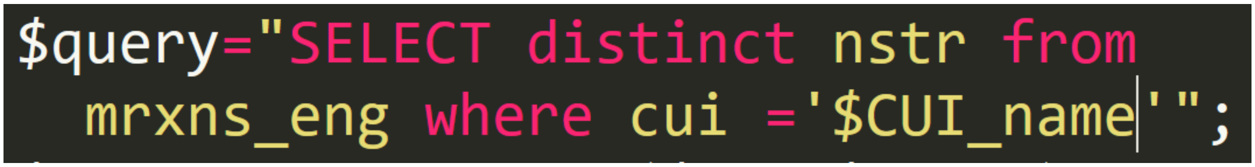
This above displayed query is used to display all the definitions related to COVID-19 from all the available source vocabularies present in the knowledge base.



cui	sab	def
1	NCI	An acute infection of the respiratory tract that is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Based on currently available information, SARS-CoV-2 is thought to mainly spread from person to person through respiratory droplets. Typically, there is a two- to 14-day incubation period and infected persons can present with no symptoms or mild to severe fever, dry cough, fatigue, and difficulty breathing. Dysgeusia, anosmia, and gastrointestinal and flu-like symptoms have also been reported. Older adults and persons of any age who have serious underlying medical conditions may be of higher risk for severe illness, including secondary infections, respiratory failure, and multi-organ dysfunction.
2	MEDLINEPLUS	COVID-19 (coronavirus disease 2019) is an illness caused by a virus. This virus is a new coronavirus that has spread throughout the world. It is thought to spread mainly through close contact from person-to-person. On this page, you'll find links to resources on important issues such as symptoms, risks, and how you can protect yourself and your family. We also have pages on COVID-19 testing and the vaccines that are being developed.
3	MSH	A viral disorder generally characterized by high FEVER; COUGH; DYSPNEA; CHILLS; PERSISTENT TREMOR; MUSCLE PAIN; HEADACHE; SORE THROAT; a new loss of taste and/or smell (see AGEUSIA and ANOSMIA) and other symptoms of a VIRAL PNEUMONIA. In severe cases, a myriad of coagulopathies often correlating with COVID-19 severity is seen (e.g., BLOOD COAGULATION; THROMBOSIS; ACUTE RESPIRATORY DISTRESS SYNDROME; SEIZURES; HEART ATTACK; STROKE; multiple CEREBRAL INFARCTIONS; KIDNEY FAILURE; catastrophic ANTIPHOSPHOLIPID ANTIBODY SYNDROME and/or DISSEMINATED INTRAVASCULAR COAGULATION). In younger patients, rare inflammatory syndromes are sometimes associated with COVID-19 (e.g., atypical KAWASAKI SYNDROME; TOXIC SHOCK SYNDROME; pediatric multisystem inflammatory disease; and CYTOKINE STORM SYNDROME). A coronavirus SARS-CoV-2 in the genus BETACORONAVIRUS is the causative agent.

Figure 5.2: COVID-19 Definitions

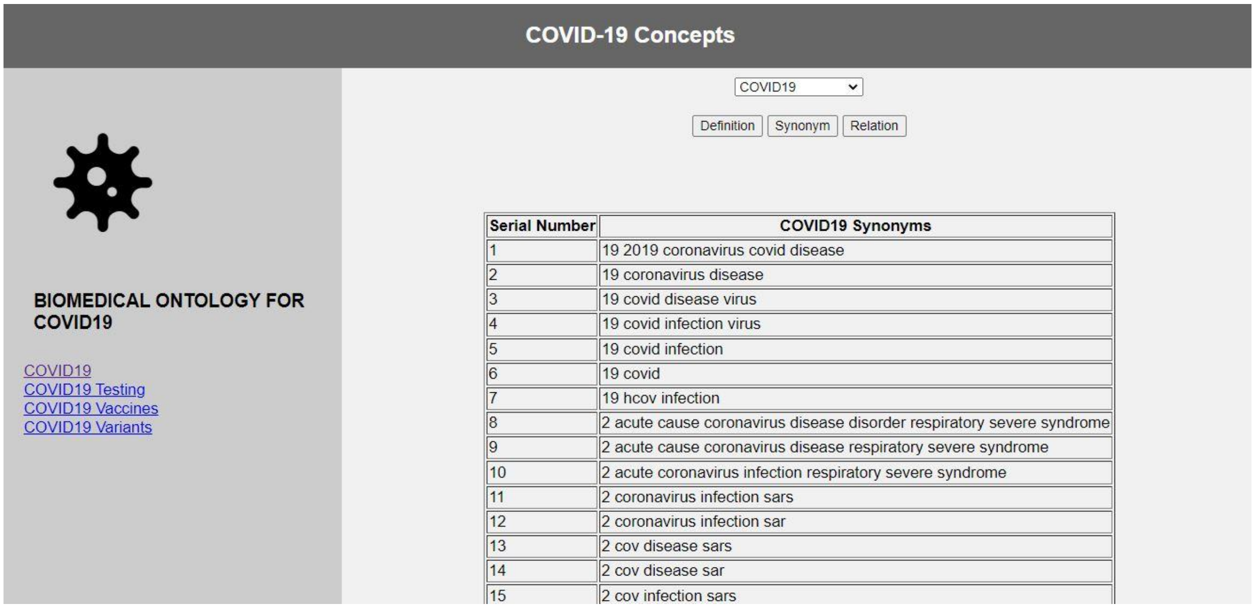
5.2.2 Display Synonym



```
$query="SELECT distinct nstr from
  mrxns_eng where cui ='$CUI_name'";
```

Figure 5.3: Query to display the synonyms of a concept

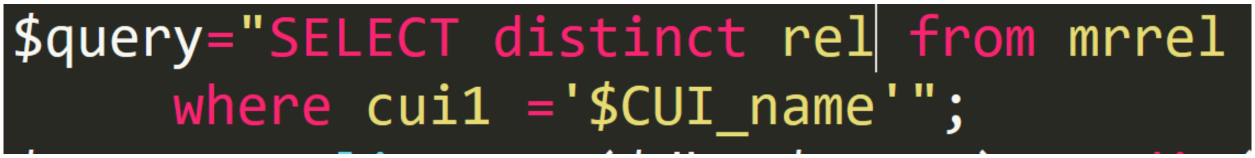
This query is used to display all the unique synonyms of a concept available in the knowledge base on the basis of CUI.



Serial Number	COVID19 Synonyms
1	19 2019 coronavirus covid disease
2	19 coronavirus disease
3	19 covid disease virus
4	19 covid infection virus
5	19 covid infection
6	19 covid
7	19 hcov infection
8	2 acute cause coronavirus disease disorder respiratory severe syndrome
9	2 acute cause coronavirus disease respiratory severe syndrome
10	2 acute coronavirus infection respiratory severe syndrome
11	2 coronavirus infection sars
12	2 coronavirus infection sar
13	2 cov disease sars
14	2 cov disease sar
15	2 cov infection sars

Figure 5.4: COVID-19 Synonyms

5.2.3 Display Broad Relations

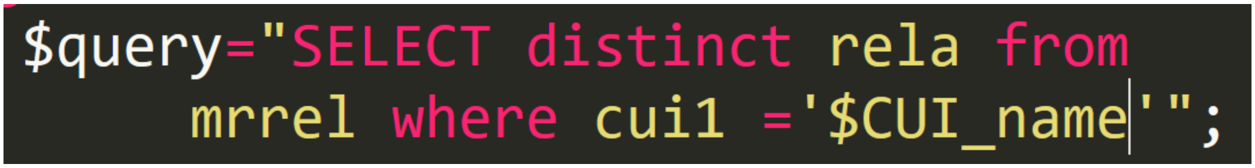


```
$query="SELECT distinct rel from mrrel
     where cui1 ='$CUI_name'";
```

Figure 5.5: Query to display broader relations

This query is used to display all the unique relations a CUI has with any other CUI. Upon selection of a particular CUI by the end user on a live portal, all the relations of that CUI with other CUIs will be displayed.

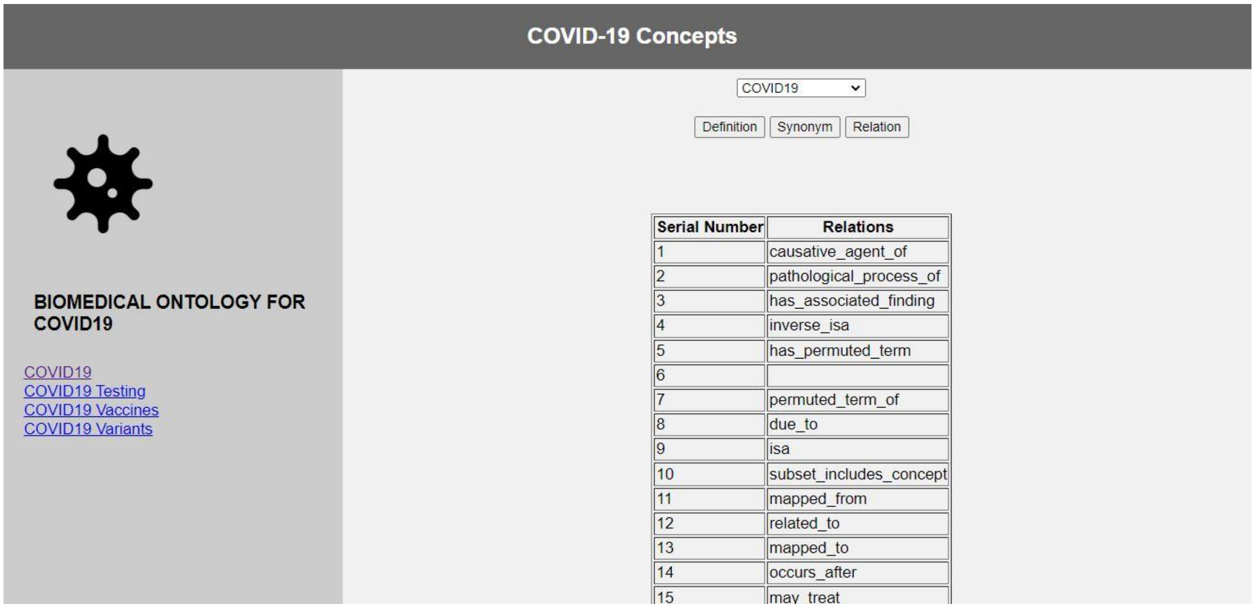
5.2.4 Display Narrow Relations



```
$query="SELECT distinct rela from
     mrrel where cui1 ='$CUI_name'";
```

Figure 5.6: Query to display narrow relations

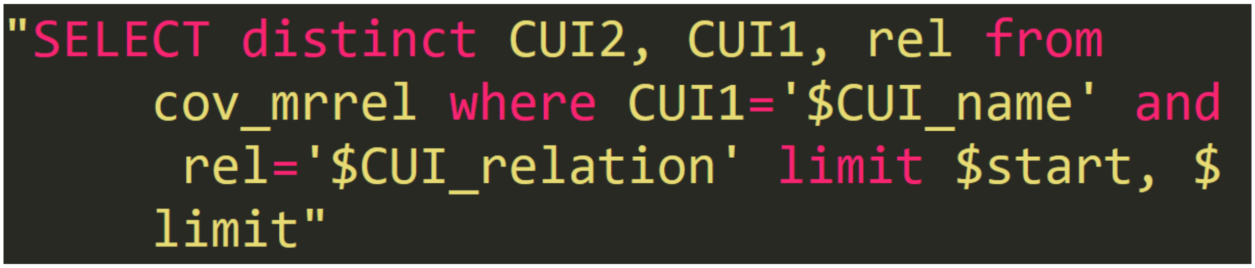
This query displays all the unique narrow relations a CUI has with other CUIs.



Serial Number	Relations
1	causative_agent_of
2	pathological_process_of
3	has_associated_finding
4	inverse_isa
5	has_permuted_term
6	
7	permuted_term_of
8	due_to
9	isa
10	subset_includes_concept
11	mapped_from
12	related_to
13	mapped_to
14	occurs_after
15	may_treat

Figure 5.7: COVID-19 Narrow Relations

5.2.5 Display Relation Details

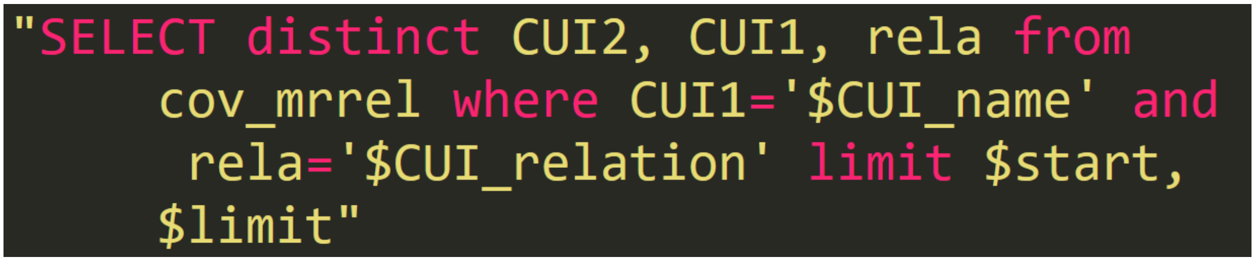


```
"SELECT distinct CUI2, CUI1, rel from
     cov_mrrel where CUI1='$CUI_name' and
      rel='$CUI_relation' limit $start, $
     limit"
```

Figure 5.8: Query to display relation details

The above query is used to display the unique CUIs related with the selected CUI. A display record limit has been imposed hence it displays only 10 records on a single page.

5.2.6 Display Narrow Relation Details



```
"SELECT distinct CUI2, CUI1, rela from
     cov_mrrel where CUI1='$CUI_name' and
      rela='$CUI_relation' limit $start,
     $limit"
```

Figure 5.9 Query to display narrow relation details

The above query displays the unique CUIs associated with the selected CUI. Here again a limit is imposed and only 10 records a displayed at a time.

CHAPTER 6

RESULTS AND VALIDATION

6. Results and Validation

This chapter discusses all the results obtained by the aforementioned methodology. This chapter covers all the details regarding the resulting domain ontology including all the concepts and its attributes. All the concepts related to COVID-19, which we have covered in the ontology along with their relations will be discussed in this chapter. Furthermore, this chapter also includes the validation methodology used for the COVID-19 ontology. The results obtained from the validation part are also discussed in this chapter.

6.1Overview

After creating our ontology, all the data is available in the live portal which is created in order to facilitate the end users along with the researchers who are working in the domain of COVID-19 as well as the clinical scholars. The link for this live portal is here.

The live portal caters all the coverage of concepts, sub-classes, relations both narrow and broad etc. In other words, the complete knowledge for the domain of COVID-19 is present in the live portal.

6.2COVID-19 Categories

The novel coronavirus also called COVID-19 is a viral infection which took the whole world by storm. As there are research being published on the regular basis, the data about COVID-19 is increasing day by day. On the basis of the data and information available for the domain of COVID-19, there are three major concepts related to COVID-19 which can be catered. Those three major concepts are then further classified into sub-concepts. The relations of these concepts among other concepts are also covered by the domain ontology. These concepts along with their CUIs and visual representation are discussed as under.

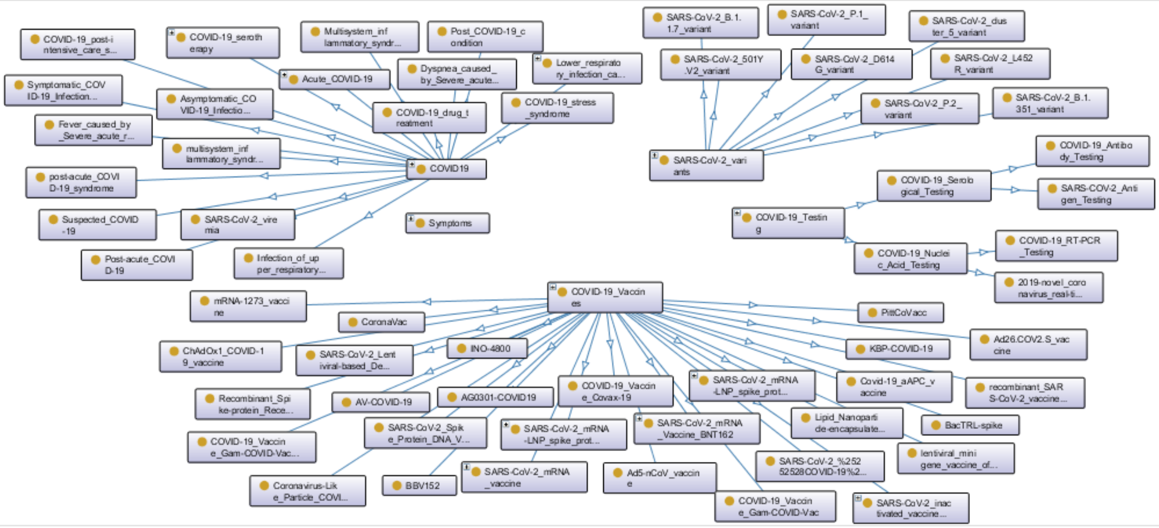

COVID-19_post-intensive_care_s...
COVID-19_serotherapy
Multisystem_inflammatory_syndr...
Post_COVID-19_condition
SARS-CoV-2_B.1.1.7_variant
SARS-CoV-2_P.1_variant
SARS-CoV-2_duster_5_variant
Symptomatic_COVID-19_Infection...
Acute_COVID-19
Dyspnea_caused_by_Severe_acute...
Lower_respiratory_infection_ca...
SARS-CoV-2_501Y.V2_variant
SARS-CoV-2_D614G_variant
SARS-CoV-2_L452R_variant
Asymptomatic_COVID-19_Infectio...
Fever_caused_by_Severe_acute_r...
COVID-19_drug_treatment
COVID-19_stress_syndrome
SARS-CoV-2_P.2_variant
SARS-CoV-2_B.1.351_variant
multisystem_inflammatory_syndr...
COVID-19_Antibody_Testing
post-acute_COVID-19_syndrome
COVID19
SARS-CoV-2_variants
COVID-19_Serological_Testing
SARS-COV-2_Antigen_Testing
Suspected_COVID-19
SARS-CoV-2_viremia
Symptoms
COVID-19_Testing
COVID-19_RT-PCR_Testing
Post-acute_COVID-19
Infection_of_upper_respiratory...
COVID-19_Nucleic_Acid_Testing
2019-novel_coronavirus_real-ti...
mRNA-1273_vaccine
COVID-19_Vaccines
CoronaVac
PittCoVacc
ChAdOx1_COVID-19_vaccine
SARS-CoV-2_Lentiviral-based_De...
INO-4800
KBP-COVID-19
Ad26.COV2.S_vaccine
Recombinant_Spike-protein_Rece...
AV-COVID-19
AG0301-COVID19
COVID-19_Vaccine_Covax-19
SARS-CoV-2_mRNA-LNP_spike_prot...
Covid-19_aAPC_vaccine
recombinant_SARS-CoV-2_vaccine...
SARS-CoV-2_Spike_Protein_DNA_V...
SARS-CoV-2_mRNA-LNP_spike_prot...
SARS-CoV-2_mRNA_Vaccine_BNT162
Lipid_Nanoparticle-encapsulate...
BacTRL-spike
COVID-19_Vaccine_Gam-COVID-Vac...
lentiviral_minigene_vaccine_of...
SARS-CoV-2_%25252528COVID-19%2...
Coronavirus-Like_Particle_COVI...
BBV152
SARS-CoV-2_mRNA_vaccine
Ad5-nCoV_vaccine
COVID-19_Vaccine_Gam-COVID-Vac
SARS-CoV-2_inactivated_vaccine...

Figure 6.1: COVID-19 Ontology

COVID-19 main concepts

Serial number	Name	Concept Unique Identifier
1	COVID-19	C5203670
2	COVID-19 Variants	C5433390
3	COVID-19 Vaccines	C5387588
4	COVID-19 Testing	C5244026

Table 6.1: COVID-19 Concepts

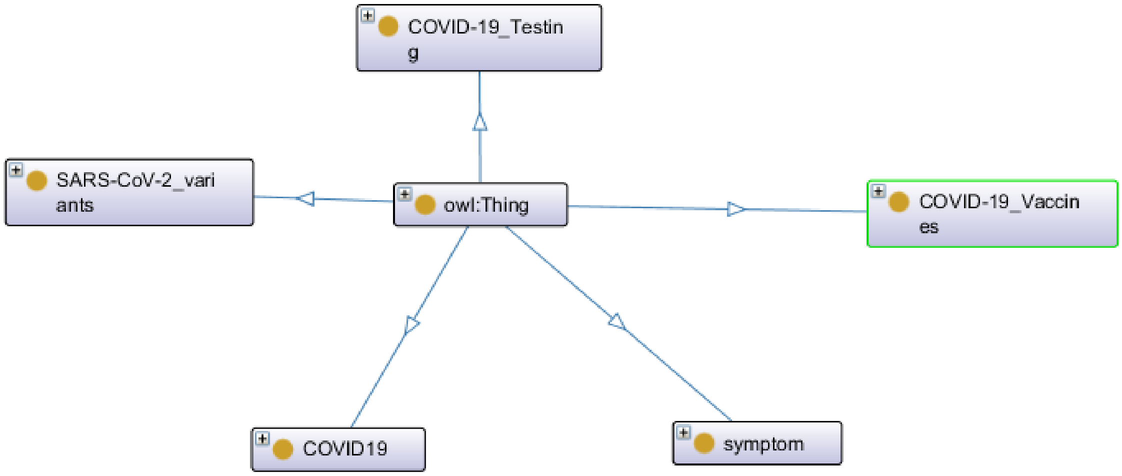


Figure 6.2: COVID-19 Concepts

COVID-19 Variants

Serial number	Name	Concept Unique Identifier
1	A.1.177	C5543782
2	A.23.1	C5543790
3	B.1.1.7	C5433393
4	B.1.351	C5433395
5	B.1.427	C5543783
6	B.1.525	C5543789
7	B.1.526	C5543787
8	B.1.617.1	C5543788
9	B.1.617.2	C5543784
10	cluster 5	C5433397
11	D614G	C5433391
12	E484K	C5543785
13	L452R	C5433396
14	P.1	C5433398

15	P.2	C5433392
16	P.3	C5543786

Table 6.2: COVID-19 Variants

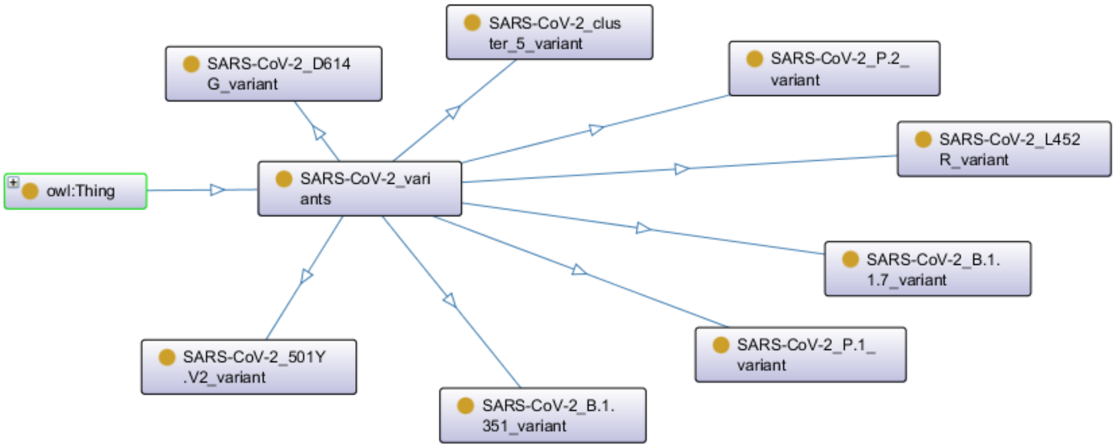


Figure 6.3: COVID-19 Variants

COVID-19 Testing

COVID-19 Testing concept is divided into two main parts, which are then further divided into sub-parts. They are mentioned below.

Serial number	Name	Concept Unique Identifier
1	Nucleic Acid Testing	C5392163
2	Serological Testing	C5244032

Table 6.3 COVID-19 Testing

Nucleic Acid Testing

Serial number	Name	Concept Unique Identifier
1	RT-PCR testing	C5392164

Table 6.4 COVID-19 Nucleic Acid Testing

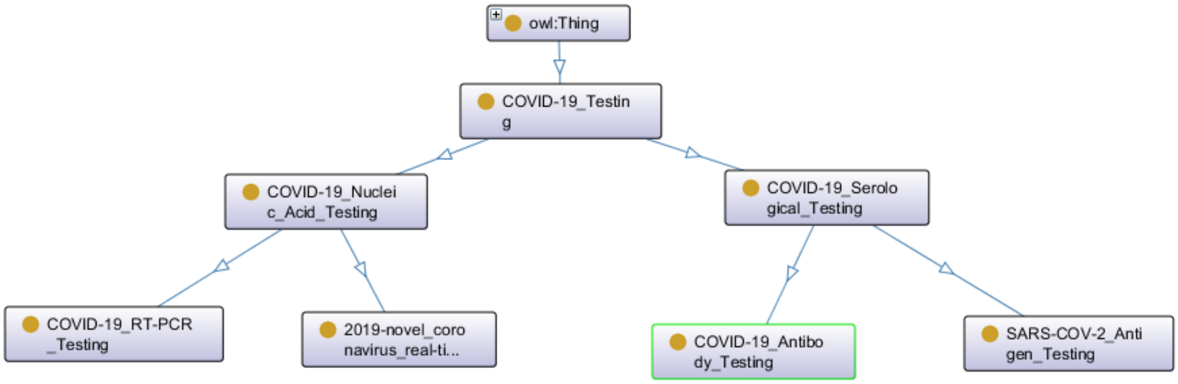

owl:Thing
COVID-19_Testing
COVID-19_Nucleic_Acid_Testing
COVID-19_Serological_Testing
COVID-19_RT-PCR_Testing
2019-novel_coronavirus_real-ti...
COVID-19_Antibody_Testing
SARS-COV-2_Antigen_Testing

Figure 6.4: COVID-19 Testing

Serological Testing

Serial number	Name	Concept Unique Identifier
1	Antibody Testing	C5244064
2	Antigen testing	C5392882

Table 6.5 COVID-19 Serological Testing

COVID-19 Vaccines

Serial number	Name	Concept Unique Identifier
1	nCoV Vaccine mRNA-1273	C5244457
2	Ad5-nCoV	C5419987
3	AG0301-COVID19	C5416832
4	AV-COVID-19	C5400891
5	BacTRL-spike	C5416823
6	BBV152	C5416846
7	ChAdOx1 nCoV-19	C5383287

8	CoronaVac	C5416819
9	Coronavirus-Like Particle	C5416845
10	aAPC	C5419986
11	Covax-19	C5420359
12	Gam-COVID-Vac	C5400622
13	Gam-COVID-Vac Lyo	C5400623
14	INO-4800	C5435003
15	KBP-COVID-19	C5416834
16	lentiviral minigene vaccine	C5433506
17	PittCoVacc	C5433507
18	NVX-cov2373	C5416844
19	Recombinant Spike-protein Receptor-binding Domain-dimer	C5420411
20	SARS-CoV-2 (COVID-19) vaccine, mRNA-BNT162b2	C5398043
21	Lentiviral-based Dendritic Cell Vaccine LV-SMENP-DC	C5417936
22	Ad26	C5416833

Table 6.6 COVID-19 Vaccines

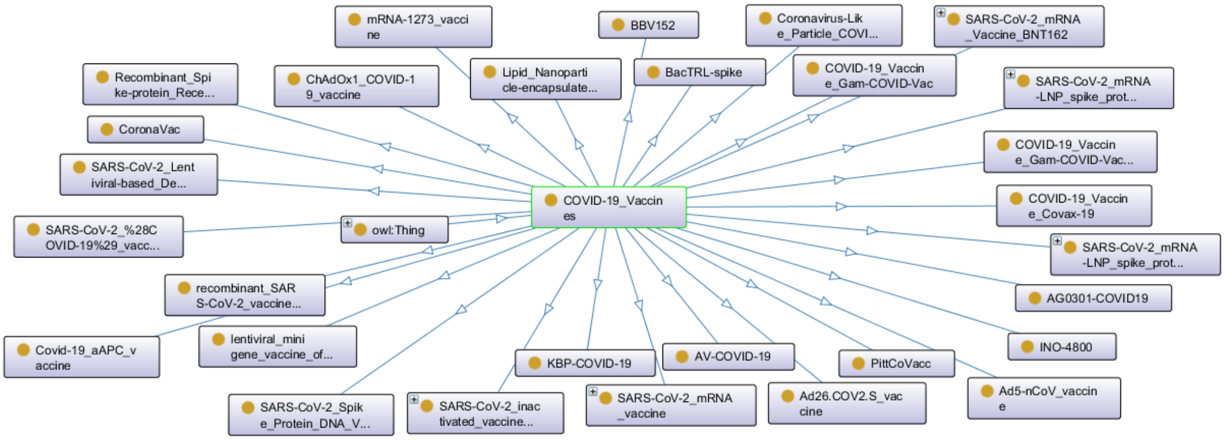

mRNA-1273_vaccine
BBV152
Coronavirus-Like_Particle_COVI...
SARS-CoV-2_mRNA_Vaccine_BNT162
Recombinant_Spike-protein_Rece...
ChAdOx1_COVID-19_vaccine
Lipid_Nanoparticle-encapsulate...
BacTRL-spike
COVID-19_Vaccine_Gam-COVID-Vac
SARS-CoV-2_mRNA-LNP_spike_prot...
CoronaVac
SARS-CoV-2_Lentiviral-based_De...
COVID-19_Vaccine_Gam-COVID-Vac...
COVID-19_Vaccines
SARS-CoV-2_%28COVID-19%29_vacc...
owl:Thing
COVID-19_Vaccine_Covax-19
SARS-CoV-2_mRNA-LNP_spike_prot...
recombinant_SARS-CoV-2_vaccine...
AG0301-COVID19
lentiviral_minigene_vaccine_of...
Covid-19_aAPC_vaccine
INO-4800
KBP-COVID-19
AV-COVID-19
PittCoVacc
Ad5-nCoV_vaccine
SARS-CoV-2_Spike_Protein_DNA_V...
SARS-CoV-2_inactivated_vaccine...
SARS-CoV-2_mRNA_vaccine
Ad26.COV2.S_vaccine

Figure 6.5: COVID-19 Vaccines

6.3Relation Count

Sr.No	Name	CUI	Number of Broader relations	Number of Narrow relations
1	COVID-19	C5203670	8	22
2	COVID-19 Testing	C5244026	3	2
3	COVID-19 Vaccine	C5387588	5	48
4	COVID-19 Variants	C5433390	1	16
5	Nucleic Acid Testing	C5392163	1	2
6	Serological Testing	C5244032	1	2
7	COVID-19 Serotherapy	C5391448	2	3
8	mRNA-BNT162	C5417954	3	5
9	Inactivated vaccines	C5416836	3	3
10	mRNA-1273	C5244457	3	1
11	Coronavac	C5416819	3	1

12	NVX cov-2373	C5416844	3	1
13	COVID-19 RNA	C5380594	1	20

Table 6.7 Relation Count

6.4Comparison to major existing knowledge bases

There are a large number of biomedical ontologies being present at the moment. Each ontology tackles the particular domain and almost all of them lacks the unified information in terms of semantics of a particular domain ontology. In other words, it can be said that the data related to a particular domain is not depicted in a single platform keeping in check the semantics related to that domain. In order to understand all the different aspects of a disease, then the user must access multiple knowledge bases instead of a single source and such a type of knowledge base does not exist. Following table shows the comparison of our ontology with the previous knowledge bases.

Existing Knowledge bases	Our ontology
The current ontologies consist of particulars of data labelling and are deficient in having the crucial semantic type associations for examples MeSH, UMLS, YAGO etc	This ontology implies information on the semantic relations (holds to, adjoining to, type of, linked with, cause of. Etc.)
Current knowledge bases comprise of the features of specific contexts relations with genotype phenotype relation, gene, drug and disease relations for instance, OMIM, GAD and Diseasome.	This Ontology involves unified data aiming for the domain of COVID-19

Table 6.8 Comparison with other ontologies

As we are well aware that in the field of medicine and bioinformatics, NLM or National Language of Medicine is considered as the most authentic and complete data source. It consists of a large number of most reliable and frequently used knowledge bases. UMLS being part of NLM is also considered a reliable source as it is updated on a regular basis and is authentic and verified thoroughly. Out data source is UMLS, which is a reliable data source, hence it makes the created ontology to be authentic and reliable.

The main differences and improvements among our ontology and previously available biomedical ontologies is mentioned in the table below.

KB	Domain	Source	Size	Semantic Relations	Causative factors of diseases
Proposed Ontology	Corona Virus/ COVID-19 disease, Definitions, Relations, Synonyms	UMLS	Different variants of COVID-19 and around 500 related concepts	Included	Included
MeSH	General medical subjects for indexing articles for PubMed database	PubMed publications	~25,186 entities	Not included	Included as a separate entry term
FMA	Human anatomy	UMLS	120,000 concepts	Included	Not 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
Diseasome	Human disease network	OMIM	>4,213 diseases, >91,182 genes	Contains relations of specialized context	Not included

Table 6.9 Detailed comparison with other Ontologies

6.5Ontology Application Research Domains

The use of ontologies in terms of using and displaying knowledge has increased over the past few years and many researchers are working to implement the domain specific ontologies. Keeping this in mind, the use of ontology in the field of biomedicine is also increasing and many biomedical ontologies are developed for the better interpretation of field knowledge as well.

The most major issue in this regard is the unstructured data available on web and other platforms in the form of literature or other sources[1]. This type of data which is dispersed and blended with a lot of noise creates a lot of inconvenience for both the researchers and the patients and cannot be used for any purpose effectively, most importantly in the field of medicine and health care. For instance, if a person decides to perform a basic search on web regarding a particular disease and intends to understand the symptoms and causes associated with that disease, the resulting material would be related to the research articles and other literature. The literature related to any disease is in such a bulk that it would require around a month to select and understand the required results and information. Therefore, if a knowledge base is prepared which is understandable by machines and all the related knowledge i.e symptoms, terms, causative factors etc, about the disease is stored in it, then many people can benefit from it.

6.5.1 Application areas

The major application domains of this developed ontology are mentioned as under:

- A complete, extensive and machine understandable vocabulary that can produce the required results on the basis of queries like a database and is based on semantic meanings.
- The knowledge available in our ontology can be used for the aim of record linkage or better known as entity resolution [43] as well as for data cleansing [82].
- Can be used in machine learning project for example [45], as it is a complete and extensive knowledge base.
- Question answering in Natural Language Processing [18], [41] can also be catered via this knowledge base.

6.6 Ontology Validation

When it comes to ontologies related to the field of biomedicine, a significant level of research and effort is required in order to build and implement it. In other words, the development of applications related to medical sciences is always crucial as human lives are dependent on it, therefore, the ontology development related to the field of biomedicine must be carried out rigorously and effectively making sure no loopholes are left. This process is crucial and complex and must be supervised closely by the medical professionals as well as professionals from the ICT domain. As we are well aware that the tools related to ICT domain and automated mechanisms to create ontologies automatically are quite efficient and time saving, but they are limited to a certain extent and does not produce high quality results. Therefore, when developing such tools and applications, the system and complete development process must be supervised by humans from the experts belonging to the ICT domain, resulting in highly efficient and quality applications. In other words, the developed ontologies must be validated by the domain experts and all the deficiencies must be rectified.

There is a huge amount of data available all over the web as well as other literary source libraries. This data is so vast and extensive that it is nearly impossible to extract the required data as it is usually and mainly available in unstructured format. Therefore, all the biomedical data available in the research articles and other sources is not useful for end users as it requires time and effort on their part.

This interpretation led to the understanding that the use of ontologies in the field of biomedicine and even other domains is increasing fast as it provides the knowledge and information related to a particular domain in a structured and comprehendible format. As the data related to biomedical ontologies must be correct and reliable therefore, the validation from the domain experts is a very necessary and crucial task. The domain experts including both medical and health professionals and ontology developers must coordinate and collaborate together in order to implement the ontology effectively and correctly. For the domain of COVID-19, the panel of doctors were contacted as the domain experts.

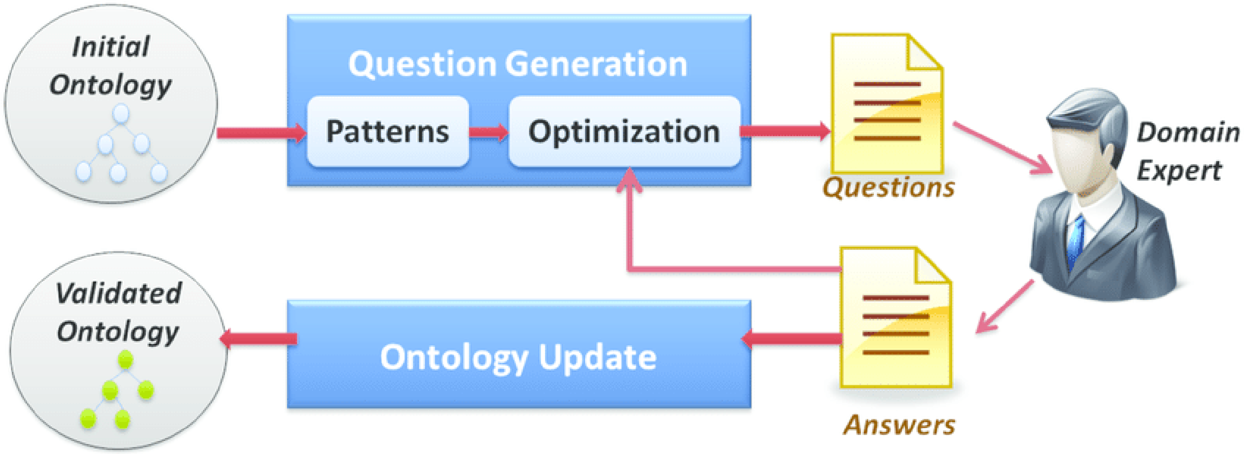


Figure 6.6: Process of Ontology Validation

The software engineers are not familiar with the domain knowledge of a particular disease therefore, they require the assistance of health care professionals. Similarly, the health care professionals are not familiar with the development techniques hence, they require the assistance of ICT experts while evaluating the ontology. After the complete development of the live portal, which is available on the following link, the ontology validation is an essential requirement.

In order to determine the correctness of ontology in all the fields, the validation is carried out in terms of data, structure and textual details available. The relationships existing in the ontology along with all the data available in the ontology is validated by the domain experts through a common technique termed as “laddering technique” [42]. The panel comprising 10 medical and health care professionals from two hospitals MH and PIMS were requested in order to validate the developed ontology. As the COVID-19 patients were catered in both the hospitals on the emergency basis and these hospitals were declared as the COIVD-19 hospitals therefore, both hospitals have seen the COVID-19 patients hence making it the reason behind the selection of these hospitals for the domain experts.

The process of validation was carried out by circulating the questionnaire regarding the details of ontology among the domain experts. The questionnaire consisted of 13 questions related to the coverage of relevant concepts, along with the relationships among these concepts. The contents of the ontology were discussed with the domain experts and after a thorough analysis it was deduced that the addition of more visual and textual data related to COVID-19 can be more effective for both the end users and clinical researchers. For this purpose, the data from prominent biomedical websites have been extracted and added into the ontology in order to facilitate the end users.

6.7 Result findings

A questionnaire consisting of 13 questions was circulated among the domain experts and after a thorough session of discussing all the visual and textual aid provided to them along with the live portal, they recorded their answers. The result findings from their evaluation are discussed as under.

6.7.1 Domain Effectiveness

The experts were asked to validate the effectiveness of ontology developed according to its specified domain and the responses are shown in the figure below. The purpose of this question is to gauge the usefulness of this domain ontology as to whether the implemented ontology provides the purpose it is created for. As a result, 75% of the domain experts agreed that the domain is useful, and the ontology provides the information it is developed for.

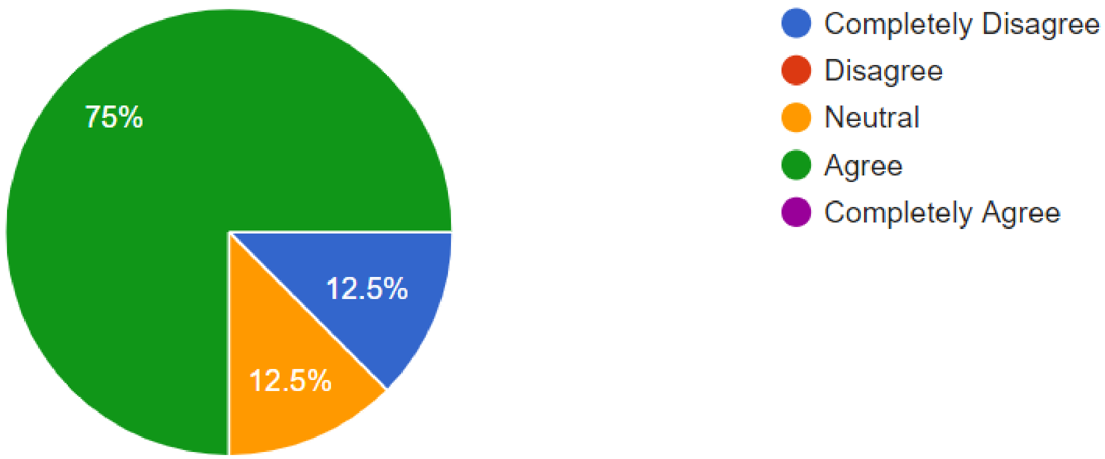


Figure 6.7: Domain Effectiveness

6.7.2 Concept Coverage

The experts were then asked to validate the number of concepts defined in the ontology. In other words, they were asked if the number of concepts were enough or not. Figure below shows the respective response. The aim regarding this question is to authenticate whether the concepts defined in the developed domain ontology enough to describe and understand the ontology.

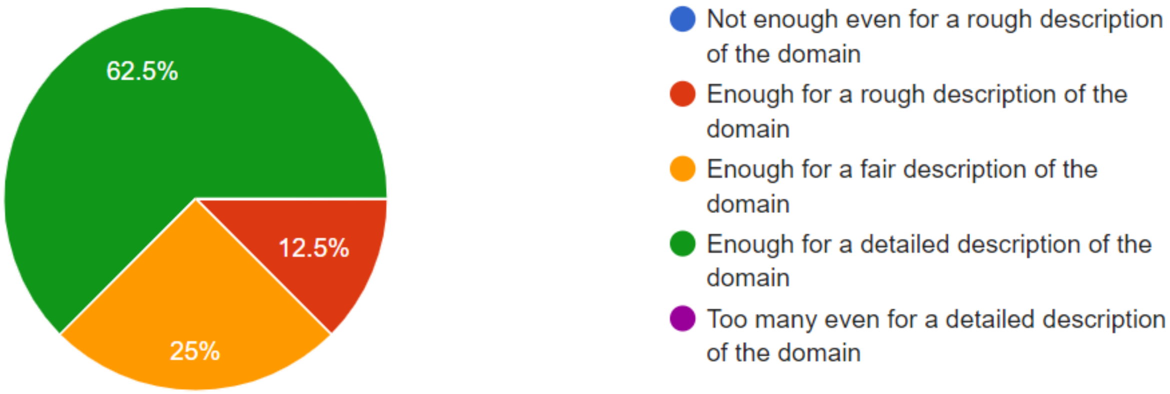


Figure 6.8: Concept Coverage

6.7.3 Concept Quality

The next question discusses the quality of concepts defined in the ontology and the responses are displayed in the following figure. This question focuses on the fact that the concepts defined are the ones that are required or not. In other words, the defined concepts are the ones that are important without which the ontology cannot be effective or not.

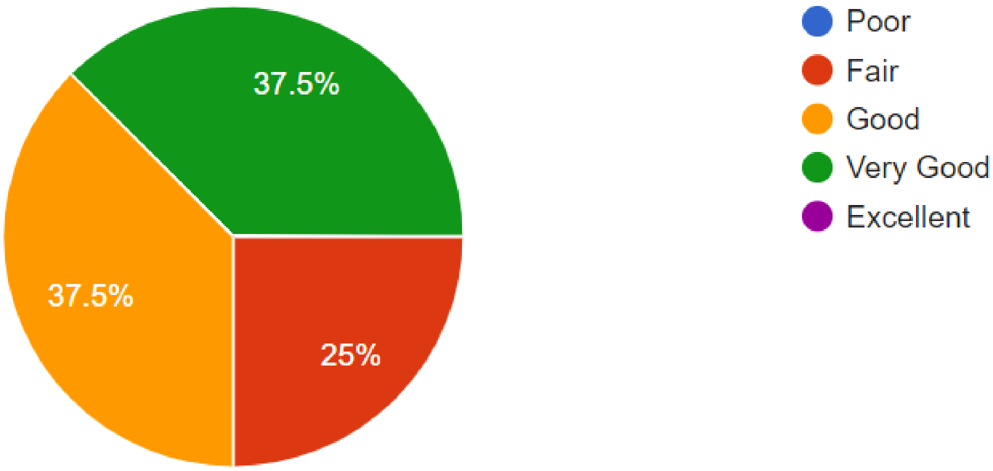


Figure 6.9: Concept Quality

6.7.4 Relation Coverage

The coverage of broader and narrower relations was validated in the next step. The answers from the domain experts are shown in the figure below. The narrow relations are the sub-relations among the concepts in an ontology whereas the broad relations are the major relations among the concepts. The question aims to validate the fact that the concepts, either broader or narrower, are covered effectively and completely and no concept relation is missing from the defined concepts. The results shows that 50% of domain experts marked the narrow relations to be covered so that it is understandable to the end users in detail. Whereas in case of broader relations, this percentage rose to around 57% claiming it to be covered for a detailed description of the specified domain.

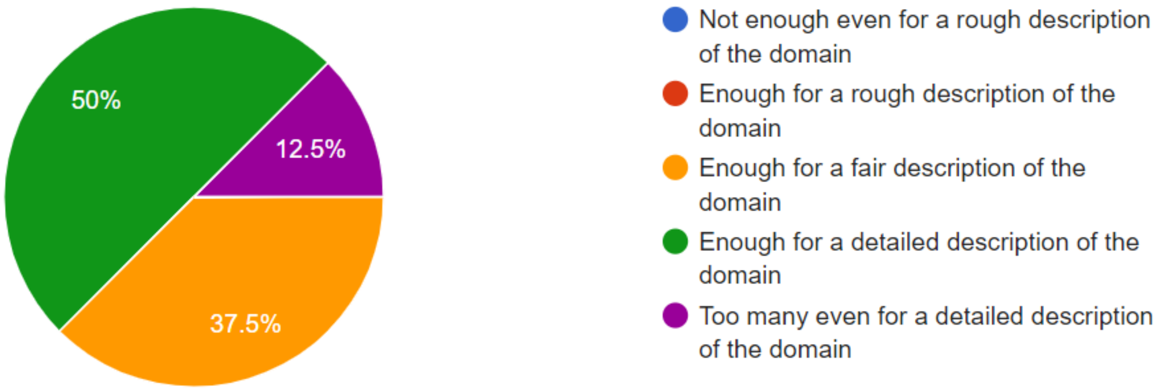


Figure 6.10: Relation Coverage (Broad)

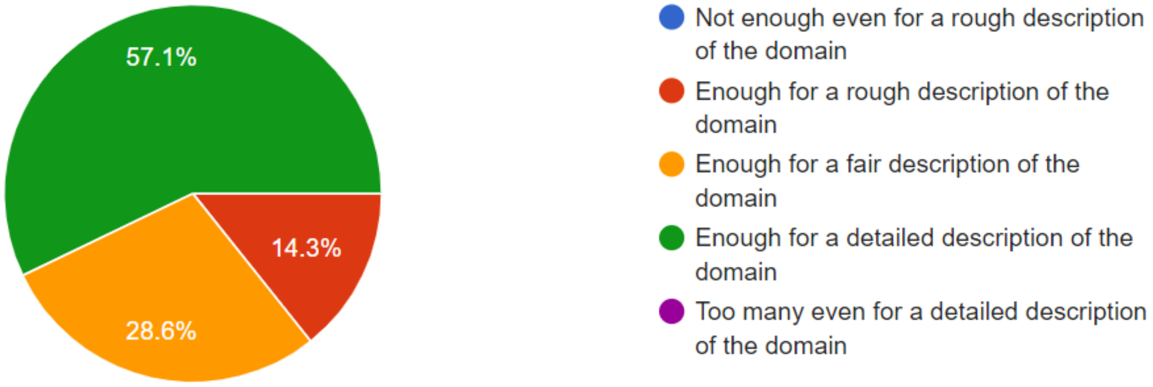


Figure 6.11: Relation Coverage (Narrow)

6.7.5 Definition Coverage

The experts were then asked to validate the definition coverage in the ontology. Following figure displays their responses. The focus of this question is to make sure the concepts defined in the ontology displays the definitions from different vocabularies. In other words, the definitions of defined concepts are covered in the ontology or not. The results shows that around 71% of domain experts agreed on the fact that the definitions of the concepts were covered effectively.

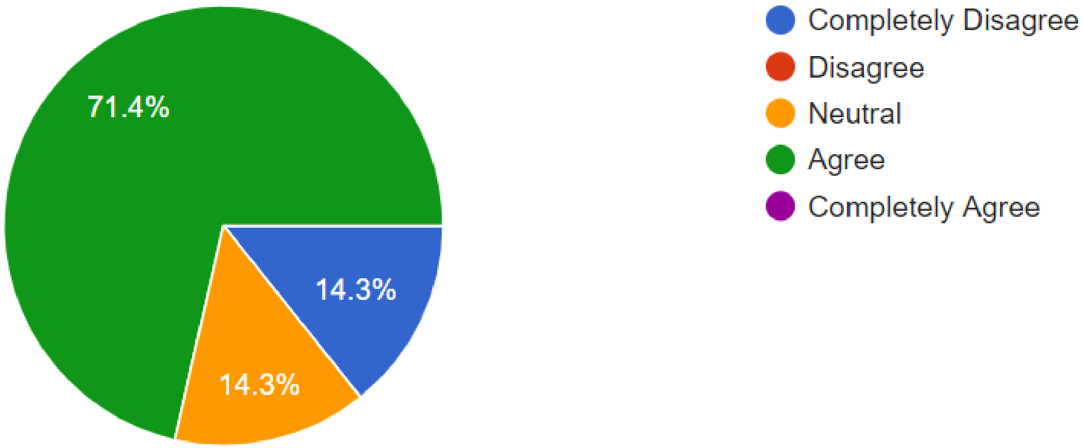


Figure 6.12: Definition Coverage

6.7.6 Synonym Coverage

Furthermore, the validation of synonym coverage is also included in ontology validation. The results are displayed in the following figure. The purpose of this question is to make sure that the concepts defined in the ontology displays all possible synonyms. Around 71% of domain experts validated that the concepts defined in the domain ontology cover all possible synonyms effectively.

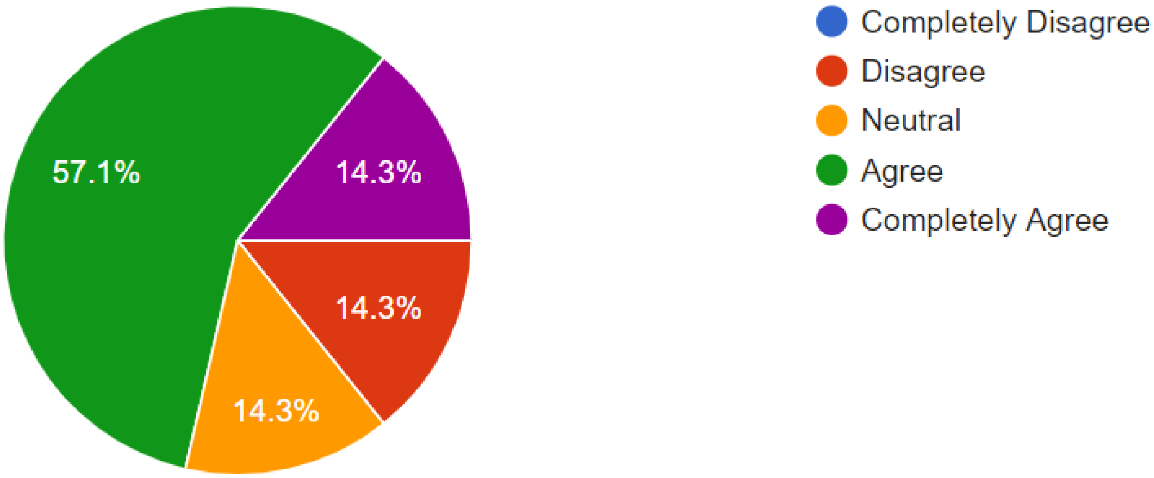


Figure 6.13: Synonym Coverage

6.7.7 Visual and Textual Details Coverage

In the end, the overall response is also collected from the domain experts in the form of text. The ontology live portal itself is evaluated in terms of its usefulness and quality. The domain experts were asked if they had all the required textual and visual details in order to understand the structure of ontology and validate it. The domain experts validated that in around 70 % of the cases, the textual and visual details were more than enough for describing the domain ontology in detail.

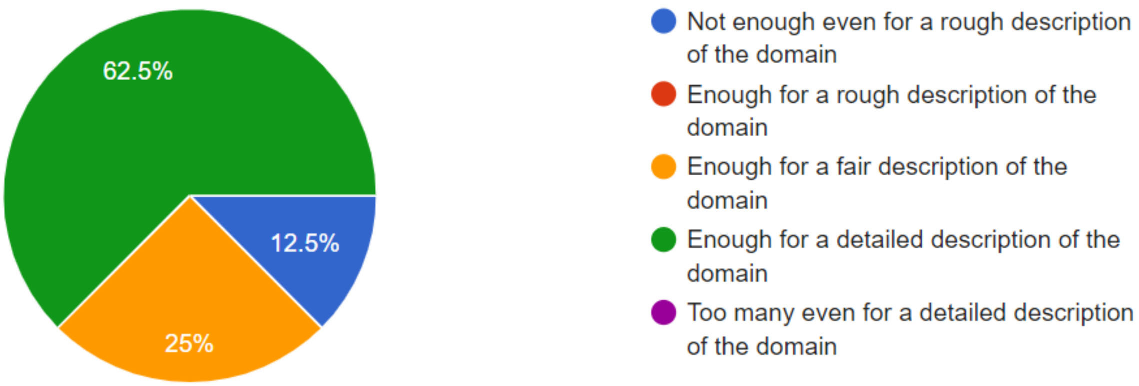


Figure 6.14: Visual Details Coverage

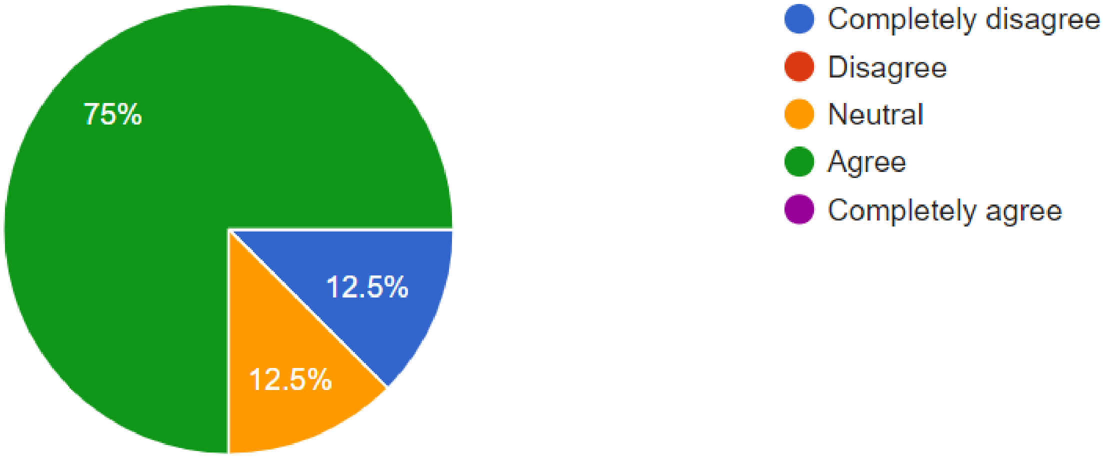


Figure 6.15: Textual Details Coverage

In order to improve the developed ontology, a comprehensive discussion was followed after the ontology validation with the domain experts. The domain experts pointed out a few suggestions in order to improve the live portal of generated ontology. They concluded that after the addition of a few visual and textual details related to COVID-19 on the live portal for the facilitation of the end users, the ontology is complete and provides the complete knowledge of the domain of COVID-19.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7 Conclusion and Future Work

7.1 Overview

A knowledge represented or shared in an impractical manner is of no use. In today's world where data and information are overflowing and are increasing on the daily basis, the knowledge bases which represents data in a structured and well-organized format are crucial and of most importance as it makes the task of data accessibility and it's processing much easier. The most common form of finding the data or information related to a biomedical domain is in the research articles or scientific studies and that too is available in an unorganized and unstructured format over the web. The availability of data is in bulk and massive, but it becomes a challenge in order to extract the required information from that bulk. Therefore, a machine understandable knowledge base is required that can store all the semantics and relationships regarding a particular domain.

When defining the term semantic web, many a times its original meaning is not interpreted. It can be defined as the network of data and knowledge woven in such a way that it is interpretable by machines universally [9], [10]. Another way to define the semantic web is that a technique used to implement the vocabularies in order to define the knowledge of a domain in a machine interpretable form [11]. As each term can have more than one meaning therefore adding semantics to the world wide web can be challenging task.

This is where ontologies come into view. This is the reason why we have created an ontology. The data is collected and turned into an ontology which caters all the basic terms, concepts, relationships among those terms, their definitions along with the semantic types for the domain of COVID-19 also known as Corona Virus. To implement an ontology for the domain of COVID-19 which includes all the data and knowledge related to the definitions, synonyms, semantic types and relations is the main goal of this study. The data related to biomedical sciences for this research is extracted form UMLS also known as Unified Medical Language System, supervised and controlled by NLM or National Library of Medicine. It contains around one hundred and seventy biomedical dictionaries.

The live portal is developed in order to display the concepts available in the domain ontology for the easy accessibility of health care professionals, researchers and end users. (Add link here).

7.2 Application areas of our ontology

The role of biomedical ontologies in the domain of biomedicine and bio informatics is very important and crucial for the scientific researchers. In the beginning the role of ontologies was only limited to the purpose of formatting knowledge in a structured way, as well the standardization of data but as the research in this field arose and the ontologies got more exposure, they are now used for extraction of data, semantic web, decision support systems, creating domain specific applications which can be reused. Furthermore, the use of ontologies in many CDSS or Clinical Decision Support Systems as the foundation is also trending. The ontologies are also used in order to increase the efficiency of CDSS, and the results obtained using various methodologies as they control and tolerate the data in Clinical Decision Support Systems.

7.3 Conclusion and Recommendations

The main aim of this research thesis is to define and explain the methodology used in order to develop an ontology from the huge dataset from UMLS resources as well as the mechanism to tailor the webapp in order to display the complete and comprehensive data and information related to a particular domain is accessible to the researchers and users, saving them the hassle of comprehending the difficult and challenging structure and format of the UMLS data. The research methodology covers all the aspects from collecting and mining data from UMLS resources to the transformation and data storage in the database using MySQL. Furthermore, the retrieval of required data from the application is also covered in this research. To gather all the data and information related to the concepts of COVID-19 and displaying them in a single live portal for the ease of the researchers related to the field of health care and medicine as well as the end users is the main aim of this research study. The live portal (https://ontologycovid.000webhostapp.com/) caters all the definitions, synonyms, relationships among the concepts as well as the semantic types of all the concepts related to the field of COVID-19. This methodology after complete development is validated by a panel of domain experts who declared this ontology as complete enough to describe the definitions and synonyms etc related to the domain of COVID-19 effectively. They agreed that the ontology provides the good understanding of all the concepts related to the COVID-19 domain. They also verified that the information available within the ontology is complete and accurate and covered on all the bases as it is extracted from the reliable resources which include various biomedical dictionaries and are up to date as they are maintained by the UMLS on a quarterly basis. In order to facilitate the users, additional information regarding COVID-19 is also included in the live portal so that the users can have the basic understanding of the disease without having to comprehend the complex terminologies of medicine. This data includes the aid in the form of visual and textual details from the renowned medical sources or websites.

7.4 Future Dimensions and Scope

This research can be extended further by adding various features to it. For the purpose of future dimensions, we have proposed that the users can get further assistance in the regards of this domain by asking the questions directly from the domain experts. The live portal can be enhanced by adding the registration feature for the end users. The users after signing up and logging into the system can ask about their queries regarding the domain of COVID-19 directly from the domain experts and they can get answers in the real time as well. The users who are unwilling to share their personal information from anyone can get their diagnosis done from the website easily as well as the people form the developing nations can get information and diagnostic data form this portal easily as they don't have the means to travel to the health care facilities or in many cases, they do not have access to the health care facilities.

Furthermore, the language restriction can also be tackled as UMLS consists of data in multiple languages. So, the generated ontology can also be updated in another language with the help of domain experts and health care professionals. In this way the incorporated ontology can be useful for the inhabitants of different languages and regions.

Another important endeavor to be added into this research can be the use of more graph and charts for the better understanding of the domain knowledge. The data from the ontology can be transformed into the format where it can be depicted in a graphical form. The data in the graphical representation or in a visual form is easier to comprehend as compared to the textual data. Therefore, the additional graphical and visual data along with the updated and improved layout of the web portal can also enhance the efficiency of our system.

CHAPTER 8

REFERENCES

References

[1] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H.-P. Kriegel, “Extraction of semantic biomedical relations from text using conditional random fields,” *BMC Bioinformatics*, vol. 9, no. 1, p. 207, Apr. 2008, doi: 10.1186/1471-2105-9-207.

[2] R. Feldman, Y. Regev, E. Hurvitz, and M. Finkelstein-Landau, “Mining the biomedical literature using semantic analysis and natural language processing techniques,” *BIOSILICO*, vol. 1, no. 2, pp. 69–80, May 2003, doi: 10.1016/S1478-5382(03)02330-8.

[3] G. Weikum and M. Theobald, “From information to knowledge: harvesting entities and relationships from web sources,” in *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, New York, NY, USA, Jun. 2010, pp. 65–76. doi: 10.1145/1807085.1807097.

[4] M. Ye, “Text Mining for Building a Biomedical Knowledge Base on Diseases, Risk Factors, and Symptoms,” p. 72.

[5] L. Soualmia, C. Golbreich, and S. Darmoni, “Representing the MeSH in OWL: Towards a Semi-Automatic Migration,” p. 9.

[6] N. F. Noy, D. L. Rubin, and M. A. Musen, “Making biomedical ontologies and ontology repositories work,” *IEEE Intell. Syst.*, vol. 19, no. 6, pp. 78–81, Nov. 2004, doi: 10.1109/MIS.2004.67.

[7] G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, and G. Antoniou, “Ontology change: classification and survey,” *Knowl. Eng. Rev.*, vol. 23, no. 2, pp. 117–152, Jun. 2008, doi: 10.1017/S0269888908001367.

[8] D. B. Lenat, “CYC: a large-scale investment in knowledge infrastructure,” *Commun. ACM*, vol. 38, no. 11, pp. 33–38, Nov. 1995, doi: 10.1145/219717.219745.

[9] U. Shah, T. Finin, A. Joshi, R. S. Cost, and J. Matfield, “Information retrieval on the semantic web,” in *Proceedings of the eleventh international conference on Information and knowledge management*, New York, NY, USA, Nov. 2002, pp. 461–468. doi: 10.1145/584792.584868.

[10] T. BERNERS-LEE, J. HENDLER, and O. LASSILA, “THE SEMANTIC WEB,” *Sci. Am.*, vol. 284, no. 5, pp. 34–43, 2001.

[11] J. Mayfield and T. Finin, “Information retrieval on the Semantic Web: Integrating inference and retrieval,” *Proc. SIGIR Workshop Semantic Web*, Aug. 2003, Accessed: Aug. 22, 2021. [Online]. Available:

https://ebiquity.umbc.edu/paper/abstract/id/88/Information-retrieval-on-the-Semantic-Web-Integrating-inference-and-retrieval

[12] “Reeve - 2000 - Substantial knowledge Aristotle’s metaphysics.pdf.” Accessed: Nov. 10, 2021. [Online]. Available: https://static1.squarespace.com/static/58d6b5ff86e6c087a92f8f89/t/5913d67a15cf7db9454a6ceb/1494472319242/Substantial_Knowledge__Aristotle__039_s_Metaphysics.pdf

[13] R. Studer, V. R. Benjamins, and D. Fensel, “Knowledge engineering: Principles and methods,” *Data Knowl. Eng.*, vol. 25, no. 1, pp. 161–197, Mar. 1998, doi: 10.1016/S0169-023X(97)00056-6.

[14] T. R. Gruber, “Toward principles for the design of ontologies used for knowledge sharing?,” *Int. J. Hum.-Comput. Stud.*, vol. 43, no. 5, pp. 907–928, Nov. 1995, doi: 10.1006/ijhc.1995.1081.

[15] N. Guarino, *Formal Ontology in Information Systems: Proceedings of the First International Conference (FOIS’98), June 6-8, Trento, Italy*. IOS Press, 1998.

[16] “WHO Coronavirus (COVID-19) Dashboard.” https://covid19.who.int (accessed Aug. 22, 2021).

[17] C. D. C. Reeve, *Substantial knowledge: Aristotle’s metaphysics*. Indianapolis: Hackett Pub, 2000.

[18] O. Bodenreider and R. Stevens, “Bio-ontologies: current trends and future directions,” *Brief. Bioinform.*, vol. 7, no. 3, pp. 256–274, Sep. 2006, doi: 10.1093/bib/bbl027.

[19] M. Uschold and M. Gruninger, “Ontologies: principles, methods and applications,” *Knowl. Eng. Rev.*, vol. 11, no. 2, pp. 93–136, Jun. 1996, doi: 10.1017/S0269888900007797.

[20] A. C. Yu, “Methods in biomedical ontology,” *J. Biomed. Inform.*, vol. 39, no. 3, pp. 252–266, Jun. 2006, doi: 10.1016/j.jbi.2005.11.006.

[21] S. Walk, S. Walk, and T. U. Graz, “kollaborativen.”

[22] P. G. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens, and A. Brass, “An ontology for bioinformatics applications.,” *Bioinformatics*, vol. 15, no. 6, pp. 510–520, Jun. 1999, doi: 10.1093/bioinformatics/15.6.510.

[23] “NCBI - Gone.” https://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0024568/ (accessed Nov. 11, 2021).

[24] C. Maria, “Aspects of Ontology Integration,” p. 87.

[25] G. Song, Y. Qian, Y. Liu, and K. Zhang, “Oasis: A Mapping and Integration Framework for Biomedical Ontologies,” in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS’06)*, Jun. 2006, pp. 611–616. doi: 10.1109/CBMS.2006.121.

[26] D. Fensel, “Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce.” 2007.

[27] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, “What are ontologies, and why do we need them?,” *IEEE Intell. Syst. Their Appl.*, vol. 14, no. 1, pp. 20–26, Jan. 1999, doi: 10.1109/5254.747902.

[28] M. A. Musen, “Scalable Software Architectures for Decision Support,” *Methods Inf. Med.*, vol. 38, no. 4/5, pp. 229–238, 1999, doi: 10.1055/s-0038-1634422.

[29] R. K. Saripalle, “Current status of ontologies in Biomedical and Clinical Informatics,” p. 15.

[30] A. Burgun, “Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System,” Dec. 2002.

[31] C. Rosse and J. L. V. Mejino, “A reference ontology for biomedical informatics: the Foundational Model of Anatomy,” *J. Biomed. Inform.*, vol. 36, no. 6, pp. 478–500, Dec. 2003, doi: 10.1016/j.jbi.2003.11.007.

[32] “protégé.” https://protege.stanford.edu/ (accessed Nov. 11, 2021).

[33] J. Michael, J. L. Mejino Jr, and C. Rosse, “The role of definitions in biomedical concept representation.,” in *Proceedings of the AMIA Symposium*, 2001, p. 463.

[34] J. L. V. Mejino, N. F. Noy, M. A. Musen, J. F. Brinkley, and C. Rosse, “Representation of Structural Relationships in the Foundational Model of Anatomy,” 2001, p. 973. Accessed: Nov. 11, 2021. [Online]. Available: http://sigpubs.si.washington.edu/id/eprint/71/

[35] P. Zweigenbaum, “MENELAS: an access system for medical records using natural language,” *Comput. Methods Programs Biomed.*, vol. 45, no. 1, pp. 117–120, Oct. 1994, doi: 10.1016/0169-2607(94)90029-9.

[36] J. Bouaud, B. Bachimont, J. Charlet, and P. Zweigenbaum, “ACQUISITION AND STRUCTURING OF AN ONTOLOGY WITHIN CONCEPTUAL GRAPHS_,” p. 25.

[37] D. a. B. Lindberg, B. L. Humphreys, and A. T. McCray, “The Unified Medical Language System,” *Yearb. Med. Inform.*, vol. 02, no. 1, pp. 41–51, 1993, doi: 10.1055/s-0038-1637976.

[38] “Cancer.” https://www.who.int/westernpacific/health-topics/cancer (accessed Nov. 11, 2021).

[39] B. Swartout, R. Patil, K. Knight, and T. Russ, “Toward Distributed Use of Large-Scale Ontologies,” p. 11.

[40] A. T. McCray and W. T. Hole, “The Scope and Structure of the First Version of the UMLS Semantic Networ,” *Proc. Symp. Comput. Appl. Med. Care*, pp. 126–130, Nov. 1990.

[41] D. L. Rubin, N. H. Shah, and N. F. Noy, “Biomedical ontologies: a functional perspective,” *Brief. Bioinform.*, vol. 9, no. 1, pp. 75–90, Jan. 2008, doi: 10.1093/bib/bbm059.

[42] C. Corbridge, G. Rugg, N. P. Major, N. R. Shadbolt, and A. M. Burton, “Laddering: technique and tool use in knowledge acquisition,” *Knowl. Acquis.*, vol. 6, no. 3, pp. 315–341, Sep. 1994, doi: 10.1006/knac.1994.1016.

[43] H. P. P. Filho, “Ontology Development 101: AGuide to Creating Your First Ontology,” p. 28.

[44] N. F. Noy and D. L. McGuinness, “Ontology Development 101: A Guide to Creating Your First Ontology,” p. 25.

[45] P. E. van der Vet and N. J. I. Mars, “Bottom-up construction of ontologies,” *IEEE Trans. Knowl. Data Eng.*, vol. 10, no. 4, pp. 513–526, Jul. 1998, doi: 10.1109/69.706054.

[46] “Cyc | The Next Generation of Enterprise AI.” https://cyc.com/ (accessed Nov. 11, 2021).

[47] “WordNet | A Lexical Database for English.” https://wordnet.princeton.edu/ (accessed Nov. 11, 2021).

[48] “GALEN’s model of parts and wholes: experience and comparisons.” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243933/ (accessed Nov. 11, 2021).

[49] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: A Bradford Book, 1998.

[50] R. Neches *et al.*, “Enabling Technology for Knowledge Sharing,” *AI Mag.*, vol. 12, no. 3, Art. no. 3, Sep. 1991, doi: 10.1609/aimag.v12i3.902.

[51] D. M. Pisanelli, A. Gangemi, M. Battaglia, and C. Catenacci, “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies,” *MEDINFO 2004*, pp. 416–419, 2004, doi: 10.3233/978-1-60750-949-3-416.

[52] V. Kashyap, A. Morales, and T. Hongsermeier, “On Implementing Clinical Decision Support: Achieving Scalability and Maintainability by Combining Business Rules and Ontologies.,” *AMIA. Annu. Symp. Proc.*, vol. 2006, pp. 414–418, 2006.

[53] P. Degoulet, D. Sauquet, M.-C. Jaulent, E. Zapletal, and M. Lavril, “Rationale and Design Considerations for a Semantic Mediator in Health Information Systems,” *Methods Inf. Med.*, vol. 37, no. 4/5, pp. 518–526, 1998, doi: 10.1055/s-0038-1634545.

[54] A. L. Rector, S. Bechhofer, C. A. Goble, I. Horrocks, W. A. Nowlan, and W. D. Solomon, “The GRAIL concept modelling language for medical terminology,” *Artif. Intell. Med.*, vol. 9, no. 2, pp. 139–171, Feb. 1997, doi: 10.1016/S0933-3657(96)00369-7.

[55] “In Hunt for Covid-19 Origin, Patient Zero Points to Second Wuhan Market - WSJ.” https://www.wsj.com/articles/in-hunt-for-covid-19-origin-patient-zero-points-to-second-wuhan-market-11614335404 (accessed Nov. 11, 2021).

[56] “Opinion | The Secret Life of a Coronavirus - The New York Times.” https://www.nytimes.com/2021/02/26/opinion/sunday/coronavirus-alive-dead.html (accessed Nov. 11, 2021).

[57] “Frontiers | Prevalence of Headache in Patients With Coronavirus Disease 2019 (COVID-19): A Systematic Review and Meta-Analysis of 14,275 Patients | Neurology.” https://www.frontiersin.org/articles/10.3389/fneur.2020.562634/full (accessed Nov. 11, 2021).

[58] M. A. Islam, S. Kundu, S. S. Alam, T. Hossan, M. A. Kamal, and R. Hassan, “Prevalence and characteristics of fever in adult and paediatric patients with coronavirus disease 2019 (COVID-19): A systematic review and meta-analysis of 17515 patients,” *PLOS ONE*, vol. 16, no. 4, p. e0249788, Apr. 2021, doi: 10.1371/journal.pone.0249788.

[59] “Prevalence and Characteristics of Taste Disorders in Cases of COVID-19: A Meta-analysis of 29,349 Patients - Jeyasakthy Saniasiaya, Md Asiful Islam, Baharudin Abdullah, 2021.” https://journals.sagepub.com/doi/full/10.1177/0194599820981018 (accessed Nov. 11, 2021).

[60] J. Saniasiaya, M. A. Islam, and B. Abdullah, “Prevalence of Olfactory Dysfunction in Coronavirus Disease 2019 (COVID-19): A Meta-analysis of 27,492 Patients,” *The Laryngoscope*, vol. 131, no. 4, pp. 865–878, 2021, doi: 10.1002/lary.29286.

[61] A. A. Agyeman, K. L. Chin, C. B. Landersdorfer, D. Liew, and R. Ofori-Asenso, “Smell and Taste Dysfunction in Patients With COVID-19: A Systematic Review and Meta-analysis,” *Mayo Clin. Proc.*, vol. 95, no. 8, pp. 1621–1631, Aug. 2020, doi: 10.1016/j.mayocp.2020.05.030.

[62] D. P. Oran and E. J. Topol, “The Proportion of SARS-CoV-2 Infections That Are Asymptomatic : A Systematic Review,” *Ann. Intern. Med.*, vol. 174, no. 5, pp. 655–662, May 2021, doi: 10.7326/M20-6976.

[63] CDC, “COVID-19 and Your Health,” *Centers for Disease Control and Prevention*, Feb. 11, 2020. https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html (accessed Nov. 11, 2021).

[64] CDC, “COVID-19 and Your Health,” *Centers for Disease Control and Prevention*, Feb. 11, 2020. https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html (accessed Nov. 11, 2021).

[65] CDC, “Healthcare Workers,” *Centers for Disease Control and Prevention*, Feb. 11, 2020. https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html (accessed Nov. 11, 2021).

[66] T. Koyama, D. Platt, and L. Parida, “Variant analysis of SARS-CoV-2 genomes,” *Bull. World Health Organ.*, vol. 98, no. 7, pp. 495–504, Jul. 2020, doi: 10.2471/BLT.20.253591.

[67] A. Rambaut *et al.*, “A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology,” *Nat. Microbiol.*, vol. 5, no. 11, pp. 1403–1407, Nov. 2020, doi: 10.1038/s41564-020-0770-5.

[68] “Tracking SARS-CoV-2 variants.” https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ (accessed Nov. 11, 2021).

[69] A. S. Lauring and E. B. Hodcroft, “Genetic Variants of SARS-CoV-2—What Do They Mean?,” *JAMA*, vol. 325, no. 6, pp. 529–531, Feb. 2021, doi: 10.1001/jama.2020.27124.

[70] S. S. Abdool Karim and T. de Oliveira, “New SARS-CoV-2 Variants - Clinical, Public Health, and Vaccine Implications,” *N. Engl. J. Med.*, vol. 384, no. 19, pp. 1866–1868, May 2021, doi: 10.1056/NEJMc2100362.

[71] “New COVID-19 variants.” https://stacks.cdc.gov/view/cdc/100425 (accessed Nov. 11, 2021).

[72] I. Zaragozá, J. Guixeres, and M. Alcañiz, “Ontologies for Intelligent e-Therapy: Application to Obesity,” in *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, Berlin, Heidelberg, 2009, pp. 894–901. doi: 10.1007/978-3-642-02481-8_136.

[73] B. Sabir, U. Qamar, and A. Muzaffar, *Ontology Development and Evaluation for Urinal Tract Infection*, vol. 2015. 2014. doi: 10.1109/CIBIM.2014.7015463.

[74] K. Wahab, U. Qamar, K. S. Arif, and U. Ali, “Building a Biomedical Ontology for Chronic Liver Disease,” in *2019 International Conference on Computer, Information and Telecommunication Systems (CITS)*, Aug. 2019, pp. 1–5. doi: 10.1109/CITS.2019.8862104.

[75] A. Iqtidar, A. W. Muzaffar, U. Qamar, and S. Rehman, “A biomedical ontology on genetic disease,” in *Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing*, New York, NY, USA, Mar. 2017, pp. 1–6. doi: 10.1145/3018896.3018966.

[76] J. Hanna, E. Joseph, M. Brochhausen, and W. R. Hogan, “Building a drug ontology based on RxNorm and other sources,” *J. Biomed. Semant.*, vol. 4, p. 44, Dec. 2013, doi: 10.1186/2041-1480-4-44.

[77] W. A. Kibbe *et al.*, “Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1071–D1078, Jan. 2015, doi: 10.1093/nar/gku1011.

[78] “protégé.” https://protege.stanford.edu/ (accessed Aug. 22, 2021).

[79] M. Horridge, T. Tudorache, C. Nuylas, J. Vendetti, N. F. Noy, and M. A. Musen, “WebProtégé: a collaborative Web-based platform for editing biomedical ontologies,” *Bioinformatics*, vol. 30, no. 16, pp. 2384–2385, Aug. 2014, doi: 10.1093/bioinformatics/btu256.

[80] A. Bellandi, S. Nasoni, A. Tommasi, and C. Zavattari, “Ontology-Driven Relation Extraction by Pattern Discovery,” in *2010 Second International Conference on Information, Process, and Knowledge Management*, Feb. 2010, pp. 1–6. doi: 10.1109/eKNOW.2010.17.

[81] D. L. Rubin, N. H. Shah, and N. F. Noy, “Biomedical ontologies: a functional perspective,” *Brief. Bioinform.*, vol. 9, no. 1, pp. 75–90, Jan. 2008, doi: 10.1093/bib/bbm059.

[82] O. Bodenreider, A. Burgun, and B. Ontologies, “Chapter 8 BIOMEDICAL ONTOLOGIES.”