Sudan University of Science and Technology

College of Graduate Studies

A Methodology for Evaluating Ontologies in the Biomedical Domain

منهجية لتقويم الأنطولوجيا في مجال الطب الحيوي

Submitted for the degree of Doctor of Philosophy
in Computer Science

Prepared By:
Abdelhakeem Mohamed Bashir Abdelrahman
Supervised By:
Prof. Dr. Ahmed Kayed

August, 2018

Declaration

I hereby certify that this material, which I now submit for assessment of the program of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others. And the work of others has been cited and acknowledged within the text of my work.

Signed

Student ID:

Date:

Dedication

To my lovely Parents,

Mohamed Bashir Abdelrahman Ahmad & Haram Mohamed Ahmed Yahia,

Wife,

Mawda Elwali Bashir Elwali

Kids,

Ghofran, Mohamed, and Amro

Brothers and Sisters.

Acknowledgments

First, I thanks Allah for given me the strength and ability to complete this study.

I also thanks my family: mother, father, brother, and sister; I also thanks my wife for support and encourage working on this thesis.

Thanks must also to faculty members of Computer Science and Information Technology College and those who taught and helped me much during my study.

Finally, I should extend special thanks to my supervisor Prof, Dr. Ahmad Kayed, without his help, guidance, and continues follow up; this research would never have been.

ABSTRACT

Infectious Disease Ontology (IDO) and its variants have been highly successful in implementing provides a consistent terminology, hierarchy, and logical representation for the domain of infectious and parasitic diseases. ICD's coverage of the domain in terms of types of infectious diseases is broad, but information about other aspects of infectious disease is limited and thus the scope of ICD-10 is considered narrow.

The great numbers, size, and complexity of biomedical ontologies make it difficult to choose appropriate ontology more adequate for given domain. The users will compare the ontologies and select higher quality ontology from more available ontologies for a single domain. Reference dataset are essential tools to check quality of any knowledge source. Currently there is no reference dataset to evaluate the quality of ontology from the perspective of semantic similarity measure, and there is no well defined reference dataset in the biomedical domain.

In this research, we proposed an approach that aids the development of a methodology for infectious and parasitic diseases. It based on biomedical domain ontology concepts/classes to compare between them using semantic similarity measure (SemDist) measure. The research approach consists of four interrelated components: select a semantic similarity measure, build reference dataset using SemDist measure, evaluate our reference dataset, and compare our reference dataset to two different ontologies. In the first part of this research, assessment of the applicability of using some measures from semantic similarity techniques has been investigated. This research builds biomedical domain taxonomy/hierarchy to be used by these measures. Several experiments have been conducted to select the best measure among all these measures. The experimental results validate the efficiency of the SemDist technique in single ontology and across ontologies, and demonstrate that the SemDist semantic similarity measure, compared with the existing techniques, gives the best overall results of correlation with experts' ratings. The reference dataset is built using ICD-10 "V1.0" ontology, infectious and parasitic diseases, named for Infectious and Parasitic DO-Reference dataset . We evaluate the approach according to a human expert in Human Disease Ontology by comparing his diseases diagnosis to those of the reference dataset, reference dataset showed good accuracy in the results were 80.6% compare to document physicians answers. We evaluate the (doid) ontology within Unified Medical Language System (UMLS) framework it indicate that the accuracy of using Infectious and Parasitic DO- Reference dataset at lexical level and conceptual level is 69cocepts (52.6) and 75% respectively. When, we evaluate the (SNOMED-CT) ontology within UMLS framework, it indicate that the accuracy of using Infectious and Parasitic DO- Reference dataset at lexical level and conceptual level is 81cocepts (62.8) and 86.3% respectively. In addition, we use the feature "compare ontologies tools" in protégé to insure the accuracy of results.

Keywords: *Semantic similarity measure, Semantic web, Infectious and Parasitic DO-Reference Dataset, Ontology evaluation, Biomedical domain, UMLS framework.*

ABSTRACT (Arabic)

المستخلص

تعتبر انطولوجيا الأمراض المعدية و الطفيلية بأنواعها المختلفة ناجحة جداً من حيث اتساق المصلحات، التسلسل الهرمي، والتمثيل المنطقي لمجال الأمراض المعدية و الطفيلية. يقوم التصنيف العالمي للأمراض بتغطية أنواع الأمراض المعدية والطفيلية بصورة واسعة، لكن المعلومات عن الجوانب الأخرى عن الأمراض المعدية محدودة و بالتالي فان نطاق التصنيف العالمي للأمراض الإصدار العاشر (ICD-10)يعتبر ضيقاً .

الأعداد الكبيرة، حجم، وتعقيد الانطولوجيات الطبية الحيوية جعل من الصعب اختيار الانطولوجيا المناسبة للمجال المعين. ليتمكن المستخدم من المقارنة واختيار الانطولوجيا الأفضل من بين الانطولوجيات، كان الاحتياج لمجموعة بيانات مرجعية كأداة أساسية للتأكد من جودة الانطولوجيا. حالياً لا يوجد مجموعة بيانات مرجعية لقياس جودة الانطولوجيا في مجال الطب الحيوي لمعرفة التشابه بين المفاهيم والمقارنة بينها، ولا توجد مجموعة بيانات مرجعية في مجال الطب الحيوي.

في هذا البحث، نقترح منهجية تساعد على تطوير مجموعة البيانات المعيارية من انطولوجيا الأمراض المعدية والطفيلية. تستند هذه المنهجية على مفاهيم/فئات انطولوجيا الطب الحيوي للمقارنة بين تلك المفاهيم باستخدام تقنية (SemDist). تتكون المنهجية من أربعة عناصر مترابطة: قياس التشابه الدلالي، وبناء قاعدة بيانات مرجعية باستخدام تقنية قياس التشابه (SemDist)، وتقييم الأداة بعد تطويرها، ومقارنتها بنوعين مختلفين من الانطولوجيا في نفس المجال. في المرحلة الأولى من هذا البحث، تم تقييم مدى إمكانية استخدام وتطبيق بعض أنواع القياس لمعرفة التشابه المعنوي بين المفاهيم في مجالات الطب الحيوي داخل إطار هيكل اللغة الطبية الموحد. في هذا البحث تم بناء تصنيف التسلسل الهرمي في المجال الطبي الحيوي لاستخدامه بواسطة طرق القياس المختلفة. وقد أجريت العديد من التجارب لتحديد أفضل مقياس من بين هذه الطرق. تؤكد النتائج التجريبية فعالية تقنية (SemDist) في تقييم الانطولوجيا الواحد وبين أكثر من انطولوجيا، وتوضح أن قياس التشابه الدلالي باستخدام SemDist، يعطي أفضل النتائج الإجمالية للارتباط مقارنة مع تقديرات الخبراء في مجال الطب الحيوي. مجموعة البيانات المعيارية هي عبارة عن بناء انطولوجيا باستخدام بعض المفاهيم من الانطولوجيا "V1.0" ICD-10، الأمراض المعدية والطفيلية، وتم تسمية هذه الأداة باسم *Infectious and parasitic DO-Reference dataset*. قمنا بتقييم المنهجية المقترحة من خلال مقارنة تشخيص الأمراض الموجودة في قاعدة البيانات المعيارية، مع " Human Disease Ontology" (Doid)، حيث أظهرت مجموعة البيانات القياسية دقة جيدة في النتائج كانت 80.6 ٪ مقارنة مع إجابات الأطباء. قمنا بتقييم الانطولوجيا (doid) داخل إطار هيكل اللغة الطبية الموحد (UMLS) وتشير النتائج إلى أن دقة استخدام *Infectious and* DO-Reference dataset *parasitic* على مستوى المعجم ومستوى المفاهيم هي (52.6) (69 concepts) و 75٪ على التوالي. أيضاً قمنا بتقييم الانطولوجيا (SNOMED-CT) ضمن إطار عمل UMLS، وتشير النتائج إلى أن دقة استخدام *Infectious and* DO- Reference dataset *parasitic* على مستوى المعجم و مستوى المقارنة بين المفاهيم هي (62.8) (81 concepts) و 86.3٪ على التوالي. بالإضافة إلى ذلك، فإننا نستخدم أدوات مقارنة مثل Protégé tool لضمان التأكد من صحة النتائج.

كلمات مفتاحية: تقنيات قياس التشابه، الويب الدلالي، مجموعة بيانات مرجعية، تقييم الانطولوجيا، الطب الحيوي، إطار اللغة الطبية الموحد.

Table of Contents

List of Tables

LIST OF FIGURES

LIST OF ABBREVIATIONS

SNOMED-CT	Systematized Nomenclature of Medicine-Clinical Terms
UMLS	Unified Modeling Language System
DOID	Human disease Ontology
OWL	Web Ontology Language
LCS	Least Common Subsume
MeSH	Medical Subject Headings
OMIM	Online Mendelian Inheritance in Man
ICD10	International Classification of Disease -10
IDO	Infectious Disease Ontology

Chapter One

Introduction

1.1 Background

Ontologies are formal specification of share conceptualization of a domain and relations among them [1, 2]. Some of the ontologies, which are formal representations of knowledge, can be used for designing and sharing conceptual models within a domain for the purpose of enhancing understanding, communication and interoperability [3]. Ontology presents a common understanding of the knowledge domain using major concepts and terms applied in that domain and identify the relationships between these concepts. Ontology can be built from scratch or it can reuse existing ontology [4]. The ontology evaluation utilities that are currently available allow the user to check the internal consistency of ontology. The whole set of tests or particular test can be executed at any time, hence, it simplifies the testing of ontology both during its development and during its evolution [5].

In the health domain, a large percentage of clinical trials are still using primary data collection tool such as paper form [6]. The great numbers, size and complexity of biomedical ontologies make it difficult to choose appropriate ontologies more adequate for given domain [7].

Important applications of ontologies include distributed knowledge-based systems, such as the semantic web, and the evaluation of modeling languages. These applications require formal ontologies of good quality. The quality of a formal ontology requires both a good conceptualization of a domain and a good specification of the conceptualization [1]. The quality of ontology is its degree of conformance to functional and non-functional requirements and we assume that such conformance can be measurable. Current work in ontology evaluation can be classified according to the particular evaluation aim: ranking, correctness, or quality evaluation [8]. The quality of ontologies, which in contrast to conceptual models have to satisfy computational requirements as well as representational requirements, has been characterized by ability to answer competency questions. Assessing the quality of ontology has become an important issue to help the ontology engineers to predict the quality of ontologies. The users will compare the ontologies and select higher quality ontology from more available ontologies for a single domain. Based on the quality metrics the ontology users can assess the quality of ontology. Metrics measure the quality of ontologies at both structure and the semantic level [9].

To be able to measure the quality of ontology we need a Reference dataset. This research will focus on how to build this reference dataset or standard definition and their measures in the health domain. Experiments are then appropriately designed to evaluate the qualities of typical ontologies to show the effectiveness of the proposed evaluation methods [10]. The use of the Semantic Web depends on the two types of evaluation: evaluation of the content of semantic web (ontology evaluation), evaluating content is a must for preventing applications from using inconsistent, incorrect, or redundant ontologies, and evaluation of the technologies that use the content of the Semantic Web (Semantic Web technology evaluation) [34]. In this thesis, evaluation is only considered in terms of the ontology content evaluation. We used RDFs and OWL as interchange language. They involve evaluating and importing ontology content.

1.2 Motivation the need for Reference dataset in the Semantic Web.

The purpose of the evaluation is to enable a system to rank ontologies returned by search engines according to how well the ontologies perform under certain measures. Due to the complex structure of ontologies and difficult terminologies of biomedical domain, the evaluation of these ontologies turns out to be a challenging task. It is utmost need of current ontology researchers and developers to evaluate the quality of these biomedical ontologies so that the applicability and reuse of these ontologies will be improved.

In motivation of this need we have proposed a methodology of evaluating the quality of biomedical ontologies with respect to basic ontology structural building blocks especially with respect to properties or relations including object properties, data properties, annotation properties, inverse properties, functional properties, symmetric properties, asymmetric properties and reflexive properties. SPARQL queries are used to extract their population frequency. Experimentations provide evidences that these structural properties/relations between the concepts are of core significance in ontology evaluation.

Any advance research is based on existing research results. In the case of ontology evaluation, the reuse and improvement of existing development after they have been evaluated and compare with others, these for any type of software, is also applicable to semantic web software.

1.3 Problem statement and its significant

The great numbers, size, and complexity of biomedical ontologies make it difficult to choose appropriate ontologies more adequate for given domain. For the enhancement of the quality of ontologies in the biomedical domain, Reference dataset or standard definitions are needed to check the quality of knowledge sources. Reference dataset is essential tools to check quality of any knowledge source. According to Hisham Al-Mubaid & Hoa A. Nguyen [11, 12], there is no standard approach to evaluate the quality of ontology from the perspective of semantic similarity measure, and there is no well defined Reference dataset in the biomedical domain. The challenge is define how to build this Reference dataset using existing resources (ICD10) as well as to define a measure to use this Reference dataset to evaluate ontologies in the health domain. This work contributes development of techniques and measures to evaluate ontologies in the biomedical domain.

1.4 Research Questions

Depend on problem definition and objectives of study it has been put some of theories: The main question that will be addressed in this research is: How we build a Reference dataset used to solved an ontology evaluation techniques problem?

There are additional sub-questions as follows:

1. How do we deploy one of the existing similarity measures to check the quality of an ontology using the proposed reference dataset?

2. How we can extract dataset from the domain knowledge sources such as ICD-10, and how we use this dataset to check the quality of ontologies?

1.5 Research Objectives

In the past several years, some ontology toolkits, such as Jena, KAON2, Protégé, and Sesame, had been developed for ontologies storing, reasoning and querying. A standard and effective reference dataset to evaluate existing systems is much needed. The main objective of this research is to build a reference dataset used it to evaluate a quality of ontologies in biomedical domain.

Other specific objectives highlight as follows:

I. To investigate in reference dataset (standard definitions) or method to evaluate ontologies.
II. To build reference dataset.
III. To evaluate proposed solution.

1.6 Research Scope

This research is mainly focus at build reference dataset of the ontology in the biomedical domain within UMLS frame work.

This is to check quality of ontologies by using semantic similarity measures to evaluate any ontology in the biomedical domain comparing to our reference dataset.

In this thesis, evaluation is only considered in terms of the ontology content evaluation. We used RDFs and OWL as interchange language. They involve evaluating and importing ontology content.

1.7 Research hypothesis

The assumptions we took to build the reference dataset is the following:

I. There are many available ontologies in the biomedical domain with in UMLS framework.
II. There are many knowledge resources in the biomedical domain with in UMLS framework.

III. There are no reference dataset in the biomedical domain with in UMLS framework.

1.8 Organization of Thesis

The rest of the thesis is structured in the following chapters: Chapter 1 (Introduction) presents a background of ontology, and the thesis problem and objectives. Chapter 2 (Background and Related Work) presents a survey of the current state of software evaluation and reference dataset; it also describes different evaluation and improvement methodologies. Chapter 3 (reference dataset Methodology for Semantic Web content) describe the development of Infectious and Parasitic DO- Reference Dataset by using clustering method. Chapter 4 (Infectious and Parasitic DO-Reference Dataset Development) phases for developing reference dataset as ontology. Chapter 5 (Semantic similarity measure). Chapter 6 (Testing and Evaluation). The last chapter (Conclusion and Future Work).

Chapter Two

Background and Related Work

2.1 Overview

This chapter provides a high level explanation of the tools and technologies used in the development of the Semantic Similarity to achieve the objectives outlined in (Chapter 1).Terms used in this thesis are defined and explained here.

2.2 The Semantic Web:

2.2.1 Ontologies:

Ontology translates from the Greek onto (begin) + logos (word). It was introduce in nineteen century by German philosophers [30]. In the context of semantic web, an ontology is the backbone of knowledge representation, and a key component of the Semantic Web [60]. It can be incorporated into computer based systems to facilitate data annotation, decision support, information retrieval, and natural-language processing [56]. Nguyen [41] define ontology as a description of the terms/concepts and relationships between them in a given domain and is used to denote for all kind of *IS-A* trees or hierarchical trees in which concepts are represented hierarchically by *IS-A* relations (*is-a-kind-of*, *is-a-part-of*) although the hierarchical relations in biomedical domain in the framework of UMLS are broader/narrow than relations.

2.3 Structure of Ontologies:

2.3.1 OWL

The Web Ontology Language (OWL) [10] is a family of knowledge representation languages standardized by the World Wide Web Consortium. It has three increasingly expressive sublanguages: OWL-Lite, OWL-DL, and OWL-Full. OWL-DL and OWL -Lite semantics are based on Description Logics, which have well-known computational properties and automated reasoning support, while OWL-Full is intended to provide compatibility with RDF Schema. We will discuss the details of those three sublanguages and the next generation of OWL (OWL 2) [32]. As knowledge representation formalism, OWL ontology consists of a set of axioms which place logical constraints on the classes (sets of individuals) and properties (relationships between individuals) in a domain of our interest.

These axioms provide Description Logics [6] based formal semantics such that the intelligent system can infer implicit knowledge from the explicitly represented knowledge.

In OWL ontologies:

- Concepts are referred to as classes.
- Class Expressions refer to the (possibly complex) concepts that are present within the ontology therefore it follows that classes are also class expressions.
- Individuals are also referred to as individuals.
- Roles are referred to as properties and can be broken down into two groups. *Object properties* that describe the relationships between objects and *Data type properties* that relate objects to built in data types

2.3.2 The OWL API

is used to create and interact with OWL Ontologies. It is open-source project developed at the University of Manchester [60]. It provides data structures that allow users model and manipulate OWL ontologies and also provides a reasoner interface allowing for "a representation that implements/understands the formal semantics of the language"[BM14]. This means that the reasoner is able to "listen" for changes to the ontologies it is reasoning over and will also respond to user queries with respect to the changed ontologies[Hor09].

2.3.3 Description Logics

In DLs there are three kinds of entities:

- **Concepts** represent the set of individuals.
- **Individuals** represent single individuals within the domain.
- **Roles** represent the relationships between individuals.

DL ontologies are comprised of a state of statements called **axioms**. These statements describe the relationships between concepts, individuals and roles. Axioms can be classified into three groups[60]:

Assertional axioms(A Box) that describe relationships between named individuals and concepts or between the individuals and roles. For example the axiom, *Lion*(*Simba*), asserts that the individual named Simba is an instance of the concept *Lion*.

Terminological axioms(T Box) that describe the relationships between concepts. For example the axiom, *Lion v Carnivores*, states that every instance of the concept *Lion* is also an instance of the concept *Carnivore*.

Relational axioms(R Box) that describe the relationships between roles. For example the axiom4, *brotherOf* ◦*parentOf v uncleOf*, states that any individual that is a brother to another individual that is a parent is an uncle.

2.4 Ontologies Classification according to a Semantic Spectrum:

Controlled vocabularies: are finite lists of terms.

Glossaries: are lists of terms whose meaning is described in natural language. The format of a glossary is similar to that of a dictionary, where terms are organized in alphabetical order, followed by their definitions.

Thesauri: are lists of terms and definitions that standardize words for indexing purposes. Besides definitions, a thesaurus also provides relationships between the terms: the hierarchical, associative, or equivalence (synonymous) relationships.

Informal is-a hierarchy: are hierarchies that use generalization (*type-of*) relationships in an informal way. In this kind of hierarchy, related concepts can be aggregated into a category, even if they do not respect the generalization relationship.

Formal *is-a* hierarchy: are hierarchies that fully respect the generalization relationship.

Frames: are models that include classes and properties after the frame representation. The primitives of the frame model are classes, or frames, that have properties, slots, or attributes. Slots do not have global scope, but they apply only to the classes for which they were defined. Each frame provides the context for modeling some aspect of the domain. Several refinements and extensions have been proposed to the frame model. Frames are largely used in modeling knowledge bases.

Ontologies that express value restrictions: are ontologies that provide constructs to restrict the values their class properties can assume.

Ontologies that express logical restrictions: are ontologies that allow first-order logic restrictions to be expressed.

Classifying Ontologies According to Their Generality Guarino (1998) proposes a classification based on the generality of the ontology, as follows:

Upper Level Ontologies describe generic concepts, such as space, time, and events. These ontologies are, in principle, domain independent and can be reused to construct new ontologies. Domain Ontologies describe the vocabulary pertaining to a given domain, by specializing the concepts provided by the upper-level ontology.

Task Ontologies describe the vocabulary required to perform generic tasks or activities, again by specializing the concepts provided by the upper-level ontology.

Application Ontologies describe the vocabulary of a specific application, whose concepts correspond, in general, to the roles performed by entities in a given domain while performing some task or activity. Guarino (1998). [44]

2.5 Biomedical Ontologies

Benefits of Ontologies in Biomedicine Ontologies can enhance how biomedical data are organized and managed, as well as enrich Web functionality [56].

2.5.1 UMLS

The Unified Medical Language System (UMLS) can be considered as an example of terminology which contains many clinical terms and integrates about 100 different vocabularies [41, 19].

What is the UMLS?

The UMLS, “is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems”

UMLS used to enhance or develop applications, such as electronic health records, classification tools, dictionaries and language translators [81].

It consists of three main knowledge sources: Metathesaurus is a terms and codes from many vocabularies, including (ICD-10-CM, MeSH, SNOMED-CT thesauruses, etc.), Semantic Network: Broad categories (semantic types) and their relationships (semantic relations), and SPECIALIST Lexicon & Lexical Tools: Natural language processing tools [19, 81].

2.5.1.1 MeSH: MeSH, stands for **Me**dical **S**ubject **H**eadings, [19, 58], is one of the source vocabularies used in UMLS. MeSH includes about 15 high-level categories, and each category is divided into subcategories and assigned a letter: A for Anatomy, B for Organisms and C for Diseases, and so on.

2.5.1.2 SNOMED-CT: SNOMED-CT, stands for Systemized Nomenclature of Medicine Clinical Term [19, 58], was included in UMLS in May 2004. It is a comprehensive clinical terminology, and the current version contains more than 360,000 concepts, 975,000 synonyms and 1,450,000 relationships organized into 18 hierarchies.

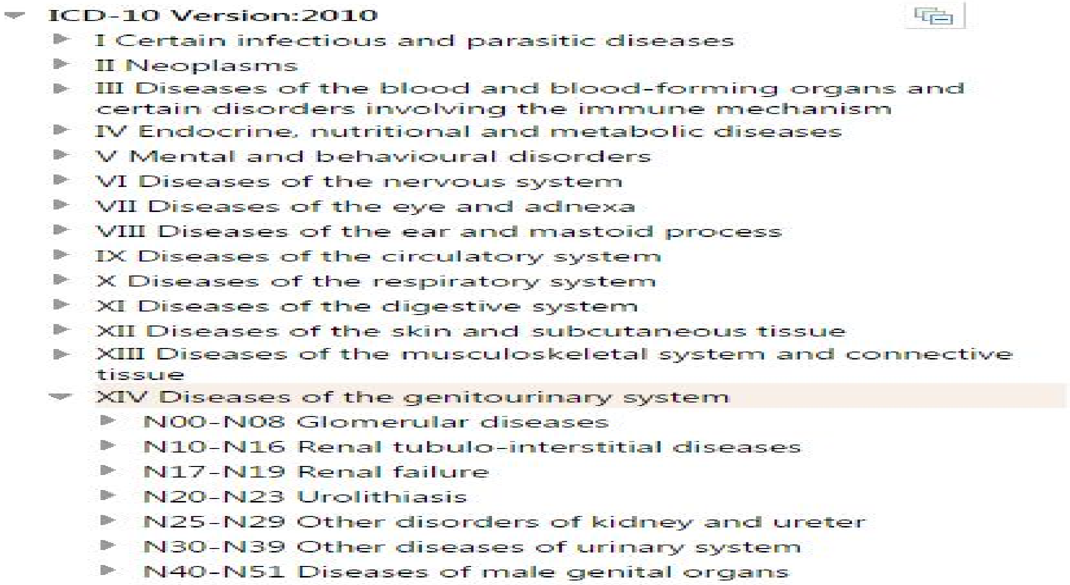
The following ontologies can be considered as known ontologies in the medical domain:

2.5.1.3 NCI Thesaurus (National Cancer Institute Thesaurus): an ontology vocabulary that includes broad coverage of the cancer domain, including cancer related disease, anatomy, genes and drugs.

2.5.1.4 ICD-10 stand for International Classification of Diseases 10threvision: An international standard used to classify diseases and other health problems adopted by World Health Organization (WHO) [42]. Its being the main indexes for disease identification and classification [39].

2.5.1.5 Human disease Ontology (DOID): an open source ontology for the integration of biomedical data that is associated with human diseases [43].

2.5.2 ICD: [42], is one of the most important international medical terminological systems; it was first issued in 1893. Its sixth revision was in 1948, and since this time it has been maintained by the World Health Organization (WHO). The current version is the tenth revision (ICD-10), which was issued in 1992. The initial aim of the ICD was to provide an international classification of death causes in order to produce internationally uniform and thus comparable mortality statistics. The WHO family of international classifications also includes other systems, notably the ICF (International Classification of Functioning, Disabilities and Health) and ICHI (International Classification of Health Inventions). The 22 main sub-categories of ICD-10 include, among others, diseases of the blood and blood-forming organs (D50–D89), endocrine, nutritional and metabolic diseases (E00–E90), mental and behavioral disorders (F00–F99), diseases of the nervous system (G00–G99) and certain infections and parasitic diseases (A00– B99). We present some preliminary observations about ICD-10 and consider the sub-domains I–XVII (codes A00 Q99). Core ontology of ICD-10 must explicate what sub-domains I–XVII address. Six of these domains are classified with respect to systems (nervous system, circulatory system, respiratory system, digestive system, musculo-skeletal system, genito-urinary system), three pertain to special organs (eye, ear, skin), and one domain relates to infectious diseases (A00–B99) and one domain addresses mental and behavioral disorders (F00–F99). Sub-domain level categories Level (i), i = I... XVII may be introduced; their instances are subsumed by the corresponding chapters. The instances of a level category level (i) in ICD-10 exhibit a taxonomic structure. Consider the domain of infections and parasitic diseases (A00–B99) and the associated domain-level category level (I),and includes about 21 high-level categories (taxonomies/sub trees) as shown in Figure2.1. The 2016 release of ICD-10 was used in our experiments.



ICD-10 Version:2010
- I Certain infectious and parasitic diseases
- II Neoplasms
- III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- IV Endocrine, nutritional and metabolic diseases
- V Mental and behavioural disorders
- VI Diseases of the nervous system
- VII Diseases of the eye and adnexa
- VIII Diseases of the ear and mastoid process
- IX Diseases of the circulatory system
- X Diseases of the respiratory system
- XI Diseases of the digestive system
- XII Diseases of the skin and subcutaneous tissue
- XIII Diseases of the musculoskeletal system and connective tissue
- XIV Diseases of the genitourinary system
  - N00-N08 Glomerular diseases
  - N10-N16 Renal tubulo-interstitial diseases
  - N17-N19 Renal failure
  - N20-N23 Urolithiasis
  - N25-N29 Other disorders of kidney and ureter
  - N30-N39 Other diseases of urinary system
  - N40-N51 Diseases of male genital organs

Figure 2.1 1 Overview of ICD-10 ontology by ICD-10 browser

2.6 Ontology Quality

The simplest methods evaluate ontologies as directed graphs in which the distance between two concepts is measured as the number of edges of the shortest path between them [58]. In this thesis we used semantic similarity measure to evaluate ontologies in the biomedical domain using ICD10 “V1.0” as knowledge source.

Brank et al. [45, 46] grouped various evaluation approaches into four categories. The first approach, called the gold-standard approach, in this approach, the gold standard ontology is regarded as a well-constructed one. We need another existing ontology, or it could be taken statistically from a corpus of documents or prepared by a domain expert. The concepts of a constructed ontology are evaluated by comparing them with those of gold standard ontology, which are considered good representations of the concepts for the problem domain under consideration [17]. Typically, the gold standard approach is used to evaluate an ontology generated by a learning process. The second one is an application-based approach in which the quality of the ontology is evaluated based on its actual use in a real-world application [62]. The output of the application or its performance on the given task might be better or worse depending on the ontology used in it. Ontologies may therefore be evaluated simply by plugging them into an application and evaluating the results of such application. However, if ontology is only a small component of the application, it is difficult to judge its quality using the application-based approach because its effect on the outcome may be relatively small and indirect [5]. The third approach is data-driven because it evaluates the quality of ontology by measuring the fit between the ontology and the corpus of a problem domain to which it refers. Thus, this approach evaluates ontology by measuring the amount of overlap between the domain-specific terms in the corpus and terms appearing in the ontology [6]. Since ontology is a fairly complex structure, it should be evaluated on the lexical, semantic, syntactic, and context levels. In the data-driven approach, however, ontology is evaluated only on the lexical level [5]. The final approach relies on human judgment. In this approach, the evaluation is done by domain experts who try to assess how well the ontology meets a set of predefined criteria, standards, and requirements. Although this evaluation requires a longer time, this approach can evaluate ontology in various perspectives including lexical, semantic, syntactic, and context levels. Our approach belongs to the last category.

2.7 Ontology Evaluation:

The goals of evaluating software depend on each specific case, but they can be summarized as follows[53-55]:

- ✓ To describe the software in order to understand it and to establish baselines for comparisons.
- ✓ To assess the software with respect to some quality requirements or criteria and determine the degree of desired quality of the software product and its weaknesses.
- ✓ To improve the software by finding opportunities for enhancing its quality. This improvement is measured by comparing the software with the baselines.
- ✓ To compare alternative software products or different versions of a same product.
- ✓ To control the software quality by ensuring that it meets the required level of quality.
- ✓ To foresee in order to take decisions, establishing new goals and plans for accomplishing them.

Multiple levels of evaluation were conducted[49].

1. Direct – evaluation of the ontology structure and content.

2. Application-based – evaluates the results from an application that uses the ontology.

3. Analysis-based – evaluates the use of the ontology as tool in scientific data analysis. The evaluation for this foundational project occurs primarily in the first level, “Direct,” through the domain conceptualization and ontology class identification.

2.7.1 Levels of Evaluation

Evaluation approach levels was addressed using either triangulation (lexical/conceptual, semantic relations), Protégé tools (hierarchy/taxonomy, syntactic, architecture), or through expert review described below (lexical/conceptual, semantic relations, context/application). No gold standard for comparison exists as this is foundational domain ontological work. Similarly, the domain source data is not normalized to a degree that could provide a standardized comparison. Thus, the approach for this work falls into the categories of application and human assessment[50].

Ontology content evaluation has three main underlying ideas[48]:

- We should evaluate ontology content during the entire ontology life cycle.
- Ontology development tools should support the content evaluation during the entire ontology-building process.
- Ontology content evaluation is strongly related to the underlying knowledge representation (KR) paradigm of the language in which the ontology is implemented.

2.8 Protégé Tool:

Started in 1987, when Mark Musen build a Meta tool for knowledge based system in the medical domain. Protégé is developed by Stanford medical informatics at the Stanford university school of medicine, with support for a number of government agencies and private institutions. Protégé is used as an ontology design interface and for collaboration, inference, and reasoning [31].

Protégé was selected as the primary tool for developing the OWL framework due to the following reasons: 1) Protégé is an open source, free ontology editor which maintains two key types of modeling ontologies via the Protégé-Frames and Protégé-OWL editors; 2) It provides a wide set of customizable user interface elements which allows easy access, hierarchical tree structure for class browsing, form interface for filling in slot values; 3) It supports several formats including RDF(S), OWL, and XML Schema; 4) Protégé which is based on Java has a great extensibility and scalability with its open modular design, which allows convenient functionality extension by adding or creating plug-ins; 5) Such a plug-and-play environment makes Protégé a flexible base for rapid prototyping and application development; 6) Protégé has been developed and tested for many years with a big group of users in bioinformatics area worldwide and with continuous support commitment [39]

The main strengths of Protégé-2000 compared to the other systems are its user interface, the extendibility using plug-ins, the functionality that the plug-ins provide (such as merging) as well as the different formats that can be imported and exported. Protégé-2000 is an old version of Protégé, and up till now, the latest version of Protégé also holds the advantages of the other three ontology editors [40]. Protégé is a free, open source ontology editor and a knowledge acquisition system. It supports ontology developers to think about domain models at a conceptual level without having to know the syntax of the language ultimately used on the Web (Noy et al., 2001). It can develop ontology in its own format, and can import or export ontology in RDF, RDFS, DAML+OIL, XML, OWL, Clips and UML. It can browse classes and properties via plug-ins (OntoViz, TGViz) and its query tab allows searching. There are many plug-ins available for extending ontology construction, constraint axiom, inferring and integration functions. Ontologies in the research have been built through Protégé to capture and represent concepts, their relationship, and instances in product design [40].

A general ontology modeling procedure in Protégé is described below:

1) Create an OWL ontology project

2) Create a new class and name it

3) Specify disjoint classes (if necessary)

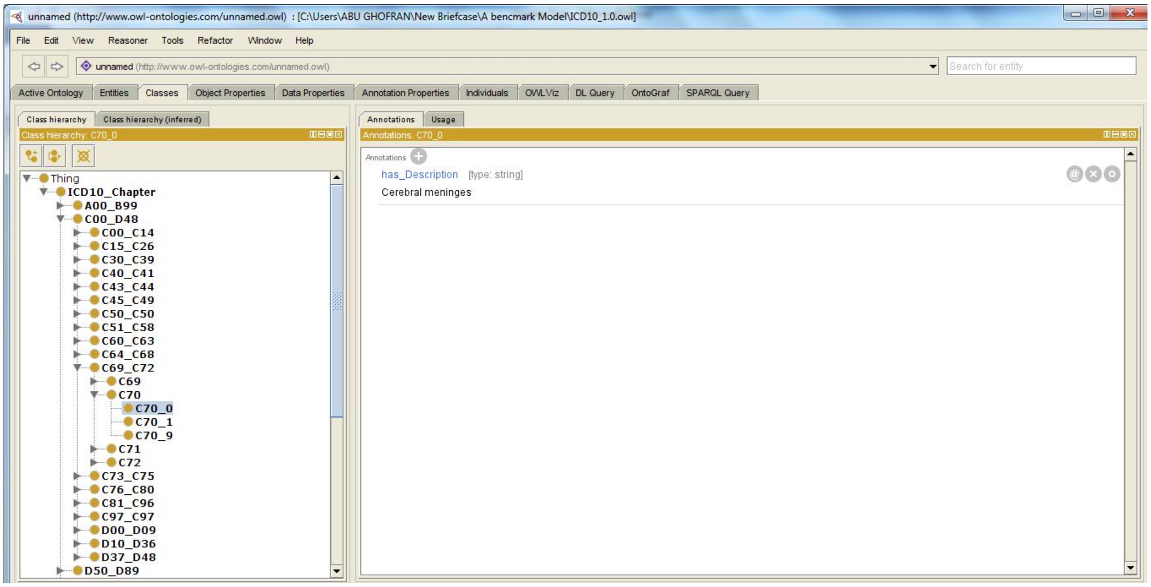
4) Create properties for the class and specify domain and range

5) Use properties to define the class and specify restrictions

6) Repeat steps 2 to 5

7) Create instances for the problem domain

Details about modeling OWL ontology in Protégé can be found in Horridge et al. (2007). The simple ontology shown in Figure 2.1 is loaded into Protégé and shown in Figure 2.2 This figure shows the basic class editing window of Protégé. The class hierarchy is represented on the tree view at the left side of the window. The right side of the window is the space for editing details of each OWL class [out (7) ontology book]



unnamed (http://www.owl-ontologies.com/unnamed.owl) : [C:\Users\ABU GHOFRAN\New Briefcase\A bencmark Model\ICD10_1.0.owl]

unnamed (http://www.owl-ontologies.com/unnamed.owl)

Annotations: C70_0

has_Description [type: string]
Cerebral meninges

Figure 2.2 1 Protégé user interface

Figure 2.2 shows the Protégé User Interface, Protégé[62] is an open-source ontology editor developed at Stanford University. Protégé was developed to be compatible with the OWL API and use it for ontology modeling and querying. It has OWL reasoned implementations that are built as plug-in. Protégé has a core API that contains many reusable UI elements and utility classes for plugin development. Within Protégé there are two possible ways to view the class hierarchy off an ontology. The first is the asserted class hierarchy which shows the subclass hierarchy that can be obtained directly from the ontology. Child nodes are subsumed by the parent(s) and anything without a parent shows up under the root node(>). The second, the inferred class hierarchy provides a more complete view of this by using the reasoner infer more relationships. In addition, the bottom concept(⊥) is added to this view. This class becomes the parent of every class within the ontology that is un satisfiable, that is, classes that can have no instances. The current release of Protégé was developed in collaboration with the University of Manchester.

Semantic Similarity Measures:

Semantic Similarity between two terms or sets of documents is defined as the degree of "sameness" between the terms as measured by comparing the information describing their properties [60]. Ontology-based semantic similarity measures are the similarity between two concepts, which is widely used in *information retrieval* and *semantic web service fields* [51].

2.9 Semantic Similarity and Relatedness

Semantic similarity is concerned about likeliness; relatedness seeks to determine relation between two terms/concepts. For example, “car” and “driver” are related, but not much similar, but “car” and “vehicle” are similar in some degree. Relatedness is thus more general than similarity. Furthermore, semantic distance is the inverse of semantic similarity that is the less distance of the two concepts, the more they are similar. To insure the conversion from semantic distance to semantic similarity do not change the absolute correlation value, the transformation function below is used:

$$Sim\ (C_1, C_2) = \text{MaxDist- Dist}\ (C_1, C_2) \qquad (1)$$

Where:

Dist is the semantic distance of two concepts, MaxDist is the maximum distance of two concepts and Sim is the converted semantic similarity of the two concepts. However, in this thesis, absolute correlation is used to evaluate performances of the approaches.

2.9.1 SEMANTIC SIMILARITY MEASURES CLASSIFICATION

Figure 3 and 4 [12]: illustrate the semantic similarity classification for single ontology and cross ontologies. To find Semantic Similarity between two concepts in ontology, by find *shortest path length* between them in the ontology (*shortest path length*) giving the length are:*is-a/part of.* Number of approaches have been developed using ontology as primary information sources. However, most of the semantic similarity techniques such as general English ontology based structure similarity measures can be adopted to be used into the biomedical domain within UMLS framework.

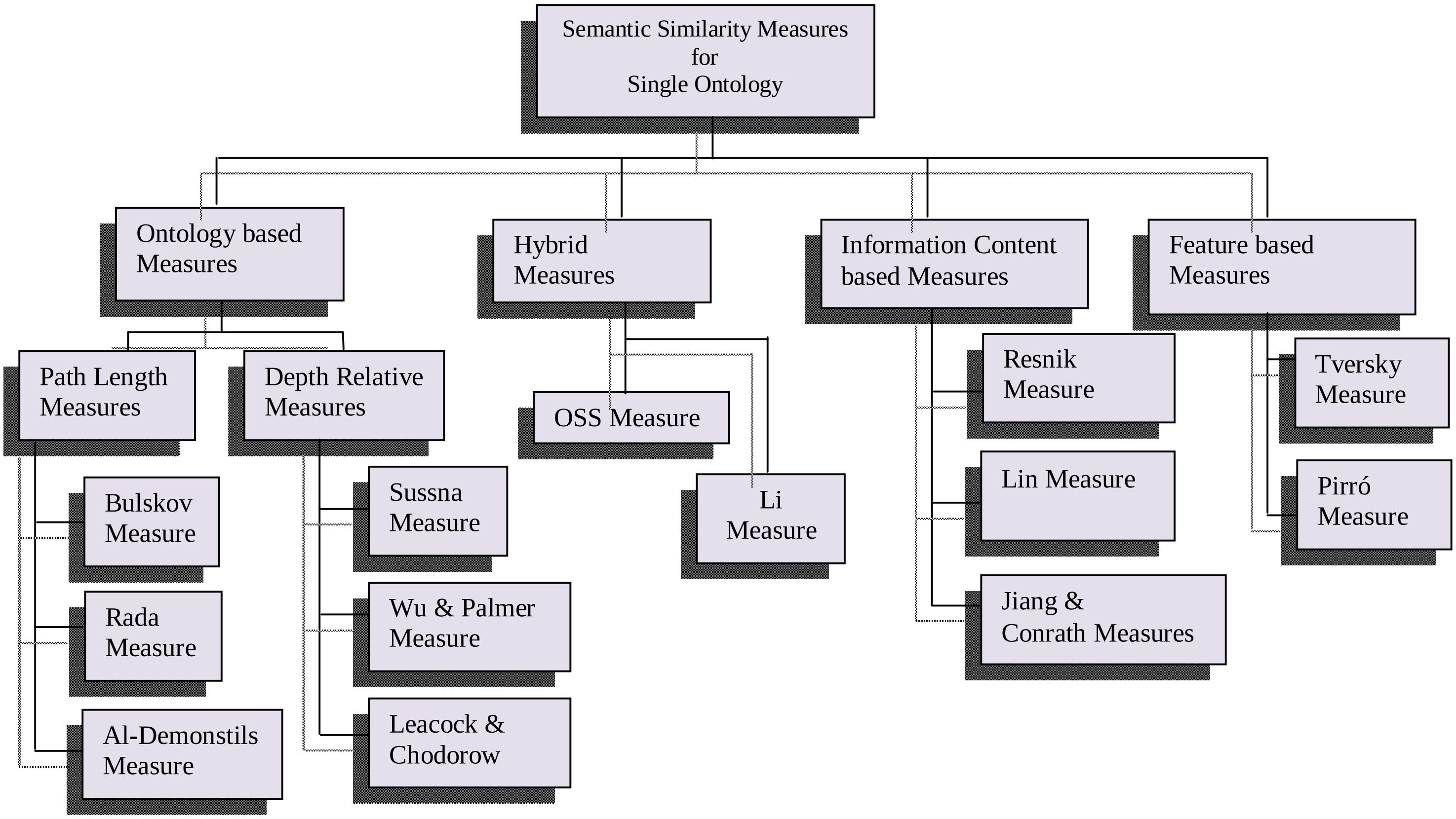


Figure 2.3 1 Classification of Semantic Similarity Measures for Single Ontology.

2.10 Semantic Similarity Measures for Single Ontology

In this work, we focus only on these semantic similarity measures that used ontology as primary information source.

2.10.1 Ontology structure –based similarity measures:

Most of these measures are based on the structure of the ontology are actually based on: path length/distance (*shortest path length*) between the two concepts nodes, and depth of concepts nodes in the ontology/*is-a* hierarchy tree. E.g. some of the measures are based on WordNet ontology includes: Path length, Wu & palmer, Leacock &Chodorow, and Li et.al [4, 12].

2.10.1.1 Path Length based Measures:

The similarity measurement among concepts is based on the path distance separating the concepts. These measures compute similarity in terms of the shortest path between the target synsets (group of synonyms) in the taxonomy.

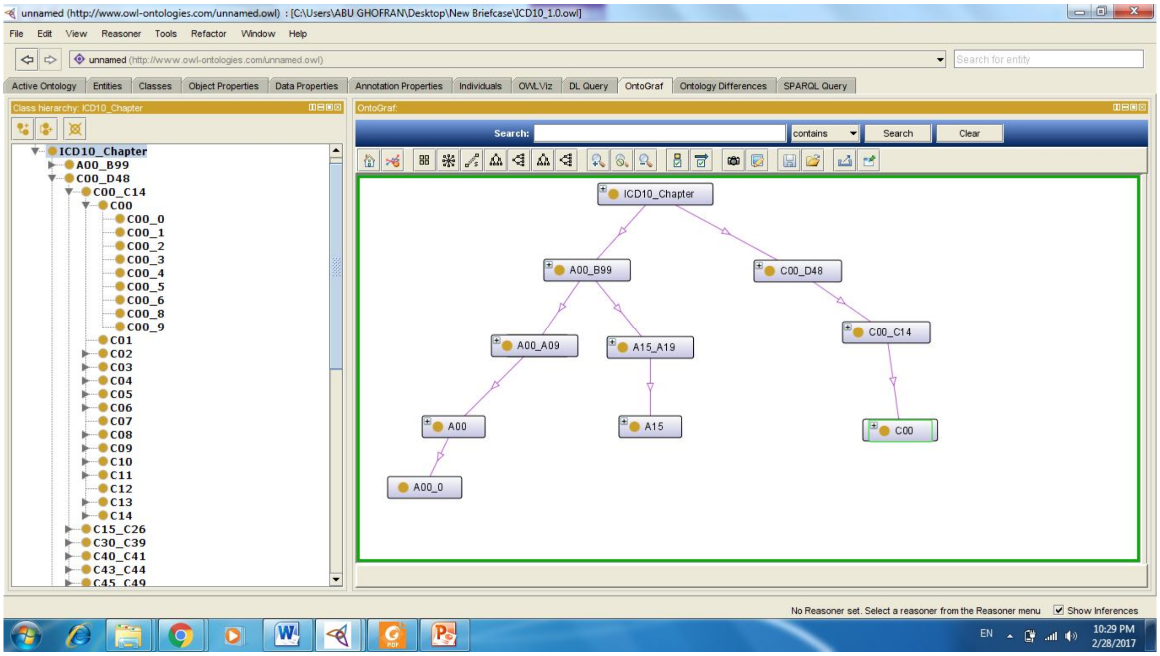
Rada measures: *[12]*In this measure the semantic distance is computed by counting the number of edges between two concepts in the taxonomy. The experiments were conducted using MeSH (Medical Subject Headings - Biomedical ontology) ontology. They are assume two concepts c1, c2 as shortest path linking them (sp(c1, c2)) as estimate distance.

distRada (c1, c2) = sp(c1, c2)   (1)

Figure.2 [2, 4, 14]: show the shortest path between two concepts a5 and b1 → → is a5

Also simple edge-counting measure proposed by ***Rada***[13]:

DisRad(c1,c2)=N_1+N_2(2)Where N_1 and N_2 are the minimum number of taxonomical links from c1 to c2to their LCS, respectively.



unnamed (http://www.owl-ontologies.com/unnamed.owl) : [C:\Users\ABU GHOFRAN\Desktop\New Briefcase\ICD10_1.0.owl]

No Reasoner set. Select a reasoner from the Reasoner menu

Figure 2.4: 1 Hierarchy tree of concepts.

2.10.1.2 Wu and Palmer Similarity Measure

[12] proposed a new method which define the semantic similarity measure between two concepts C1 and C2 as:

$$\mathrm{Sim}(c1, c2) = \frac{2N3}{N1+N2+2N3} \qquad (2)$$

Where N1 is the length given as number of nodes in the path from C1 to C3 which is the least common super concept of C1 and C2, and N2 is the length given in number of nodes on a path from C2 to C3. N3 represents the global depth of the hierarchy and it serves as the scaling factor. For example: (LCS (M08.0 ,M08.1) = M08 and LCS(M08 ,M09) = M05_M14) of two concept nodes and N1, N2 are the path lengths from each concept node to LCS, respectively.

2.10.1.3 leacok and chodorow [12] are proposed non linear adaptation of Rada's distance:

$$\mathrm{SimL}\&C = -\log\left[\frac{\mathrm{Sp}\ (c1, c2)}{2(\mathrm{Max\_depth})}\right] \qquad (3)$$

Max_depth is longest of the shortest path linking two concepts, which subsumed all others. The Least Common Ancestor (LCA) of conceptsN00_N99 and M08 is ICD10 Chapter in Figure 2.

2.10.2 Information Content-based similarity measures:

These measures use Information Content (IC) of concept nodes drive from ontology hierarchy structure and corpus statistics. Some of Information Content-based similarity measures in WordNet include: [4, 2].

2.10.2.3 Resnik Similarity Measure:

The similarity between a pair of concepts (c1 and c2) is estimated as the amount of taxonomical information they share. In a taxonomy, this information is represented by the least common subsume of both terms (LCS(c1, c2)), which is the most specific taxonomical ancestor common to c1 and c2 in a given ontology. Formally:

$$\mathrm{Sim}res = -\log(P(\mathrm{LCS}\ (c1, c2)) = \mathrm{IC}(\mathrm{LCS}\ (c1, c2)) \qquad (4)$$

Where:

$$IC(C) = \frac{\log(\text{depth}(C))}{\log(\text{deep\_max})}$$

2.10.2.3 Lin Similarity Measure

This measure depends on the relation between information content (IC) of the LCS of two concepts and the sum of the information content of the individual concepts [15, 7, 12].

Formally:

$$SimLin\ (c1, c2) = \frac{2\times \mathrm{IC}(\mathrm{LCS}\ (\mathrm{C1,C2}))}{\mathrm{IC(C1)} + \mathrm{IC(C2)}} \qquad (5)$$

2.11 SEMANTIC SIMILARITY MEASURES FOR CROSS ONTOLOGY

In this case the concepts for which similarity is to be assessed belong to two different ontologies. The secondary ontology is connected to the primary ontology through the common nodes. Two nodes in two ontologies are equivalent if they refer to the same concept.

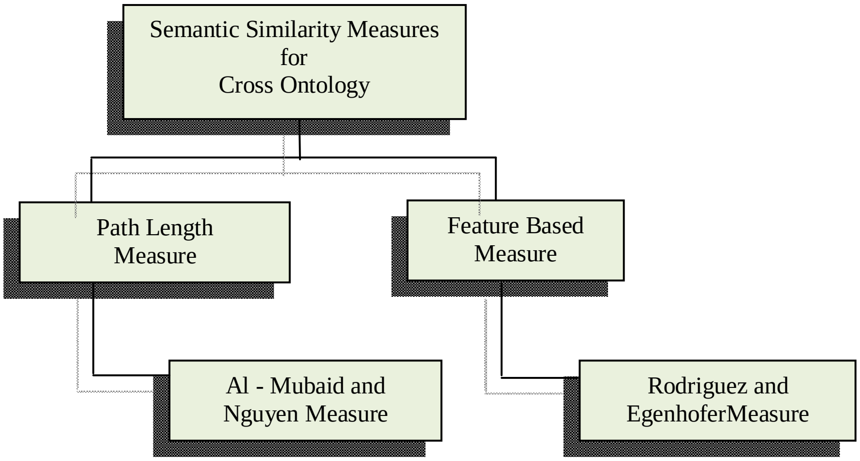

Semantic Similarity Measures for Cross Ontology
Path Length Measure
Feature Based Measure
Al - Mubaid and Nguyen Measure
Rodriguez and EgenhoferMeasure

Figure 2.5 1 Classification of Semantic Similarity Measures for Cross Ontology.

2.11.1 Al- Mubaid and Nguyen Similarity Measure

Their proposed measure is ontology-based semantic similarity measure that account for the depth of the concept nodes as well as distance (path length) between them. To compute these mantic similarity distance between two concepts, the method takes the depth of their Least Common Subsume (LCS),and the distance of shortest path of between them. The method assigns higher similarity when the two concepts are in a lower level of the hierarchy. The similarity measure is:

$$\text{Sim}(c1, c2) = \log_2 ([L(c1, c2) - 1] \times [D - \text{depth}(L(c1, c2)] + 2) \qquad (6)$$

Where:

L(c1, c2) is shortest distance between c1 and c2.

Depth L(c1, c2) is depth of L(c1, c2) using node counting.

L(c1, c2) lowest common subsume of c1 and c2.

D is maximum depth of the taxonomy.

The similarity equal 1, where two concept nodes are in the same cluster/ontology. The maximum value of this measure occur when one of the concepts is the left most leaf node, and the other concept is right leaf node in the tree. Path distance between two concepts, when two pairs of two concepts have the same path distance, they have the same value of semantic similarity. In figure2: similarity (n1, n5) = similarity (n2, n4) but (n2, n4) share more information and attributes, so they are more similar than (n1, n5). In this measure the high numeric similarity result between (c1, c2) means the lower semantic similarity between two concept. In this thesis, the term “semantic measure” is used to denote to semantic similarity measure.

In this measure they are put rules and assumptions which satisfied their proposed measure. They wont to combine all semantic features in one measure in an effective and logical way.

Rule 1: The semantic similarity scale system reflects the degree of similarity of pairs of concepts comparably in single ontology or in cross-ontology. This rule ensures that the mapping of one ontology (called secondary ontology) to another ontology (called primary ontology) does not deteriorate the similarity scale of the primary ontology. [4, 2]

Rule 2: The semantic similarity must obey local ontology’s similarity rule as follow:

Rule 2.1: The shorter the distance between two concept nodes in the ontology, the more they are similar.

Rule 2.2: Lower level pairs of concept nodes are more similar than higher level pairs.

Rule 2.3: The maximum similarity is arises when the two concept nodes are the same node in the ontology.

Assumptions:

They used logarithms (inverse of exponential for semantic distance). In rule 2.3 the semantic similarity reached higher similarity when the two concept nodes are in the same node regardless of any other features, hence, should used non linear approach to combine the features.

Non linear function is universal combination low of semantic similarity features.

New common specificity features:

Proposed by [4, 2], they used path length and depth of concept nodes to improved performance. The least common subsume (LCS) of two concepts node in the ontology is lowest node that connect pairs of concepts. It used to determine common specificity of two concept nodes in the cluster. So finding the depth of their LCS node and then scaling this depth by depth D of the cluster as follow:

$$CSpec\ (c1, c2) = \mathrm{D} - \mathrm{depth}\left(\mathrm{LCS}(\mathrm{c1}, \mathrm{c2})\right) \qquad (8)$$

Where: D is depth of the cluster. The smaller common specificity of two concept nodes, means that they are more similar and share more information.

Single cluster similarity:[2, 4] proposed their measure for single cluster:

$$SimDis(\mathrm{c1, c2}) = \log\left((\mathrm{path}-1)^{\alpha} \times (\mathrm{Cspec})^{\beta} + \mathrm{k}\right) \qquad (9)$$

Where

$\alpha > 0$ and $\beta > 0$, k constant and must be ($k >= 1$), and Cspec calculate in Eq 8.

Sem = 0 when depth = 1 regardless of (CSpec).

Cross –cluster semantic similarity:

The cluster has largest depth is main cluster (primary cluster) and all remaining cluster is secondary.

Case 1:*(Similarity within primary ontology):*

When two concept nodes in the primary ontology, in this case the similarity is calculate as similarity within single ontology using Eq (9) given before.

Case 2:*Cross-Ontology similarity (primary -secondary):The common specificity feature:* In this case, the two concepts belong to two different ontologies, and one of the two concepts belong to the primary ontology while another belong to secondary ontology, and the LCS of two concept nodes is the global root node, which belongs to the two ontologies. This technique does not affect the scale of the *CSpec* feature of the primary ontology. The common specificity is then given as:

$$CSpec\ (c1, c2)\ = \mathrm{Cspec\ primary}\ =\ \mathrm{Dprimary}\ - 1 \qquad (10)$$

where D primary is the depth of the primary ontology. The root is the LCS of the two concept nodes in this case. The path between the two concept nodes passes through two ontologies having different granularity degrees. The portion of the path length that belongs to the secondary ontology is in scale of granularity different from that of the primary ontology, and thus, we need to convert it (level it) into primary cluster scale-level as follows:

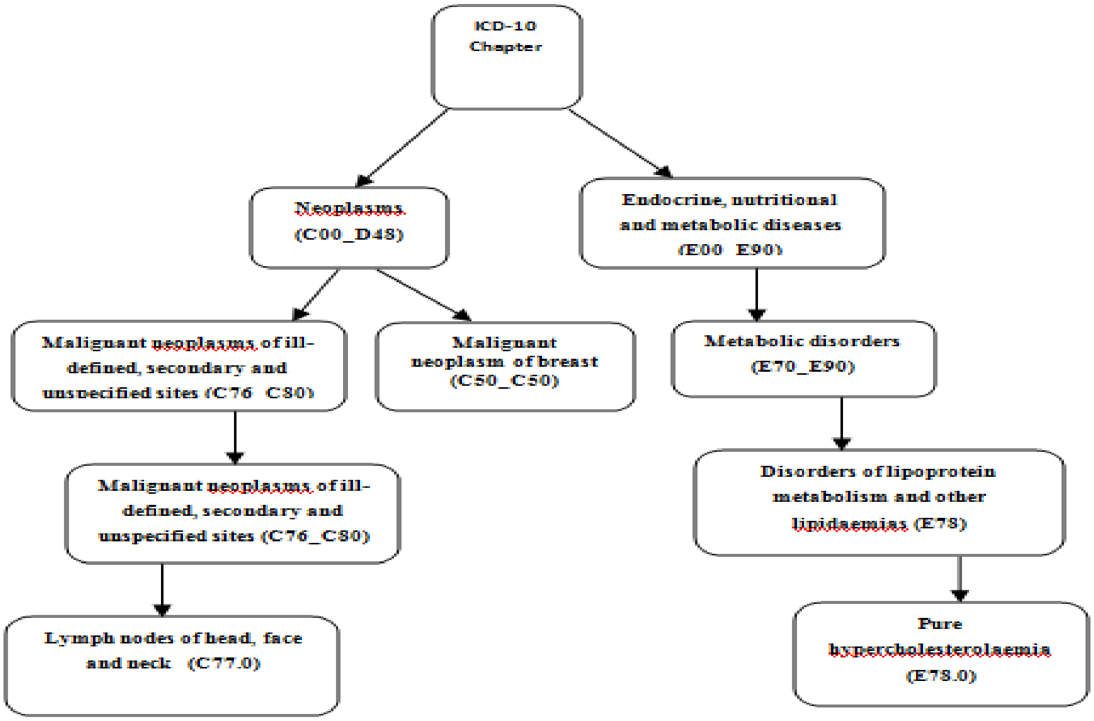

ICD-10 Chapter
Neoplasms (C00_D48)
Endocrine, nutritional and metabolic diseases (E00 E90)
Malignant neoplasms of ill-defined, secondary and unspecified sites (C76 C80)
Malignant neoplasm of breast (C50_C50)
Metabolic disorders (E70_E90)
Malignant neoplasms of ill-defined, secondary and unspecified sites (C76_C80)
Disorders of lipoprotein metabolism and other lipidaemias (E78)
Lymph nodes of head, face and neck (C77.0)
Pure hypercholesterolaemia (E78.0)

Figure 2.6: 1 A fragment of two clusters in ICD10 ontology (C77.0, E78.0).

The Cross-Cluster Path length Feature:

The path length between two concept nodes (c1, c2) is computed by adding up the two shortest path lengths from the two nodes to their LCS node (their LCS is the root(ICD10_Chapter)). For example, in Figure 2.6, for the two concept nodes (A00, C00), the LCS is the root ICD10_Chapter. So, we measure the path length between A00 and C00 as:

Path $(C_1, C2) = d_1 + d_2 - 1$ (11)

In this case: d_1 = d(A00, root) and d_2 = d(C00, root), where d(A00, root) is the path length from the root ICD10_Chapter to node A00, and similarly d(C00, root) is the path length from ICD10_Chapter to C00.Note: we subtract one in Eq.(11), because the root counted twice. The cluster containing A00 has higher depth, and then it’s the primary cluster, and the cluster containing C00is the secondary. The granularity rate of the primary cluster over the secondary cluster for the common specificity feature is:

$$CSpecRate = \frac{D_1 - 1}{D_2 - 1} \qquad (12)$$

Where: $(D_1\text{-}1)$ and $(D_2\ \text{-}1)$ are maximum common specificity values of the primary and secondary clusters respectively. The granularity rate, PathRate, of path length feature for the primary cluster over the secondary cluster is given by:

$$PathRate = \frac{2D_1 - 1}{2D_2 - 1} \qquad (13)$$

where $(2D_1\text{-}1)$ and $(2D_2\text{-}1)$ are maximum path values of any two nodes in the primary

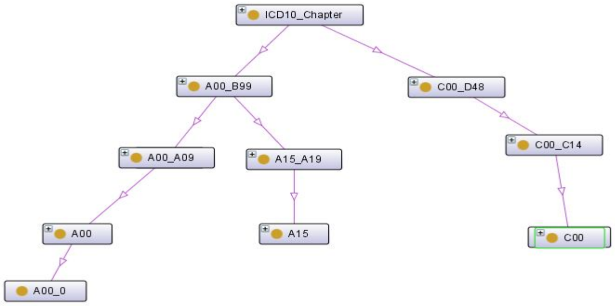


Figure 2.7: 1 fragment of concepts (A00 and C00)

and secondary ontologies respectively. Following Rule R1, we convert d2 in Eq.(11) to the primary cluster as follows:

$$d'_2 = PathRate \times d_2 (14)$$

This new path length d’2 reflects path length of the second concept to the LCS relative to primary cluster’s path length feature scale. Applying Eq.(14), we obtain path length between 2 concept nodes in primary cluster scale as follow:

$$Path(C1,C2) = d_1 + PathRate \times d_2 - 1 \ (15)$$

$$Path(C1,C2) = d_1 + \frac{2D_1 - 1}{2D_2 - 1} \times d_2 - 1 \qquad (16)$$

Finally, the semantic distance between two concept nodes is given as follow:

$$CSpec\ (C1,\ C2) = D_{primary} - 1\ (17)$$

$$Sem\ (C_1,\ C_2) = \log\ ((path\text{-}1)^{\alpha}\ \ (Cspec)^{\beta} + k)\ \ (18)$$

Case 3: Similarity within a single secondary ontology: when two concept nodes are in single secondary ontology. Then the semantic features, in this case, must be converted to primary ontology's scales for the two features, Path and CSpec, as follow:

$\mathrm{Path}(C_1, C_2) = \mathrm{Path}(C_1, C_2)_{\mathrm{secondary}} \times \mathrm{PathRate}$ (19)

$\mathrm{CSpec}(C_1, C_2) = \mathrm{CSpec}(C_1, C_2)_{\mathrm{secondary}} \times \mathrm{CSpecRate}$ (20)

$\mathrm{Sem}\ (C_1, C_2) = \log\ ((\mathrm{path}\text{-}1)^{\alpha}\ (\mathrm{Cspec})^{\beta}+ k)$ (21)

Where: $\mathrm{Path}(C_1, C_2)_{\mathrm{secondary}}$ and $\mathrm{CSpec}(C1, C2)_{\mathrm{secondary}}$ are the Path and CSpec between C_1& C_2 in the secondary ontology.

***Case 4*:** Similarity within multiple secondary ontology:

One of the secondary ontologies acts temporarily as the primary ontology to calculate Cspec and path using cross-cluster approach as in *case2* above. Then semantic distance is computed using *case3*.

Chapter Three

Reference Dataset Methodology for Semantic Web content

3.1 Overview

3.1 Introduction

This chapter presents a methodology to build a reference dataset for the ontologies. In this chapter the idea of reference dataset has been adopted from the clustering methods used in data mining explaind by: LiorRokach et. al, [70]. Farley and Raftery [71] dividing the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber [72] suggest categorizing the methods into additional three main categories: density-based methods, model-based clustering and grid based methods. The general methodology for clustering methods are illustrated in the following figure 3.1.

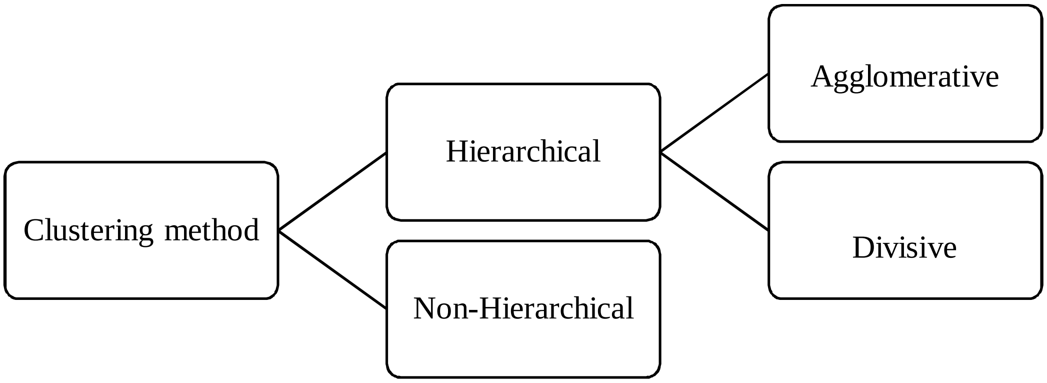


Figure 3.1: 1 Clustering method groups.

3.2 Hierarchical Methods:

In these methods, the clusters are construct by recursively partitioning the instances in either a top-down or bottom-up fashion. These methods can be sub-divided as the following:

1) **Agglomerative hierarchical clustering:** Each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained.

2) **Subdivided (Divisive) hierarchical clustering:** All objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained. The desire cluster is last cluster that contains the last class (concept) in

our experiment using *Al-Mubaid and Nguyen's measure (SemDist)* equation. We will used this method to create all clusters.

The result of the hierarchical methods is called dendrogram. Dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering [73]. This result represents nested group of objects and similarity levels among groupings. A clustering of the data objects is obtained by cutting the tree diagram (dendrogram) at the desired similarity level. We will use dendrogram to build our reference dataset.

The merging or division of clusters is performed according to some similarity measures. One way choose to optimize some criterion such as a sum of squares. The hierarchical clustering methods could be further divided according to the manner that the similarity measure is calculated [70]. There are three well-known methods used to measure the distance, which have been used extensively, namely activity distribution, metabolic clearance and equilibrium time method.

- **i.** **Single-link clustering** (Connectedness) methods that consider the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster [74].
- **ii.** **Complete-link clustering** (Diameter) methods that consider the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster [75].
- **iii.** **Average-link clustering** (Minimum variance) methods that consider the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster [76, 77].

3.3 Hierarchical Methodology Stages

A) Developing the reference dataset

The first step of our reference dataset development is to find a good source of clinical knowledge to construct the reference dataset based on them. After evaluating some known resources such as MeSH (Medical Subject Headings), SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms), OMIM (Online Mendelian Inheritance in Man), ICD10 (International Classification of Disease -10), NCI Thesaurus and UMLS. We selected the International of Classification of Diseases (ICD-10), chapter I, as the source of knowledge of diseases for our reference dataset design due to the following reasons: The ICD-10 is open source ontology for the integration of biomedical data. ICD-10 has a formally correct, semantically computable structure. Terms/concepts in ICD-10 are well.

B) Component of the clustering tasks: Typical pattern clustering activity involves the following stages:

Stage 1: Pattern representations (Choosing data source)

Stage 2: Pattern Measure (appropriate to the data domain)

Stage 3: Three: Grouping

Figure 3.3 depicts a typical sequencing of these stages, including a feedback path where the grouping process results could influence consequent feature selection or extraction and similarity calculations [78].

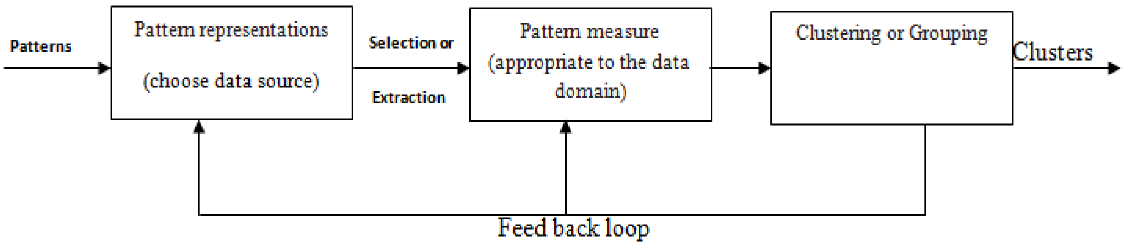


Figure 3.2 1 Stages in clustering [70].

Stage 1: Pattern representations (Choosing data source)

At this stage, we have investigated the choosing data source which refers to the number of classes (concepts). Considering the biomedical domain ICD-10 ontology version 1.0 “ICD10_1.0” [36]. ICD is stand for the International Statistical Classification of Diseases. The 22 main sub-categories of ICD-10 include, among others, diseases of the blood and blood-forming organs (D50–D89), endocrine, nutritional and metabolic diseases (E00–E90), mental and behavioral disorders (F00–F99), diseases of the nervous system (G00–G99) and certain infections and parasitic diseases (A00–B99). Sub-domain level categories Level (i), i = I,. ..,XVII, may be introduced; their instances are subsumed by the corresponding chapters. The instances of a level category level (i) in ICD-10 exhibit a taxonomic structure. Consider the domain of infections and parasitic diseases (A00–B99) and the associated domain-level category level(I). One of the classification principles is based on the pathogens that cause the disease. Hence, the concepts in level(I) have a taxonomic concept, “infectious and parasitic diseases” (diseases caused by pathogens).

Stage 2: Pattern Measure (appropriate to the data domain)

This stage follows the pattern representation stage, it highlights the feature selection is the process of identifying the most effective subset of the original features to use in clustering. Pattern Measure is usually measured by a distance function defined on pairs of patterns (concepts). The following figure 3.3 shows small portion of our taxonomy. We collect the concepts in stage (1) to get the feature in stage (2) then we build the taxonomy as shown in figure 3.3. It consists of many levels and the relationships between each level. This is important since it will be used to calculate the similarity among classes (concepts).

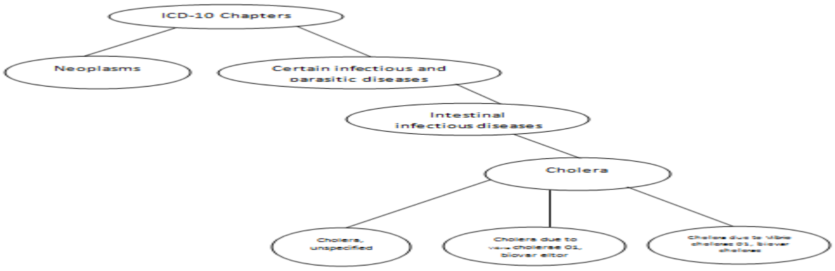

ICD-10 Chapters
Neoplasms
Certain infectious and parasitic diseases
Intestinal infectious diseases
Cholera
Cholera, unspecified
Cholera due to vibrio cholerae 01, biover eltor
Cholera due to Vibrio cholerae 01, biovar cholerae

Figure 3.3: 1 Taxonomy for cholera disease.

3.4 Semantic Similarity Measures:

Clustering is the grouping of similar concepts/classes, some sort of measure that can determine whether two concepts are similar or dissimilar using. There are two main type of measures used to estimate this relation: distance measures and similarity measures. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.

3.5 Experiments:

Our experiment was applied on the hierarchical taxonomy to determine the similarity value between two classes (concepts). The similarity between pairs of concepts of the biomedical domain has been calculated using: *Al-Mubaid and Nguyen's measure (SemDist)* equation. SemDist measure showed higher correlations with Experts scores than with Physicians scores (for more details see chapter five). For that reason it has been chosen as the best measure to be utilized in the proposed reference dataset.

Our proposed approach can be used to generate dataset from biomedical domain (ICD-10 Ontology). We used "Certain Infectious and Parasitic Diseases (A00 - B99). Because it is a largest and most widely used vocabulary resources relevant to the study of infectious diseases and conclude with a description of the Infectious Disease Ontology (IDO) suite of interoperable ontology modules that together cover the entire infectious disease domain.

3.6 Calculate Similarity between pair of concepts (classes):

The whole steps of calculate similarity between pair of concepts to create our reference dataset is shown in features, read classes list, set all the classes in one cluster, begin with one cluster (all classes together) compare the first concepts with all other concepts, split the most dissimilar classes (concepts), and repeat step two until all concepts (classes) are in their own clusters.

Here in this step we will discuss the architecture of divisive hierarchical clustering with SemDist measure. From the below architecture the implementation of divisive hierarchical clustering with SemDist measure is understood where first data taken which has objects and their measured features. First data will be read from the cluster list, where initially the whole data is taken as one big cluster which consists of all the concepts. Then the next steps are shown in Figure 3.4.

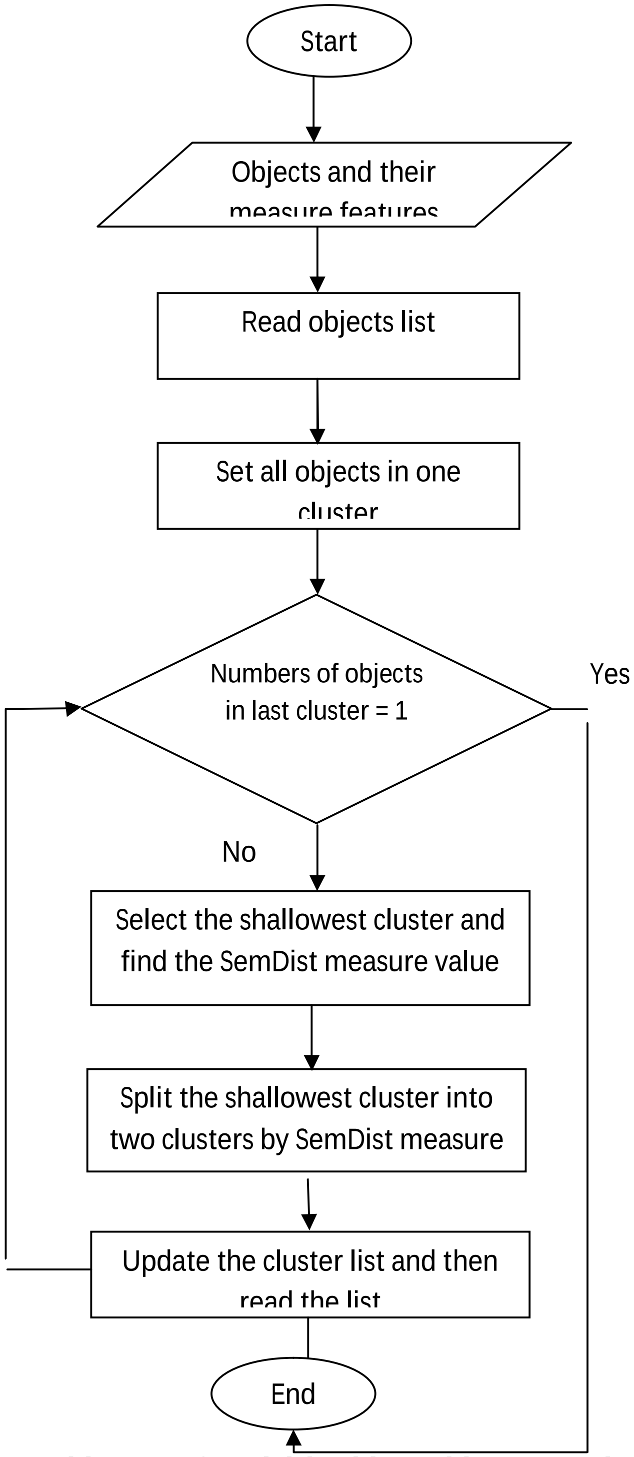

Start
Objects and their measure features
Read objects list
Set all objects in one cluster
Numbers of objects in last cluster = 1
Yes
No
Select the shallowest cluster and find the SemDist measure value
Split the shallowest cluster into two clusters by SemDist measure
Update the cluster list and then read the list
End

Figure 3.4: 1 Architecture for Divisive hierarchical clustering with SemDist-measure [79]

3.6.1 Select the shallowest cluster and find the SemDist measure value

In this experiment we considered the biomedical domain type (ICD-10 Ontology (Chapter I)) taxonomy which is being shown above in Figure3.2 as the Data source that we used in our experiment. Its contains 738 pairs of terms (leaf nodes) were chosen to compute the similarity between them by applying *Hisham Al-Mubaid & Nguyen* (SemDist) similarity measure equation.

$$\text{Sim (C1, C2)} = \log_2([\text{Path Length(C1, C2)} - 1]^{\alpha} \times [\text{CSpec(C1, C2)}]^{\beta} + k) \quad (1)$$

$$\text{CSpec}(C1, C2) = D - \text{depth}\,(\text{LCS}(C1, C2)) \quad (2)$$

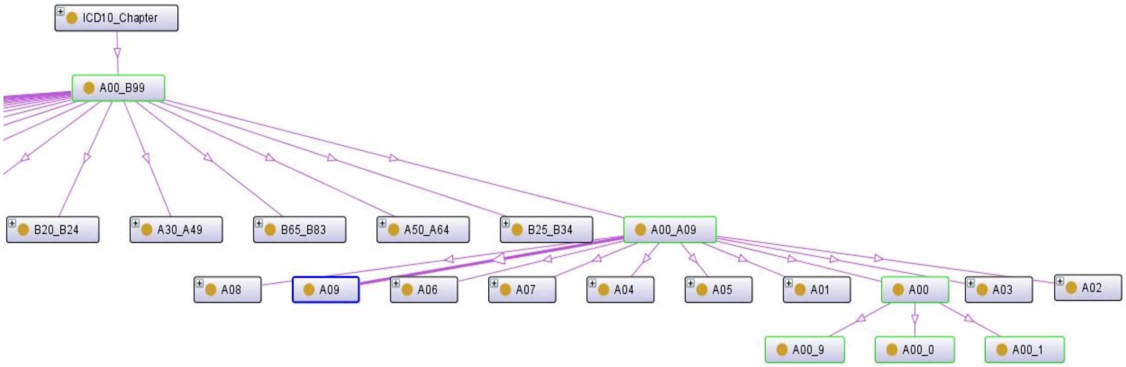


Figure 3.5 1: fragment of A00 class in ICD10 "V1.0"ontology.

For example, To compute the similarity between*"Cholera due to Vibrio cholerae 01, biovar cholera [A00.0]"* and *"Cholera due to Vibrio cholerae 01, biovareltor [A00.1]"* the shortest pathlength = 3 "using node counting"and the shortest pathlength between *"Cholera due to Vibrio cholerae 01, biovar [A00.0]"* and *"Cholera, unspecified [A00.9]"* is also 3 (from figure3.2). The depth of Least Common Subsume (LCS) is:

LCS (A00.0, A00.0)   = A00.0

CSpec(A00.0, A00.0)   = D – depth (LCS (A00.0))

= 5 – 5 = 0

So, similarity:

$$\text{SemDist (A00.0, A00.0)} = \log_2([1 - 1]^1 \times [0]^{1} + 1) = \log_2(1) = 0$$

The higher similarity arises when the two concepts are in the lower level of the hierarchy. Classes that are more similar with have a lower similarity score than classes that are less similar with this measure. Table 3.1 shows the similarity score.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec(c1, c2)	SemDist(c1, c2)
1	A00.0	A00.0	A00.0	1	0	0
2	A00.0	A00.1	A00	3	1	1.6
3	A00.0	A00.9	A00	3	1	1.6
4	A00.0	A01.0	A00_A09	5	2	3.2
. . 58	A00.0	A08.5	A00_A09	5	2	. . 3.2

Table 3.1 1: The similarity between the concepts (classes) using SemDist (Cluster One).

[A00.0] = *Cholera due to Vibrio cholerae 01, biovar.*

[A00.1] = *Cholera due to Vibrio cholerae 01, biovareltor.*

[A00.9] =*Cholera, unspecified.*

[A01.0] =*Typhoid fever.*

[A08.5] =*Other specified intestinal infections.*

[A00] =*Cholera.*

[A00_A09] =*Intestinal infectious diseases.*

3.6.2 Split the shallowest cluster into two clusters by SemDist measure

Hierarchical clustering algorithms produce a nested series of partitions based on a criterion for merging or splitting clusters based on similarity. Table3.1, show the similar concepts which have small value of SemDist measure. The smaller the common specificity value of two concept nodes, the more they share information, and thus the more they are similar.

3.6.3 Update the cluster list and then read the list

Repeat step two until all concepts (classes) are in their own clusters. For example, we select the last concept (class) “B95.0 = *Streptococcus, group A, as the cause of diseases classified to other chapters*” as leaf node and compare it with all other remaining leaf nodes (26 concepts) in “*chapter I*” using *Al-Mubaid and Nguyen’s measure* (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create the last cluster (cluster twenty). As shown in Table 3.2. For more details information of clusters, please refer to [Appendix C].

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B95.0	B95.0	B95.0	1	0	0
2	B95.0	B95.1	B95	3	1	1.6
3	B95.0	B95.2	B95	3	1	1.6
. . . 26	B95.0	B97.7	B95_B97	5	2	3.2

Table 3.2 1: Compare between concepts (classes) using SemDist (Cluster Twenty).

Stage 3: Grouping

Data Summary:

Cluster Name	Number of concepts in each cluster	Number of concepts in desired cluster	Cluster Name	Number of Concepts in each Cluster	Number of Concepts in desired Cluster
Cluster #1	58	3	Cluster #11	38	9
Cluster #2	37	10	Cluster #12	17	2
Cluster #3	46	7	Cluster #13	25	10
Cluster #4	75	8	Cluster #14	33	5
Cluster #5	48	9	Cluster #15	92	9
Cluster #6	23	10	Cluster #16	35	3
Cluster #7	6	3	Cluster #17	71	6
Cluster #8	15	5	Cluster #18	18	5
Cluster #9	39	6	Cluster #19	10	5
Cluster #10	26	7	Cluster #20	26	9
			Overall		131

Table 3.3 1 provides summaries of concepts in each cluster and in desired cluster.

Table 3.3: shows the summaries of classes/concepts in each cluster and in the desired cluster. From all clusters, we select the minimum similarity value scores, and collect them in one group and call them (*Infectious and Parasitic DO-* Reference dataset). As shown in Table 3.4. For more details information of clusters, please refer to [Appendix A].

ID	Concept1(Class)	ICD-10 Code	Concept2 (Class)	ICD-10 Code	Sem Dist
1	Cholera due to Vibrio cholerae 01, biovar cholera	A00.0	Cholera due to Vibrio cholerae 01, biovar cholera	A00.0	0
2	Cholera due to Vibrio cholerae 01, biovar cholera	A00.0	Cholera due to Vibrio cholerae 01, biovareltor	A00.1	1.6
3	Cholera due to Vibrio cholerae 01, biovar cholera	A00.0	Cholera, unspecified	A00.9	1.6
4	Tuberculosis of lung, confirmed by sputum microscopy with or without culture	A15.0	Tuberculosis of lung, confirmed by sputum microscopy with or without culture	A15.0	0
5	Tuberculosis of lung, confirmed by sputum microscopy with or without culture	A15.0	Tuberculosis of lung, confirmed by culture only	A15.1	1.6
131	. . . . Streptococcus, group A, as the cause of diseases classified to other chapters	. . . B95.0	Unspecified staphylococcus as the cause of diseases classified to other chapters	B95.8	1.6

Table 3.4 1 Infectious and Parasitic DO- Reference dataset concepts (classes).

Chapter Four
Design & Implementation
Infectious and Parasitic DO-Refernce Dataset Development
4.1 Overview

From the hypothesis and the literature review have been established, an appropriate design is required to meet the requirements of the research. It is important to build a reference dataset which used existing resources (*Certain infectious and parasitic diseases* (Chapter I) in ICD-10 Ontology). In this chapter we identify and discuss the seven steps to develop the biomedical domain reference dataset (*Infectious and Parasitic DO*-Reference Dataset) to be used as a basis to evaluate the ontology in the biomedical domain by comparing them to our reference dataset. We present specific steps on developing the reference dataset and applying it in a specific development environment namely, protégé. We evaluate our approach using the *Infectious and Parasitic DO-Reference Dataset.*

4.2 Design the reference dataset

In order to evaluate the biomedical or health domain ontology, we need a reference dataset or standard definition. According to Hisham Al-Mubaid & Hoa A. Nguyen [11, 12] as far as I know, there is no standard approach to evaluate the quality of ontology from the perspective of semantic similarity measure, and there is no well define of a reference dataset or standard definition in the biomedical domain. For these reasons, we had to acquire a new reference dataset. There are many different tools available for developing ontology such as Top Braid Composer, OBO-Edit, and Protégé etc. We use Protégé which is one of the most widely used in biomedical ontology development editor that defines ontology concepts (classes), properties, taxonomies, various restrictions and class instances. We present methodology for a build our reference dataset as ontology (adapted from Noy and McGuiness (2001)). According to this method developing a reference dataset include:

- Defining classes in the dataset or reference dataset.
- Arranging the classes in a taxonomic (subclass-super class) hierarchy.
- Defining slots and describing allowed values for these slots.
- Filling in the values for slots for instances.

However, there is no single way to correctly model a domain. The ontology development process is not linear [81]. This process can also be responsible for modulate new information and modifications into the reference dataset, which requires returning to previous stages. Building the reference dataset (dataset) consists of the following steps:

Phase One: Determine the domain and scope of the ontology:

Defining ontology domain and scope requires answering the following questions:

1. *What is the domain that the ontology will cover?*

Our domain of the ontology is ***certain infectious and parasitic diseases.***

2. *What is the use of the ontology?*

The ontology is to provide a knowledge base of diseases. This ontology (*certain infectious and parasitic diseases)* will be used as reference dataset (dataset) to compare it with other ontology in the biomedical domain.

3. *What types of questions should the information in the ontology provide answers?*
4. *Who will use and maintain the ontology?*

This ontology used by physician, experts and specialist in the biomedical domain. The users who are interested in creating and developing biomedical ontologies. This version of dataset is Beta version, and its open source for modifying. The developed dataset can be reused in the future for other purposes.

Phase Two: Consider reusing existing ontologies:

There is no single standard way to develop ontology. It is not necessary to start from scratch always. We use ICD-10 Ontology[36] as a basis for developing the *Infectious and Parasitic DO-* Reference dataset.

Phase Three: Overview of Infectious and Parasitic DO- Reference dataset:

We identify some diseases, *"Intestinal infectious diseases"*, *"Tuberculosis"*, …… and *"Bacterial, viral and other infectious agents"*. These concepts represent biomedical domain (ICD-10 version 1.0) types taken from the our taxonomy described in chapter five. Then the classes in Table 4.1 are top level concepts (classes) in our reference dataset (dataset).

ID	***Class Name***	***Code in ICD-10***	***Note***
1	*Intestinal infectious diseases*	A00_A09	
2	*Tuberculosis*	A15_A19	
3	*Certain zoonotic bacterial diseases*	A20_A28	
4	*Other bacterial diseases*	A30_A49	
5	*Infections with a predominantly sexual mode of transmission*	A50_A64	
6	*Other spirochaetal diseases*	A65_A69	
7	*Other diseases caused by chlamydiae*	A70_A74	
8	*Rickettsioses*	A75_A79	
9	*Viral diseases of the central nervous system*	A80_A89	
10	*Arthropod-borne viral fevers and viral haemorrhagic fevers*	A90_A99	
11	*Viral infections characterised by skin and mucous membrane lesions*	B00_B09	
12	*Viral hepatitis*	B15_B19	
13	*Human immunodeficiency virus [HIV] disease*	B20_B24	
14	*Other viral diseases*	B25_B34	
15	*Mycoses*	B35_B49	
16	*Protozoal diseases*	B50_B64	
17	*Helminthiases,*	B65_B83	
18	*Pediculosis, acariasis and other infestations,*	B85_B89	
19	*Sequelae of infectious and parasitic diseases,*	B90_B94	
20	*Bacterial, viral and other infectious agents*	B95_B97	

Table 4.1 1: Top level concepts (classes)

We name our reference dataset dataset *Infectious and Parasitic DO*- Reference dataset as a short name for certain infectious and parasitic diseases Ontology. Figure 4.1 highlights the main classes of the *Infectious and Parasitic DO-Reference dataset*, as well as relationships among them. It has 174 classes, 15 object properties.

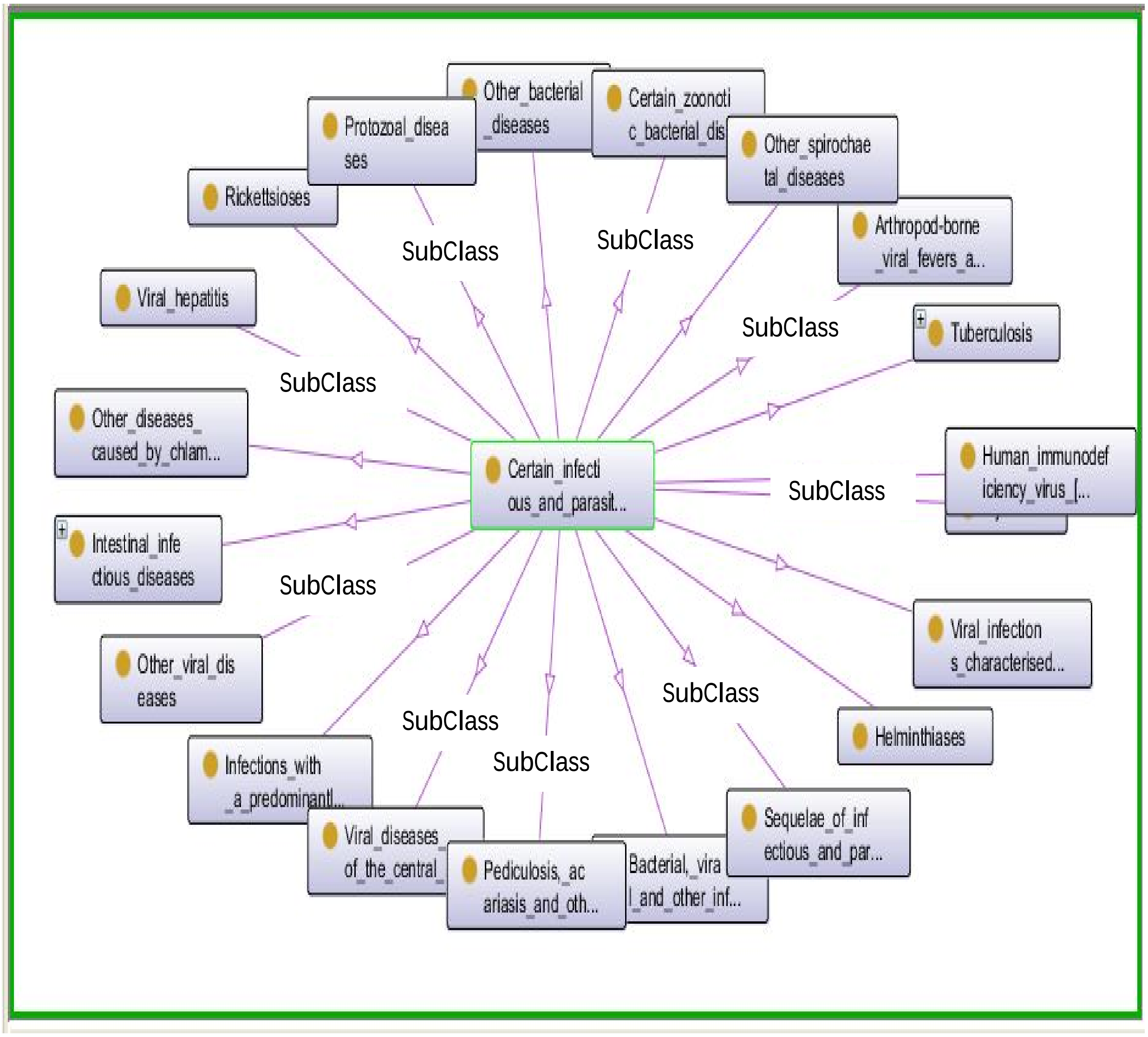

Other_bacterial_diseases
Certain_zoonotic_bacterial_dis
Protozoal_diseases
Rickettsioses
Other_spirochaetal_diseases
Arthropod-borne_viral_fevers_a...
Viral_hepatitis
Tuberculosis
Other_diseases_caused_by_chlam...
Certain_infectious_and_parasit...
Human_immunodeficiency_virus_[...
Intestinal_infectious_diseases
Other_viral_diseases
Viral_infections_characterised...
Helminthiases
Infections_with_a_predominantl...
Viral_diseases_of_the_central_
Pediculosis,_acariasis_and_oth...
Bacterial,_viral_and_other_inf...
Sequelae_of_infectious_and_par...
SubClass
SubClass
SubClass
SubClass
SubClass
SubClass
SubClass
SubClass
SubClass
SubClass

Figure 4.1 1 Main classes of Infectious and Parasitic DO-Bench

Phase Four: Enumerate important terms in the Infectious and Parasitic DO-Bench:

The symptom term is an important terms in our reference dataset dataset, the relationship of disease taxonomy pattern to the diagnosis of medical data is created by the symptom terms. The main symptom term for classes from A00-B99 are: *Chronic vapor, chest pain, low fever, panting, tiredness, headache, high fever, vapor, prickly heat, muscle, cephalitis, aches* *and pains, fever, itch, ache, irritated, fidgeting, shock, excitable.* The information of symptom terms is taken from a number of relevant research papers and documentations of Domain Ontology Health Informatics Classification (DOHIC) diseases domain [39].

Phase Five: Define classes and the class hierarchy of *Infectious DO-Bench*

The classes in Table 4.2 are sub classes in our reference dataset (dataset) and organized into a hierarchical taxonomy, also represent an infectious and parasitic disease *"chapter I"* taken from Table 3.3 described previously in chapter Three. This Phase starts by defining 131 classes.

ID	Class Name (Concept)	ICD-10Codes	Notes
1	*Cholera due to Vibrio cholerae 01, biovarcholera*	*A00.0*	
2	*Cholera due to Vibrio cholerae 01, biovareltor*	*A00.1*	
3	*Cholera, unspecified*	*A00.9*	
4	*Tuberculosis of lung, confirmed by sputum microscopy with or without culture*	*A15.0*	
5	*Tuberculosis of lung, confirmed by culture only*	*A15.1*	
. . *130*	*Other staphylococcus as the cause of diseases classified to other chapters*	*B95.7*	
131	*Unspecified staphylococcus as the cause of diseases classified to other chapters*	*B95.8*	

Table 4.2 1: Low level Infectious and Parasitic DO-Bench sub classes, leaf nodes.

There are at three common approaches in building class hierarchy (Uschold & Gruninger, 1996): top-down approach, bottom-up approach and mixed approach. In this thesis we use mixed approach to define our classes and classes' hierarchy. Our hierarchy is taxonomic hierarchy, and we follow the built in semantics of primitives such as *owl:subClassOf* and *rdfs:subClassOf*. In our approach, the ontology provides a broad conceptual structure consisting in the top of twenty portions illustrated in the following figures: figure 4.2. (more details show Appendix A). Although *Infectious and Parasitic DO-Bench*, is containing about 174 concepts, is still under development, *it's* connect the major categories used in reference dataset (Intestinal infectious diseases, Tuberculosis, Certain zoonotic bacterial diseases, Other bacterial diseases, Infections with a predominantly sexual mode of transmission, Other spirochaetal diseases, Other diseases caused by chlamydiae, Rickettsioses, Viral diseases of the central nervous system, Arthropod-borne viral fevers and viral haemorrhagic fevers, etc.). Then we generate all other classes that could expand from the top level concepts (classes).

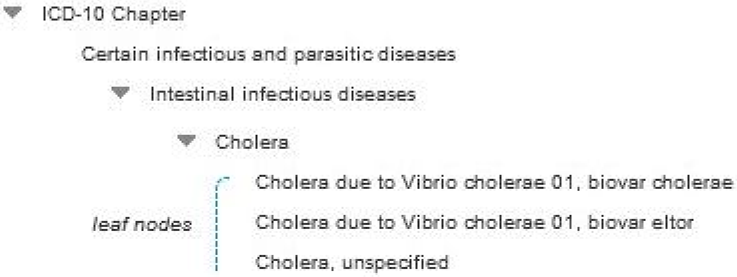

ICD-10 Chapter
Certain infectious and parasitic diseases
Intestinal infectious diseases
Cholera
Cholera due to Vibrio cholerae 01, biovar cholerae
leaf nodes
Cholera due to Vibrio cholerae 01, biovar eltor
Cholera, unspecified

Figure 4.2: 1 Class hierarchy for ClassA00.0 "Cholera due to Vibrio cholerae 01, biovarcholerae" and class synonymous.

Phase Six: Define properties of classes (or Slots):

Properties define the relationships between two objects. There are two types of properties. Object properties and data properties. Object properties are used to link object to objects. Data Properties are used to link objects to xml schema data type. Once we defined the classes, we clarify and reflect the internal structure of concepts. This is considered as the property of the developed classes. These properties are extracted from classes that are illustrated

ID	Top Object Properties	Domain	Range	Characteristics
1	Complicated_by			
2	Composed_of			
3	Drives_from			
4	Has_material_basis_in			
5	Has_symptom			
6	In_heres_in			
7	Is_a			
8	Located_in			
9	Occurs_with			
10	Pasrt_of			
11	Realized_by			
12	Realized_by_suppression_with			
13	Result_in			
14	Result_in_formation_of			
15	Transmitted_by			

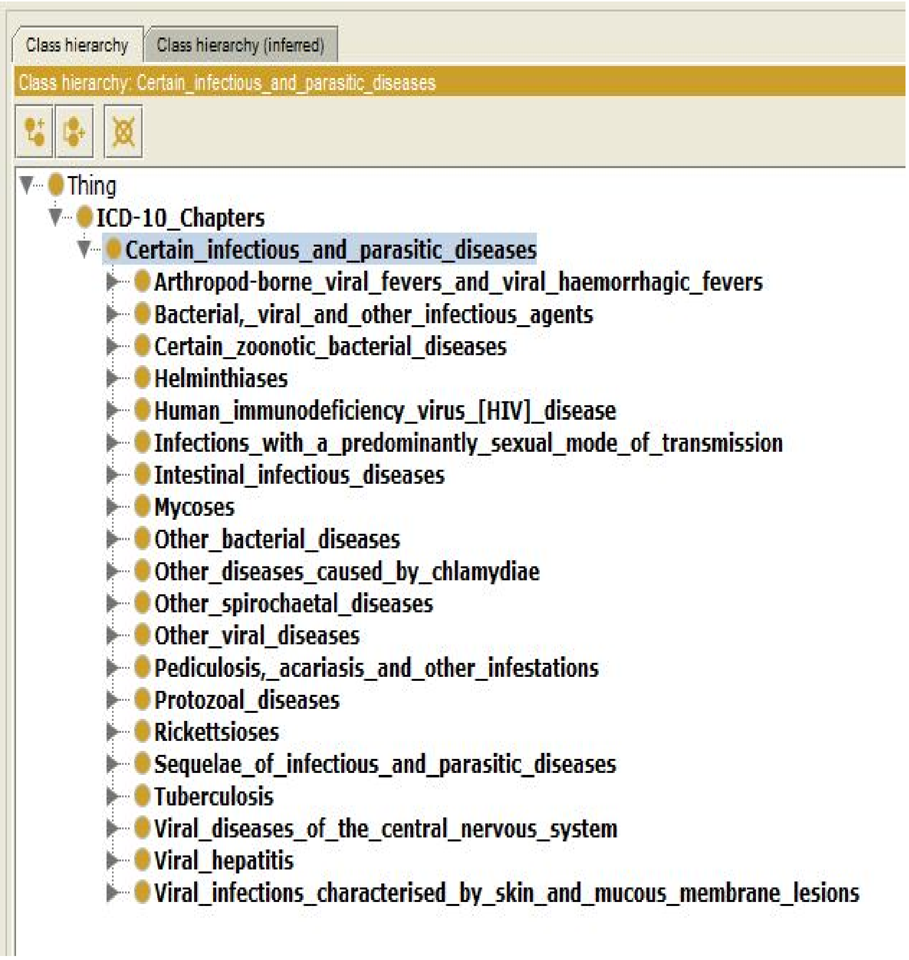
Table 4.3 1 Object Properties

4.3 Infectious and Parasitic DO-Bench implementation in Protégé:

Protégé is an ontology editor application developed by the Stanford Medical Informatics at Stanford University School of Medicine. It is a free, open source ontology editor and knowledge-based framework. It is based on Java, is extensible, and provides a foundation for customized knowledge-based applications. Protégé supports Frames, XML Schema, RDF(S) and OWL. It provides a plug-and-play environment that makes it a flexible base for rapid prototyping and application development [34, 56]. This section describes the development of Infectious and Parasitic DO-Bench in protégé as OWL Ontology.

4.3.1 Classes and Subclasses:

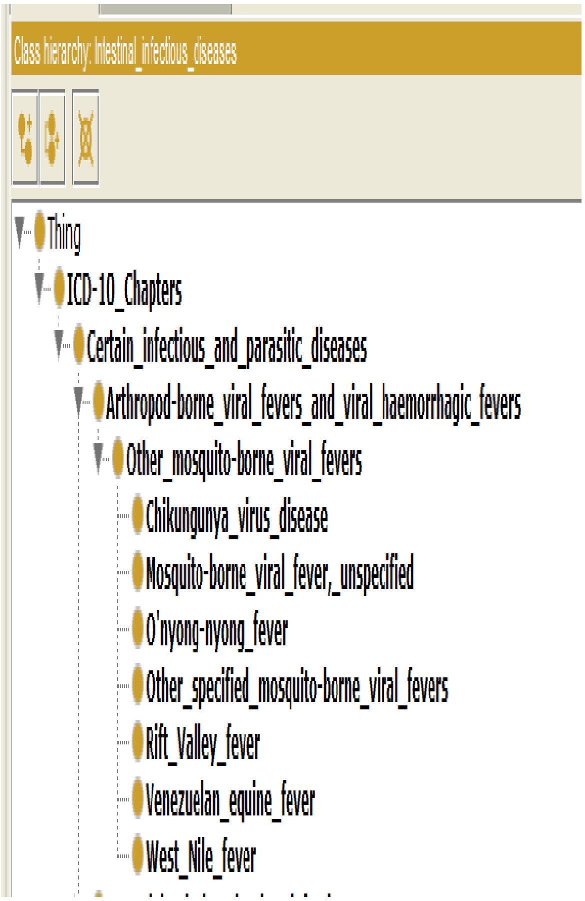
Classes are the core of ontology, which describes the concepts in some domain. In the *Infectious and Parasitic DO-Bench*, Intestinal infectious diseases, Tuberculosis, Certain zoonotic bacterial diseases, Other bacterial diseases, Infections with a predominantly sexual mode of transmission, Other spirochaetal diseases, Other diseases caused by chlamydiae, Rickettsioses, Viral diseases of the central nervous system, Arthropod-borne viral fevers and viral haemorrhagic fevers, Viral infections characterised by skin and mucous membrane lesions, Viral hepatitis, Human immunodeficiency virus [HIV] disease, Other viral diseases, Mycoses, Protozoal diseases, Helminthiases, Pediculosis, acariasis and other infestations, Sequelae of infectious and parasitic diseases, Bacterial, viral and other infectious agents are the subclasses of Certain infectious and parasitic diseases.


Class hierarchy
Class hierarchy (inferred)
Class hierarchy: Certain_infectious_and_parasitic_diseases

```
▼ Thing
  ▼ ICD-10_Chapters
    ▼ Certain_infectious_and_parasitic_diseases
      ► Arthropod-borne_viral_fevers_and_viral_haemorrhagic_fevers
      ► Bacterial,_viral_and_other_infectious_agents
      ► Certain_zoonotic_bacterial_diseases
      ► Helminthiases
      ► Human_immunodeficiency_virus_[HIV]_disease
      ► Infections_with_a_predominantly_sexual_mode_of_transmission
      ► Intestinal_infectious_diseases
      ► Mycoses
      ► Other_bacterial_diseases
      ► Other_diseases_caused_by_chlamydiae
      ► Other_spirochaetal_diseases
      ► Other_viral_diseases
      ► Pediculosis,_acariasis_and_other_infestations
      ► Protozoal_diseases
      ► Rickettsioses
      ► Sequelae_of_infectious_and_parasitic_diseases
      ► Tuberculosis
      ► Viral_diseases_of_the_central_nervous_system
      ► Viral_hepatitis
      ► Viral_infections_characterised_by_skin_and_mucous_membrane_lesions
```

Figure 4.3: 1 Top level Infectious and Parasitic DO-Bench taxonomy.

A class can have subclasses which represent the middle level Taxonomy. Figure 4.3 shows a taxonomy of *cholera disease.* It has subclasses such as [*Cholera due to Vibrio cholerae 01, biovar cholera*], [*Cholera due to Vibrio cholerae 01, biovareltor]* and [*Cholera, unspecified*].



Class hierarchy: Intestinal_infectious_diseases

- Thing
  - ICD-10_Chapters
    - Certain_infectious_and_parasitic_diseases
      - Arthropod-borne_viral_fevers_and_viral_haemorrhagic_fevers
        - Other_mosquito-borne_viral_fevers
          - Chikungunya_virus_disease
          - Mosquito-borne_viral_fever,_unspecified
          - O'nyong-nyong_fever
          - Other_specified_mosquito-borne_viral_fevers
          - Rift_Valley_fever
          - Venezuelan_equine_fever
          - West_Nile_fever

Figure 4.4 1 Middle level Infectious and Parasitic DO-Bench taxonomy.

Chapter Five

Semantic Similarity Measures

5.1 Introduction

In this chapter, We will give a high level overview of the evaluation of semantic similarity measures to determine the best one that is suitable to be used in our reference dataset model. We evaluated the applicability of using six different semantic similarity measures, these measures are Path length based measure (Shortest path length), Wu and Palmer Measure, Leacock and Chodorow similarity measure, information content-based similarity measure (Resnik’s Measures, Lin’s Measure, Lin’sMeasure) and semantic similarity measure in the biomedical domain (Al-Mubaid and Nguyen’s measure (SemDist)) equation. Several experiments have been conducted by deploying the six different semantic similarity measures to calculate the similarity between 30 pairs of (classes) concepts. These classes represent biomedical domain taken from the taxonomy describe in Figure 5.1. Then the same 30 concepts are evaluated by human expert, results have been compared, in order to determine the best measure to be used in build of our reference dataset model to evaluate ontology in biomedical domain.

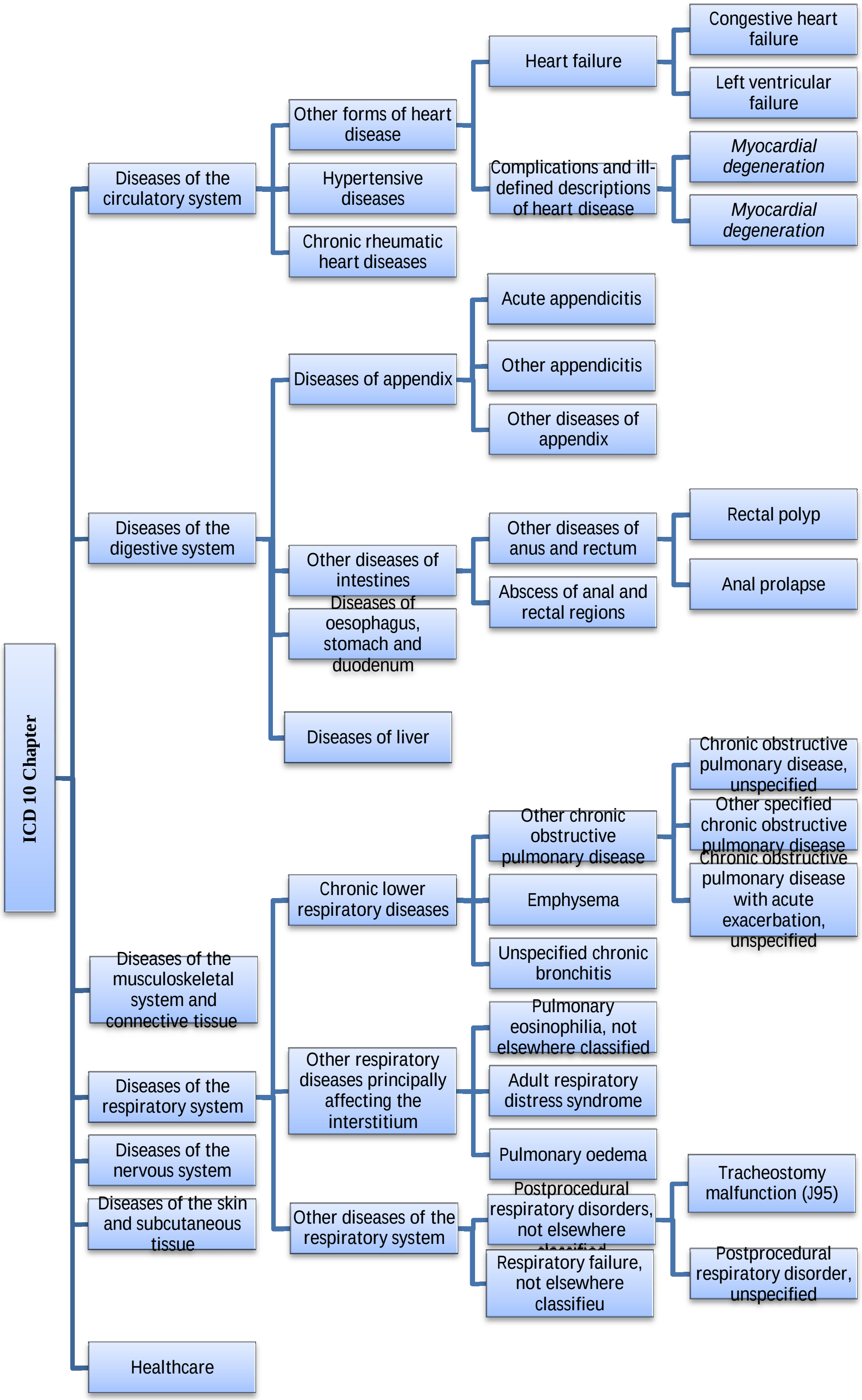

ICD 10 Chapter
Diseases of the circulatory system
Other forms of heart disease
Hypertensive diseases
Chronic rheumatic heart diseases
Heart failure
Complications and ill-defined descriptions of heart disease
Congestive heart failure
Left ventricular failure
Myocardial degeneration
Myocardial degeneration
Diseases of the digestive system
Diseases of appendix
Acute appendicitis
Other appendicitis
Other diseases of appendix
Other diseases of intestines
Diseases of oesophagus, stomach and duodenum
Other diseases of anus and rectum
Abscess of anal and rectal regions
Rectal polyp
Anal prolapse
Diseases of liver
Diseases of the musculoskeletal system and connective tissue
Diseases of the respiratory system
Diseases of the nervous system
Diseases of the skin and subcutaneous tissue
Healthcare
Chronic lower respiratory diseases
Other chronic obstructive pulmonary disease
Emphysema
Unspecified chronic bronchitis
Chronic obstructive pulmonary disease, unspecified
Other specified chronic obstructive pulmonary disease
Chronic obstructive pulmonary disease with acute exacerbation, unspecified
Other respiratory diseases principally affecting the interstitium
Pulmonary eosinophilia, not elsewhere classified
Adult respiratory distress syndrome
Pulmonary oedema
Other diseases of the respiratory system
Postprocedural respiratory disorders, not elsewhere
Respiratory failure, not elsewhere classifieu
Tracheostomy malfunction (J95)
Postprocedural respiratory disorder, unspecified

Figure 5.1 1: Fragment of the ICD-10 “V1.0” taxonomy

5.2 Semantic Similarity Measures:

Semantic similarity techniques are becoming essential components of most of the information retrieval (IR), information extraction (IE), and other intelligent knowledge-based systems. For example, in IR, similarity measures play a crucial role in determining an optimal match between query terms and the retrieved document in ranking the results such as plagiarism detection [19].

5.3 Ontology-Based Semantic Similarity Measures:

Ontology-based semantic similarity measures are those use ontology source as the primary information source. They are can be roughly grouped into two groups as follows:

5.3.1 Ontology Structure-Based Measure:

In this method, the similarity measurement among concepts is determined according to the path distance, which separates the concepts on the taxonomy or ontology structure and it includes the following types:

5.3.1.1 The Path length based measure (Shortest path length):

The similarity measurement among concepts is based on the path distance separating the concepts. These measures compute the similarity in terms of the shortest path between two concepts (classes) (group of synonyms) in the taxonomy. Rada et al, [64] proposed their measure as potential measure in the biomedical domain. Their experiments were conducted using MeSH (Medical Subject Headings) biomedical ontology. In this measure the similarities between two concepts C1 and C2 can be calculated as follows [64]:

$$\text{DistRada}\,(c1, c2) = \text{Sp}(c1, c2) \qquad (1)$$

$$\text{Shortest Path}(C1,\ C2) \quad = 2 * \text{Max}_{\text{depth}} - \text{length}(c1, c2) \qquad (2)$$

Where:

SP is Shortest Path

$\text{Max}_{\text{depth}}$ is the maximum depth of our taxonomy.

Length (c1, c2) is the shortest path length between C1 and C2.

For example, to compute the similarity between *“Hypertensive renal disease with renal failure”* (I12.0) and *“Hypertensive renal disease with renal failure”* (I12.0) the shortest path length between them equal 1 “Using node counting”

Max_Depth of our Taxonomy = 5

So:

Sim (*Hypertensive renal disease with renal failure*, *Hypertensive renal disease with renal failure*) = 2*5 – 0 = 10 = 100%

.

.

Sim (*Pure hypercholesterolaemia*, *Lymph nodes of head, face and neck*) = 2*5 – 2 = 8= 20%

Id	Concept1	Concept2	LCA(c1 c2)	Length	Similarity
4	Hypertensive renal disease with renal failure	Hypertensive renal disease with renal failure	Hypertensive renal disease with renal failure	0	100%
11	Congestive heart failure	Left ventricular failure	Heart failure	2	80%
8	Lymph nodes of head, face and neck	Major salivary gland, unspecified	Neoplasms	6	40%
17	Calcification and ossification of muscle	Stenosis and insufficiency of lacrimal passages	ICD10_Chapter	7	30%
19	Mitral stenosis	Atrial fibrillation and flutter	Diseases of the circulatory system	5	50%
. . . . 30	Pure hypercholesterolaemia	Lymph nodes of head, face and neck	ICD10_Chapter	8	20%

Table 5.1 1: Similarity values for two concepts from our taxonomy (Figure 5.1) using Path Length Based Measures (shortest path).

5.3.1.2 Wu and Palmer Measure: the measure of Wu and Palmer [80] measures semantic similarity of concepts by taking into account the depths of concept nodes only. The formula of Wu and Palmer measure is rewritten as follows:

$$\mathrm{Sim}\,(C1, C2) = 2 * \mathrm{depth}(\mathrm{LCS}(C1, C2)) \;/\; \big(\mathrm{depth}(C1) + \mathrm{depth}(C2)\big) \quad (3)$$

Or:

$$\text{Sim }(c1, c2) = \frac{2N}{N1+N2+2N} \quad (4)$$

Where: N is the depth of the least common subsumer (The least common subsumer, LCS(C_1,C_2), of two concept nodes C1 and C2 is the lowest node that can be a parent for C1 and C2.

The score can never be 0 because the depth of the LCS is never 0 (the depth of the root is 1) So the score is 0<score<=1. When the two classes are the same the score is 1.

From our taxonomy (figure5.1), We can calculate the similarity between classes C_1 and C_2 as shown in Table 5.2:

Similarity (*Hypertensive renal disease with renal failure*, *Hypertensive renal disease with renal failure*) = $\frac{2*5}{0+0+(2*5)}$ = 1 = 100%

.

.

Similarity (*Pure hypercholesterolaemia*, *Lymph nodes of head, face and neck*) = $\frac{2*1}{4+4+(2*1)}$ = 0.2 = 20%

Id	Concept1	Concept2	LCS(c1 c2)	Wu & Palmer	Similarity
4	Hypertensive renal disease with renal failure	Hypertensive renal disease with renal failure	Hypertensive renal disease with renal failure	1.00	100%
11	Congestive heart failure	Left ventricular failure	Heart failure	0.80	80%
8	Lymph nodes of head, face and neck	Major salivary gland, unspecified	Neoplasms	0.25	25%
17	Calcification and ossification of muscle	Stenosis and insufficiency of lacrimal passages	ICD10_Chapter	0.22	22%
19	Mitral stenosis	Atrial fibrillation and flutter	Diseases of the circulatory system	0.44	44%
. 30	Pure hypercholesterolaemia	Lymph nodes of head, face and neck	ICD10_Chapter	0.20	20%

Table 5.2 1:Similarity values for two concepts from the ICD-10 taxonomy (Figure 5.1) using Path Length Based Measures (Wu & Palmer).

5.3.1.3 Leacock and Chodorow similarity measure:

The similarity between two concepts is determined by the shortest path length between two concepts node, which connects these two concepts in the taxonomy. The similarity is calculated as the negative algorithm of this value. They proposed a measure that has formula as follows:

$$\mathrm{SimL\&}C = -\log\left[\frac{\mathrm{Sp}\ (\mathrm{c1,c2})}{2(\mathrm{Max\_depth})}\right] \qquad (5)$$

Where:

SP is Shortest Path

Max_depth is the maximum depth of our taxonomy.

From our taxonomy (Figure 5.1), We can calculate the similarity between classes C_1 and C_2 as shown in Table 5.3

Similarity (*Hypertensive renal disease with renal failure*, *Hypertensive renal disease with renal failure*) $= -\log\left(\frac{1}{2(5)}\right) = 1.00$

.

Similarity (Congestive heart failure, Left ventricular failure) $= -\log\left(\frac{3}{2(5)}\right) = 0.52287874528$

.

Similarity (*Pure hypercholesterolaemia*, *Lymph nodes of head, face and neck*) $= -\log\left(\frac{9}{2(5)}\right) = 0.045757490560$

ID	Concept1	Concept2	Length (c1 c2)	Leacok and Chodorow	Sim
4	*Hypertensive renal disease with renal failure*	*Hypertensive renal disease with renal failure*	1	1.00	100%
11	*Congestive heart failure*	*Left ventricular failure*	3	0.52287874528	52%%
8	*Lymph nodes of head, face and neck*	*Major salivary gland, unspecified*	7	0.154901959986	15%
17	*Calcification and ossification of muscle*	*Stenosis and insufficiency of lacrimal passages*	8	0.0969100130081	10%
19	*Mitral stenosis*	*Atrial fibrillation and flutter*	6	0.221848749616	22%
. 30	*Pure hypercholesterolaemia*	*Lymph nodes of head, face and neck*	9	0.0457574905607	5%

Table 5.3 1: Similarity values for two concepts from the ICD-10 taxonomy (Figure 5.1) using Path Leng1th Based Measures (Leacok and Chodorow).

5.3.2 Information Content-Based Similarity Measure:

The information content of a concept c can be quantified as the negative log probability [#]

$$\mathrm{IC}(c) = -\log p(c) \qquad (6)$$

5.3.2.1 Resnik's Measures

Resnik [65] the similarity between a pair of Classes (C1 and C2) is estimated as the amount of taxonomical information they share. In a taxonomy, this information is represented by the Least Common Subsume of both classes (LCS (C1, C2)), which is the most specific taxonomical ancestor common to C1 and C2 in a given ontology. Formally:

$$\mathrm{Sim}res = -\log(\mathrm{P}(\mathrm{LCS}\ (\mathrm{C1}, \mathrm{C2})) = \mathrm{IC}(\mathrm{LCS}\ (\mathrm{C1}, \mathrm{C2})) \qquad (7)$$

Where:

$$\mathrm{IC}(\mathrm{C}) = \frac{\log(\mathrm{Depth}(\mathrm{C}))}{\log(\mathrm{Deep}_{\max})} \qquad (8)$$

From our taxonomy (Figure 5.1), We can calculate the similarity between classes C1 and C2 as shown in Table 5.4

$$\mathrm{IC}\left(\mathrm{LCS}\begin{pmatrix} Hypertensive\ renal\ disease\ with\ renal\ failure, \\ Hypertensive\ renal\ disease\ with\ renal\ failure \end{pmatrix}\right)$$

$$= \mathrm{IC}(Hypertensive\ renal\ disease\ with\ renal\ failure)$$

Depth (*Hypertensive renal disease with renal failure*) = 5 “using node counting”

Deep_max = 5 the maximum depth of ICD10 Ontology.

Then:

$$\mathrm{Sem}_{\mathrm{res}} = \mathrm{IC}(Hypertensive\ renal\ disease\ with\ renal\ failure) = \frac{\log(\mathrm{depth}(\mathrm{C}))}{\log(\mathrm{deep}_{\max})} =$$

$$\log\frac{(5)}{\log(5)} = 1.00$$

.

.

$$\mathrm{IC}(\mathrm{LCS}(Congestive\ heart\ failure, \quad Left\ ventricular\ failure))$$

$$= \mathrm{IC}(Heart\ failure)$$

Depth (*Heart failure*) = 4 “using node counting”

Deepmax = 5 the maximum depth of ICD10 Ontology.

Then:

$$\text{Simres} = \text{IC}(Heart\ failure) = \frac{\log(\text{depth}(\text{C}))}{\log(\text{deep}_{\text{max}})} = \log\frac{(4)}{\log(5)} = 0.86$$

.

.

$$\text{IC}\big(\text{LCS}(Pure\ hypercholesterolaemia, \quad Lymph\ nodes\ of\ head, face\ and\ neck)\big) = \text{IC}(\text{ICD10\_Chapter})$$

Depth ($ICD10\_Chapter$) = 1 “using node counting”

Deepmax = 5 the maximum depth of ICD10 Ontology.

Then:

$$\text{Simres} = \text{IC}(ICD10\_Chapter) = \frac{\log(\text{depth}(\text{C}))}{\log(\text{deep}_{\text{max}})} = \log\frac{(1)}{\log(5)} = 0.00$$

ID	Concept1	Concept2	LCS(c1 c2)	SimResink	Similarity
4	Hypertensive renal disease with renal failure	Hypertensive renal disease with renal failure	5	1.00	100%
11	Congestive heart failure	Left ventricular failure	4	0.86135311614	86%
8	Lymph nodes of head, face and neck	Major salivary gland, unspecified	2	0.43067655807	43%
17	Calcification and ossification of muscle	Stenosis and insufficiency of lacrimal passages	1	0.0	0.00
19	Mitral stenosis	Atrial fibrillation and flutter	2	0.43067655807	43%
. . 30	Pure hypercholesterolaemia	Lymph nodes of head, face and neck	1	0.00	0.00%

Table 5.4 1: Similarity values for two concepts from the ICD-10 taxonomy (Figure 5.1) using information content based Measures (Resink).

5.3.2.2 Lin's Measure:

This measure depends on the relation between information content (IC) of the LCS of two concepts and the sum of the information content of the individual concepts [40].

$$\text{Sim}Lin\ (\text{c1}, \text{c2}) = \frac{2 \times \text{IC}(\text{LCS}\ (\text{C1}, \text{C2}))}{\text{IC}(\text{C1}) + \text{IC}(\text{C2})} \qquad (9)$$

From Resink's measure:

$$\text{IC}(\text{LCS}(\text{Hypertensive renal disease with renal failure}, \text{Hypertensive renal disease with renal failure}))$$

$$= \text{IC}(\text{Hypertensive renal disease with renal failure}) = \frac{\log(5)}{\log(5)} = 1.00$$

$$\text{IC}(\text{Hypertensive renal disease with renal failure}) = \frac{\log(\text{depth}(\text{C}))}{\log(\text{Deep}_{\text{max}})}$$

$$= \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\text{SimLin}\ (\text{Hypertensive renal disease with renal failure}, \text{Hypertensive renal disease with renal failure}) = \frac{2 \times 1}{1 + 1} = 1.00$$

.

.

.

From Resink's Measure:

$$\text{IC}(\text{LCS}(\text{Congestive heart failure}, \quad \text{Left ventricular failure})) = \text{IC}(\text{Heart failure})$$

$$= \frac{\log(4)}{\log(5)} = 0.86$$

$$\text{IC}(\text{Congestive heart failure}) = \frac{\log(\text{depth}(\text{C}))}{\log(\text{Deep}_{\text{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

$$\text{IC}(\text{Left ventricular failure}) = \frac{\log(\text{depth}(\text{C}))}{\log(\text{Deep}_{\text{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\text{SimLin}(\text{Congestive heart failure}, \quad \text{Left ventricular failure}) = \frac{2 \times 0.86}{1 + 1} = 0.86$$

.

.

.

From Resink’s Measure:

$$\mathrm{IC}(\mathrm{LCS}(\text{Pure hypercholesterolaemia}, \quad \text{Lymph nodes of head, face and neck}))$$

$$= \mathrm{IC}(\mathrm{ICD10\_Chapter}) = \frac{\log(0)}{\log(5)} = 0.00$$

$$\mathrm{IC}(\text{Lymph nodes of head, face and neck}) = \frac{\log(\mathrm{depth}(C))}{\log(\mathrm{Deep}_{\mathrm{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

$$\mathrm{IC}(\text{Pure hypercholesterolaemia}) = \frac{\log(\mathrm{depth}(C))}{\log(\mathrm{Deep}_{\mathrm{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\mathrm{SimLin}(\text{ure hypercholesterolaemia}, \quad \text{Lymph nodes of head, face and neck})$$

$$= \frac{2 \times 0.00}{1 + 1} = 0.00$$

ID	Concept1	Concept2	IC(c1)	IC(c2)	IC(LCS(c 1,c2))	SimLin
4	Hypertensive renal disease with renal failure	Hypertensive renal disease with renal failure	1.00	1.00	1.00	100%
11	Congestive heart failure	Left ventricular failure	1.00	1.00	0.86	86%
8	Lymph nodes of head, face and neck	Major salivary gland, unspecified	1.00	1.00	0.43	43%
17	Calcification and ossification of muscle	Stenosis and insufficiency of lacrimal passages	0.86	1.00	0.00	0%
19	Mitral stenosis	Atrial fibrillation and flutter	1.00	0.86	0.43	46%
. . 30	Pure hypercholesterolaemia	Lymph nodes of head, face and neck	1.00	1.00	0.0	0%

Table 5.5 1: Similarity values for two concepts from the ICD-10 taxonomy (Figure 5.1) using information content based Measures (Lin).

5.3.3 Semantic Similarity Measures in the Biomedical Domain:

5.3.3.1 Rada et al. [64] first proposed a semantic distance measure and applied it into the biomedical domain using MeSH ontology. The semantic distance between two classes is the shortest path length between them.

5.3.3.2 Caviedes and Cimino [63] implemented the shortest Path length measure, called CDist, based on the shortest distance between two classes' nodes in the ontology. They evaluated their measure *(CDist measure)* on MeSH, SNOMED, ICD9 ontology based on correlation with human ratings.

5.3.3.3 Pedersen et al. [60] proposed semantic similarity and relatedness in the biomedicine domain in which they applied a corpus-based context vector approach to measure similarity between concepts in SNOMED-CT. Their context vector approach is ontology free but requires training text, for which, they used text data from Mayo Clinic corpus of medical notes.

5.3.3.4 Hisham Al-Mubaid & Nguyen measure [19] [21] proposed measure take the depth of their Least Common Subsume (LCS) and the distance of the shortest path between them. The higher similarity arises when the two concept are in the lower level of the hierarchy. Classes that are more similar with have a lower similarity score than classes that are less similar with this measure. Their similarity measure is:

$$\text{Sim}\ (C1, C2) = \log_2([L\ (C1, C2) - 1]^{\alpha} \times [\text{CSpec}(C1, C2)]^{\beta} + k) \qquad (10)$$

$$\text{CSpec}(C1, C2) = D - \text{depth}\left(\text{LCS}(C1, C2)\right) \qquad (11)$$

Where:

α> 0 and β > 0 are contribution factors of two features (Path and CSpec).

Depth (LCS(C1, C2)) is depth of LCS(C1, C2) using node counting.

L(C1, C2) is shortest path length between the two concept nodes.

D is maximum depth of the taxonomy.

K is constant, and CSpec feature is calculated as in (11). We use logarithm function (inverse of exponentiation) for semantic distance (10), which is the inverse of semantic similarity.

To insure the distance is positive and the combination is non-linear, k must be greater or equal to one (k >= l). In this thesis, k=l is used in experiments. When two concept nodes have path length of 1 (Path=l) using node counting (i.e., they are in the same node in the ontology), they have a semantic distance (*SemDist*) equals to zero (i.e. maximum similarity) regardless of common specificity feature.

The maximum value of this measure occurs when one concept is the left-most leaf node, and the other concept is the right-most leaf node in the tree. In ICD10 terminology the maximum value is $\log_2$ ([22-1]*[5-1] + 2) equal 6.4262647547. Therefore, the similarity distance values will be in [1.0000, 6.4262647547] in ICD10 terminology.

The single-cluster path length feature:

From our taxonomy (Figure 5.1), We can calculate the similarity between classes C1 and C2 as the following:

Path length (*Hypertensive renal disease with renal failure*, *Hypertensive renal disease with renal failure*) = 1 “using node counting”

CSpec (*Hypertensive renal disease with renal failure, Hypertensive renal disease with renal failure*) = D – depth (LCS (I12.0))

= 5 – 5 = 0

So, similarity

Sim (*Hypertensive renal disease with renal failure, Hypertensive renal disease with renal failure*)) = $\log_2$([1 - 1]1 × [0]1 + 2) = $\log_2$(2) = 1

.
.
.

The cross-cluster path length feature:

Let us conceder the example, shown in Figure 5.1. The root is node that connects all the clusters. The path length between two concept nodes (C1 and C2) is computed by adding up the two shortest path lengths from the two nodes to their LCS node (their LCS is the root). For example, in Figure 5.1, for the two concept nodes (*Pure hypercholesterolaemia, Lymph nodes of head, face and neck*), the LCS is the root ICD-10. So, the path length between *Pure hypercholesterolaemia,* and *Lymph nodes of head, face and neck* is calculated as follows:

Path (*Pure hypercholesterolaemia, Lymph nodes of head, face and neck*) = d1 + d2 -1

Where d1 = d (*Pure hypercholesterolaemia*, root) and d2 = d (*Lymph nodes of head, face and neck*, root), where d (*Pure hypercholesterolaemia*, root) is the path length from the root ICD-10 to node *Pure hypercholesterolaemia*, and similarly d (*Lymph nodes of head, face and neck*, root) is the path length from ICD-10 to node *Lymph nodes of head, face and neck.* One is subtracted in the above equation, because the root node is counted twice.

$$\text{Path (Pure hypercholesterolaemia, Lymph nodes of head, face and neck)}$$

$$= d1 + \frac{2D1 - 1}{2D2 - 1} \times d2 - 1$$

$$\text{Path(Pure hypercholesterolaemia, Lymph nodes of head, face and neck)} =$$

$$5 + \frac{10-1}{10-1} \times 5 - 1 \quad = 9$$

CSpec (Pure hypercholesterolaemia, Lymph nodes of head, face and neck) = D primary - 1 = 5 – 1 = 4

ID	Concept1	Concept2	L (c1,c2)	CSPec(c1, c2)	SemDist	Note
4	Hypertensive renal disease with renal failure	Hypertensive renal disease with renal failure	1	0	1	Same code
11	Congestive heart failure	Left ventricular failure	3	1	2	Same group
8	Lymph nodes of head, face and neck	Major salivary gland, unspecified	7	3	4.32	Same code
17	Calcification and ossification of muscle	Stenosis and insufficiency of lacrimal passages	8	4	4.91	Same chapter
19	Mitral stenosis	Atrial fibrillation and flutter	6	3	4.09	Same section
30	Pure hypercholesterolaemia	Lymph nodes of head, face and neck	9	4	5.09	Different chapter

Table 5.6 1: Similarity values for two classes from the ICD-10 taxonomy (Figure 5.1) using Path Length Based Measure (Al-Mubaid and Nguyen).

So, similarity

SemDist (Pure hypercholesterolaemia, Lymph nodes of head, face and neck) = $\log_2($ $[Path - 1]^{\alpha} \times [CSpec]^{\beta} + k)$

$= \mathrm{Log}_2 ((9 - 1) \times (4) + 2) = \log2 (34) = 5.09$

Table 5.7 1: Dataset 1: 30 medical term pairs sorted in the order of the averag

Id	***Concept1***	***Concept2***	***Phys***	***Expert***
4	Renal failure I12.0	Kidney failure I12.0	4.0000	4.0000
5	Heart I51.5	Myocardium I51.5	3.3333	3.0000
1	Stroke I64	Infarct I64	3.0000	2.7778
7	Abortion O03	Miscarriage O03	3.0000	3.3333
9	Delusion (F06.2)	Schizophrenia (F06.2)	3.0000	2.2222
11	Congestive heart failure (I50.0)	Pulmonary edema (I50.1)	3.0000	1.4444
8	Metastasis (C77.0)	Adenocarcinoma (C08.9)	2.6667	1.7778
17	Calcification (M61)	Stenosis (H04.5)	2.6667	2.0000
10	**Diarrhea**	**Stomach cramps**	**2.3333**	**1.3333**
19	Mitral stenosis (I05.0)	Atrial fibrillation (I48)	2.3333	1.3333
20	Chronic obstructive pulmonary disease (J44.9)	Lung infiltrates (J82)	2.0000	1.8889
2	Rheumatoid arthritis (M05.3)	Lupus (L93)	2.0000	1.1111
3	Brain tumor (G94.8)	Intracranial hemorrhage(I69.2)	2.0000	1.3333
15	Carpal tunnel Syndrome (G56.0)	Osteoarthritis (M19.9)	2.0000	1.1111
18	Diabetes mellitus (E10-E14)	Hypertension (I10-I15)	2.0000	1.0000
27	**Acne**	**Syringe**	**2.0000**	**1.0000**
12	Antibiotic (Z88.1)	Allergy (Z88.1)	1.6667	1.2222
13	**Cortisone**	**Total knee replacement**	**1.6667**	**1.0000**
14	**Pulmonary embolus**	**Myocardial infarction**	**1.6667**	**1.2222**
16	Pulmonary Fibrosis (E84.0)	Lung Cancer (C34.1)	1.6667	1.4444
6	**Cholangiocarcinoma**	**Colonoscopy**	**1.3333**	**1.0000**
29	Lymphoid hyperplasia (K38.0)	Laryngeal Cancer (C32.0)	1.3333	1.0000
21	Multiple Sclerosis (F06.8)	Psychosis (F06.8)	1.0000	1.0000
22	Appendicitis (K35)	Osteoporosis (M80)	1.0000	1.0000
23	Rectal polyp (K62.1)	Aorta (I70.0)	1.0000	1.0000
24	Xerostomia (K11.7)	Alcoholic cirrhosis (K70.3)	1.0000	1.0000
25	Peptic ulcer disease (K21.0)	Myopia (H52.1)	1.0000	1.0000
26	Depression (F20.4)	Cellulitis (H60.1)	1.0000	1.0000
28	**Varicose vein**	**Entire knee meniscus**	**1.0000**	**1.0000**
30	Hyperlipidemia (E78.0)	Metastasis (C77.0)	1.0000	1.0000

5.4 Evaluation

5.4.1 Dataset

There are no standard human rating sets of concepts/terms for semantic similarity in the biomedical domain. Thus, to evaluate semantic similarity measures, the dataset of 30 concept pairs from Pedersen T. et al. (2006) [60], (Dataset 1) which was annotated by 3 physicians and 9 medical index experts. Each pair was annotated on a 4-point scale: “practically synonymous, related, marginally related, and unrelated”.

Table 1 contains whole pairs of this dataset. The average correlation between physicians is 0.68, and between experts is 0.78. Because the experts are more than the physicians, and the correlation (agreement) between experts (0.78) is higher than the correlation between physicians (0.68), it can be assumed that the experts’ rating scores are more reliable than the physicians’ rating scores.

Only 24 out of the 30 term pairs are found in ICD-10 using ICD-10 browser version 2010 [61] as some terms cannot be found, 24 pairs was used in the experiments (Pedersen et. al. [60] tested 29 out of the 30 concept pairs as one pair was not found in SNOMED-CT).

The term pairs in bold, in Table 5.7, are the ones that contains a term that was not found in ICD-10 and they were excluded in experiments.

Table 5.7 Dataset 1: 30 medical term pairs sorted in the order of the average

5.4.2 Experiments and Results

In these experiments, only one dataset was used, ICD10 Ontology was used as information source for the semantic similarity measures and one dataset are used for evaluation. All measures use node counting for path lengths and depths of concept nodes. Out of the 30 pairs of Dataset 1, only 24 pairs in ICD10 were found. For the six pairs that were not found in ICD10, average distance/similarity values of the most related concept nodes to each one of them were calculated, so there were 24 pairs in ICD10 in total. The results of absolute correlations with human scores using dataset, experimented on ICD10 Ontology, are shown in Tables 5.8 and Figure 5.2. Table 5.8 shows for the six measures the results of correlation with human ratings of physicians, experts, and both (phys. and experts), with the ranks between parentheses. These correlation values (Table 5.8) show that *Al-Mubaid and Nguyen's* (SemDist) measure is ranked #1 in correlation relative to experts' judgments and relative to both (expert and phys. judgments). But relative to physician judgments, the SemDist approach is ranked #2. The experimental results demonstrated that *Al-Mubaid and Nguyen's* (SemDist) measure can achieve high correlations with human similarity scores.

Measure	Phys. (rank)	Expert (rank)	Both (rank)
SemDist	0.6007 (3)	**0.6641 (1)**	**0.6548 (1)**
Lin	0.6045 (2)	0.6563 (2)	0.6526 (2)
Path Length	**0.6118 (1)**	0.6505 (5)	0.6436 (4)
Wu Palmer	0.5865 (4)	0.6508 (4)	0.6451 (3)
L&C	0.5801 (5)	0.6558 (3)	0.6401 (5)
Resink	0.5576 (6)	0.6207 (6)	0.6096 (6)

Table 5.8 1: Absolute correlations with human scores for all measures using ICD10 on Dataset1

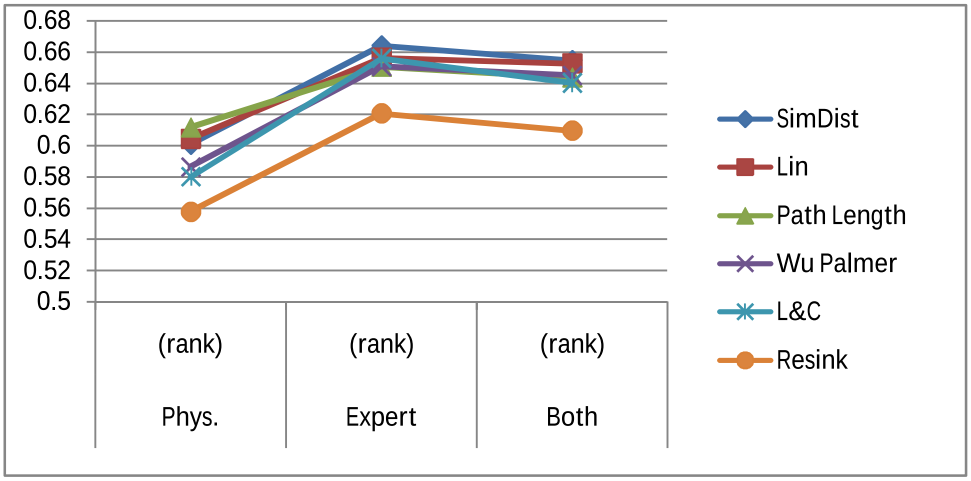


Figure 5.2 1: Results of correlations with human scores for six measures using ICD10 "V1.0" Ontology.

Chapter Six
Testing and Evaluation
6.1 Overview

This chapter will give an overview of the testing practices adopted over the project lifecycle along with the evaluations made of the final model.

6.2 Ontology evaluation approaches:

Various approaches to the evaluation of ontologies have been considered in the literature, depending on what kind of ontologies is being evaluated. Most evaluation approaches fall into one of the following categories:

- ✓ Based on comparing the ontology to a "golden standard".
- ✓ Based on using the ontology in an application and evaluating the results.
- ✓ Involving comparisons with a source of data (e.g. a collection of documents) about the domain to be covered by the ontology.
- ✓ Where evaluation is done by humans who try to assess how well the ontology meets a set of predefined criteria, standards, requirements.

6.3 Ontology evaluation at different levels:

Ontology is a fairly complex structure and it is often more practical to focus on the evaluation of different levels of the ontology separately rather than trying to directly evaluate the ontology as a whole. This is particularly true if we want a predominantly automated evaluation rather than entirely carried out by human users/experts. Another reason for the level-based approach is that when automatic learning techniques have been used in the construction of the ontology, the techniques involved are substantially different for the different levels.

The individual levels have been defined variously by different authors, but these various definitions tend to be broadly similar and usually involve the following levels:

a) *Hierarchy or taxonomy*: Ontology typically includes a hierarchical is-a relation between concepts.

b) *Other semantic relations:* The ontology may contain other relations besides is-a, and these relations may be evaluated separately. This typically includes measures such as precision and recall.

c) *Context or application level:* An ontology may be part of a larger collection of ontologies, and may reference or be referenced by various definitions in these other ontologies. In this case it may be important to take this context into account when evaluating it. Another form of context is the application where the ontology is to be used, evaluation looks at how the results of the application are affected by the use of the ontology.

d) *Syntactic level:* Evaluation on this level may be of particular interest for ontologies that have been mostly constructed manually. The ontology is usually described in a particular formal language and must match the syntactic requirements of that language. Various other syntactic considerations, such as the presence of natural-language documentation, avoiding loops between definitions, etc., may also be considered.

e) *Structure, architecture, design:* This is primarily of interest in manually constructed ontologies. We want the ontology to meet certain pre-defined design principles or criteria; structural concerns involve the organization of the ontology and its suitability for further development.

f) *Lexical, vocabulary, or data layer:* Here the focus is on which concepts, instances, facts, etc. have been included in the ontology, and the vocabulary used to represent or identify these concepts. Evaluation on this level tends to involve comparisons with various sources of data concerning the problem domain (e.g. domain-specific text corpora), as well as techniques such as string similarity measures (e.g. edit distance). This sort of evaluation usually proceeds entirely manually.

6.4 Evaluating the reference dataset:

The reference dataset test set aims at assessing the strengths and the weaknesses of matching systems, depending on the availability of ontology features, i.e., the availability of instances, properties or labels in the ontology.

6.5 Testing the reference dataset

- Compare the original ontology with itself.
- Compare the original ontology with the ontology obtained by applying the following set of modifications.
- Compare the original ontology with real ones found on the web.

6.5.1 Comparing a biomedical ontology with Infectious and Parasitic DO-Reference Dataset:

The golden standard could be in fact another ontology or it could be taken statistically from a corpus of documents or prepared by domain experts. In our work we using approach based on comparing the ontology to a *"golden standard"*. The "gold standard" based ontology evaluation depends on calculating the similarity between concepts in two different ontologies in the same domain such as (*doid* and *Infectious and Parasitic DO-Bench*) using *Al-Mubaid and Nguyen's measure (SemDist).*

In this thesis we can use the terms *"gold standard"* and *"Infectious and Parasitic DO-Bench"* interchangeably to refer to the same thing.

6.5.2 Manual Testing

Testing the semantic similarity measures (SemDist) were done manually. The first version of the reference dataset includes around one hundred and thirty one concepts. compare the original ontology with itself

6.5.2.1 Lexical, vocabulary or data comparison level:

In the real case of evaluation:

ID	Concept1(Class) doid Ontology	Code ICD-10	Concept2 (Class) Our Dataset	Code ICD-10
1	Cholera	A00.0	Cholera due to Vibrio cholerae 01, biovarcholerae	*A00.0*
2			***Cholera due to Vibrio cholerae 01, biovareltor***	***A00.1***
3	Cholera	A00.9	Cholera, unspecified	*A00.9*
4	pulmonary tuberculosis	A15, A15.0	Tuberculosis of lung, confirmed by sputum microscopy with or without culture	*A15.0*
5 .			***Tuberculosis of lung, confirmed by culture only***	***A15.1***
131			***Unspecified staphylococcus as the cause of diseases classified to other chapters***	***B95.8***

Table 6.1 1: Compare between concepts (classes) using our reference dataset and doid ontology.

The similarity between our *Infectious and Parasitic DO-Benc*h and *doid* ontology is 53%, we find 69 concepts similar to our dataset. Out of the 131 concepts tested in our approach, only 69 concepts were included in the test. This was because some concepts in the pairs were not present in the ***doid*** ontology.

ID	Concept1(Class) SNOMED-CT Ontology	Code in SNOMED-CT	Concept2 (Class) Our Dataset	Code in ICD-10
1	Cholera due to Vibrio choleraeEl Tor		Cholera due to Vibrio cholerae 01, biovarcholerae	A00.0
2			***Cholera due to Vibrio cholerae 01, biovareltor***	***A00.1***
3			***Cholera, unspecified***	***A00.9***
4	Tuberculosis of lung, confirmed by sputum microscopy with or without culture		Tuberculosis of lung, confirmed by sputum microscopy with or without culture	A15.0
5	Tuberculosis of lung, confirmed by culture only		Tuberculosis of lung, confirmed by culture only	A15.1
131			***Unspecified staphylococcus as the cause of diseases classified to other chapters***	***B95.8***

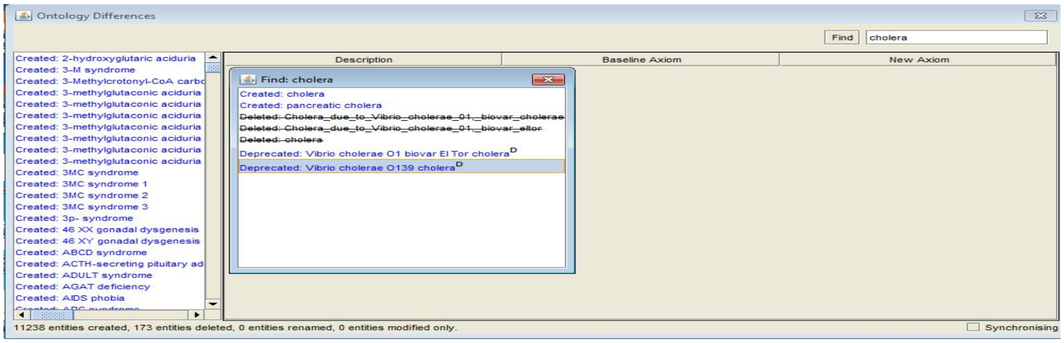
Table 6.2 1: Compare between concepts (classes) using our reference dataset and SNOMED-CT ontology.

The similarity between pair of classes in our *Infectious and Parasitic DO-Reference Dataset*, which are used as “gold standard” and SNOMED-CT ontology is 62%, we find 81 concepts similar to our dataset. Out of the 131 concepts tested in our approach, only 81cocepts were included in the test. This was because some concepts in the pairs were not present in the SNOMED-CT ontology.

6.5.3 Testing using Protégé tool:

We compare between two ontologies (***diod*** ontology and our reference dataset) using protégé tool.

In this test, we used the first concept “A00.0” and we compared it with other two concepts A00.1, A00.2, and with them self.



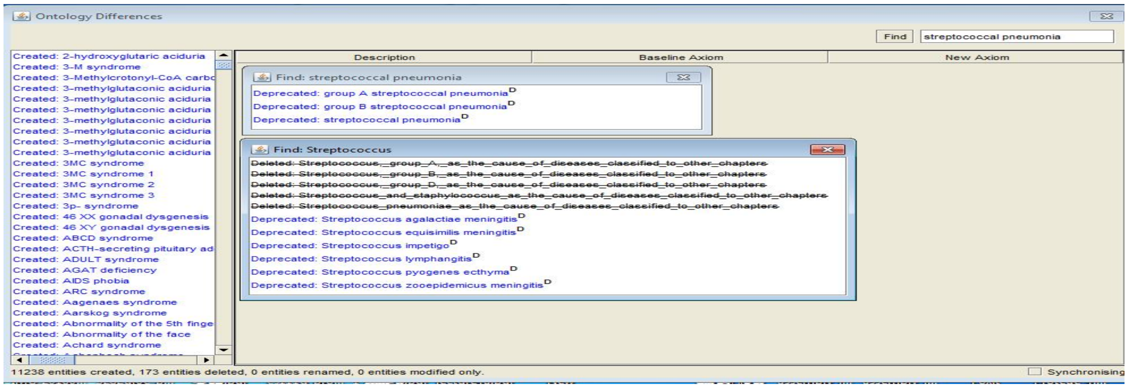
Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia

Created: cholera
Created: pancreatic cholera
~~Deleted: Cholera_due_to_Vibrio_cholerae_01,_biovar_cholerae~~
~~Deleted: Cholera_due_to_Vibrio_cholerae_01,_biovar_eltor~~
~~Deleted: cholera~~
Deprecated: Vibrio cholerae O1 biovar El Tor cholera D
Deprecated: Vibrio cholerae O139 cholera D

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

Id	Concept1(Class) doid Ontology	Code ICD-10	Concept2 (Class) Our Dataset	CodeICD-10
1	Vibrio Cholera O139, Cholera		Cholera due to Vibrio cholerae 01, biovarcholerae	A00.0
2	Vibrio Cholera choleraeO1, biovareltor Cholera		Cholera due to Vibrio cholerae 01, biovareltor	A00.1
3	Cholera		Cholera, unspecified	A00.9

Table 6.3 1: Compare between concepts (classes) using our reference dataset and doid ontology



Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome
Created: Aarskog syndrome
Created: Abnormality of the 5th finge
Created: Abnormality of the face
Created: Achard syndrome

Find: streptococcal pneumonia

Deprecated: group A streptococcal pneumonia^{D}
Deprecated: group B streptococcal pneumonia^{D}
Deprecated: streptococcal pneumonia^{D}

Find: Streptococcus

~~Deleted: Streptococcus,_group_A,_as_the_cause_of_diseases_classified_to_other_chapters~~
~~Deleted: Streptococcus,_group_B,_as_the_cause_of_diseases_classified_to_other_chapters~~
~~Deleted: Streptococcus,_group_D,_as_the_cause_of_diseases_classified_to_other_chapters~~
~~Deleted: Streptococcus_and_staphylococcus_as_the_cause_of_diseases_classified_to_other_chapters~~
~~Deleted: Streptococcus_pneumoniae_as_the_cause_of_diseases_classified_to_other_chapters~~
Deprecated: Streptococcus agalactiae meningitis^{D}
Deprecated: Streptococcus equisimilis meningitis^{D}
Deprecated: Streptococcus impetigo^{D}
Deprecated: Streptococcus lymphangitis^{D}
Deprecated: Streptococcus pyogenes ecthyma^{D}
Deprecated: Streptococcus zooepidemicus meningitis^{D}

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

123	*group A streptococcal pneumonia*	-	*Streptococcus, group A, as the cause of diseases classified to other chapters*	*B95.0*
124	*group B streptococcal pneumonia*	-	*Streptococcus, group B, as the cause of diseases classified to other chapters*	*B95.1*
125			***Streptococcus, group D, as the cause of diseases classified to other chapters***	***B95.2***
126			***Streptococcus pneumoniae as the cause of diseases classified to other chapters***	***B95.3***
127			***Other streptococcus as the cause of diseases classified to other chapters***	***B95.4***
128			***Unspecified streptococcus as the cause of diseases classified to other chapters***	***B95.5***
129			***Staphylococcus aureus as the cause of diseases classified to other chapters***	***B95.6***
130			***Other staphylococcus as the cause of diseases classified to other chapters***	***B95.7***
131			***Unspecified staphylococcus as the cause of diseases classified to other chapters***	***B95.8***

Table 6.4 1: Compare between concepts (classes) using our reference dataset and (doid) ontology. For more details see [Appendix B]

Chapter Seven
Conclusion and Future Work
7.1 Thesis Contribution

In this thesis, we proposed a reference dataset from the perspective of the semantic similarity measure defined for the biomedical domain based on the UMLS frame work. We deployed the SemDist measure to development the reference dataset. We extracted the concepts of reference dataset from the domain knowledge (ICD-10 version 1.0). We used this reference dataset to check the quality of ontology in biomedical domain.

7.2 Conclusion

Our study demonstrated the usefulness of our approach to evaluate the ontology quality. The results discussed in this thesis has shown that, the SemDist(C1, C2) similarity (proposed by Al-Mubaid and Hoa A. Nguyen) has achieved high matching score by the expert's judgment to measure similarity between concepts in the biomedical domain. This is an important step that can affect the reusability of the ontology. Our study also demonstrated the usefulness of our reference dataset model to evaluate the quality of ontologies from the perspective of a similarity measure. Our approach can be reused to support the evaluation of additional ontologies in the biomedical domain.

7.3 Recommended Future Work

We recommend the following ideas that can be used for future:

The results we found in the my work in chapter six is very interesting, since our ontology evaluation combines new techniques and procedures that were never used before, we gathered our concepts from trusted sources in the ICD-10 domain.

In our future wok we will avoid some of the limitations of the manual procedure we had to make in measuring our semantic similarity of the ICD10 ontology. In case we did not find an ontology tool that solves our problem, we propose the development of a computerized program that is designed specifically to make our ontology evaluation methodology fully automated.

Moreover, in order to further improve the accuracy of semantic similarity measuring, we will attempt to introduce more factors which have effect on the semantic similarity, such as the relationship between concept nodes and the strength of edge in the ICD-10 taxonomy.

7.4 From a methodological perspective, there are at least three open problems:

1. The lack of a software for reference dataset methodology.
2. The difficulty of using current evaluation and improvement methodologies with Semantic Web content.
3. The absence of integrated methods and techniques supporting the complex task of reference dataset Semantic Web content.

The main features of our proposed approach are that it focuses on fully automated evaluation of ontologies, based semantic similarity measures. We used SemDist (C_1, C2) ontology similarity measure, designed by Al-Mubaid and Hoa A. Nguyen that is commonly used and adapted for biomedical domain.

References

[1] Sara Garcia-Ramos, Abraham Otero, and Mariano Fernandez-Lopez."Ontology Test: A tool to evaluate ontologies through Tests Defined by the User" , Part II, LNCS 5518, pp. 91-98, 2009.

[2] Ahmad Kayed, et al. "Ontology Evaluation: Which Test to Use" 2013 5th International Conference on Computer Science and Information Technology (CSIT), pp. 45-48, 2013.

[3] JiabinRuan, Yubin Yang, "Assess Content Comprehensiveness of Ontologies" 2010 Second International Conference on Computer Modeling and Simulation. pp.536-539, 2010.

[4] ZurErlangung des akademischen Grades eines, et al. “Ontology Evaluation” PhD thesis, June, 2010.

[5] Van-Anh Tran, et al. "OnWARD: Ontology-driven web-based framework for multi-center clinical studies" Journal of Biomedical Informatics 44 (2011) S48–S53.

[6] SébastienHarispe, et al. “A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain” Journal of Biomedical Informatics (2013).

[7] J. Kulandai Josephine Julina, D. Thenmozhi. "Ontology Based EMR for Decision Making in Health Care Using SNOMED CT", IEEE, pp. 514- 519, ICRTIT-2012.

[8] HebaAyeldeen, et al. “Evaluation of Semantic Similarity across MeSH Ontology: A Cairo University Thesis - Mining Case Study” 2013 12th Mexican International Conference on Artificial Intelligence pp. 139 -144.

[9] D.Sathya, K.R.Uthayan, Assistant Professor. “Proposal for Semantic Metric to Assess the Quality of Ontologies” Proceedings of 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN 2011) pp 754 -756.

[10] Stefano David, “Defining a benchmark suite for evaluating the import of OWL Lite ontologies” ” Master thesis, July, 2006.

[11] Wilson Yiksen Wong, “Learning Lightweight Ontologies from Text across Different Domains using the Web as Background Knowledge” PhD thesis, September 2009.

[12] Samir Tartir. “ONTOLOGY-DRIVEN QUESTION ANSWERING AND ONTOLOGY QUALITY EVALUATION” PhD thesis, May, 2009.

[13] Marcos Martínez-Romero, et al. “BiOSS: A system for biomedical ontology selection” computer methods and programs in biomedicine 114 (2 0 1 4) 125–140.

[14] FaezehEnsan, Weichang Du. “A semantic metrics suite for evaluating modular ontologies” Information Systems 38 (2013) 745–770. January 2013.

[15] Hliaoutakis, “Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline”. Master’s thesis, Technical University of Crete, Greek. 2005.

[16] Qing Lu, “Onto KBEval: A Support Tool for OWL Ontology Evaluation” Master thesis of Computer Science, September, 2006.

[17] Dekang Lin “An Information-Theoretic Definition of Similarity”

[18] David Sanchez, et al. “Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain” Journal of Biomedical Informatics 45 (2012) 141–155.

[19] Hisham Al-mubaid& Hoa A. Nguyen “Cluster-Based Approach for Semantic Similarity in the Biomedical Domain” Proceeding of the 28th IEEE, New York City, Aug 30-Sep 3, 2006.

[20] Montserrat Batet, et al. “An ontology-based measure to compute semantic similarity in biomedicine” Journal of Biomedical Informatics 44 (2011) 118–125.

[21] Hisham Al-mubaid& Hoa A. Nguyen “measuring Semantic Similarity between Biomedical concepts within multiple ontologies” IEEE Trans Syst Man Cybern Part C: Appl Rev 2009, 39.

[22] Vijay N Garla1 and Cynthia Brandt."Semantic similarity in the biomedical domain: an evaluation across knowledge sources" Garla and Brandt BMC Bioinformatics 2012, 13:261.

[23] D.Sathya&K.R.Uthayan, "Proposal for Semantic Metric to Assess the Quality of Ontologies" Proceedings of 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN 2011).

[24] Sheau-Ling Hsieh, et al. "Semantic Similarity Measures in the Biomedical Domain by Leveraging a Web Search Engine" IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 17, NO. 4, JULY 2013.

[25] Alexander C. Yu, "Methods in biomedical ontology" Journal of Biomedical Informatics 39 (2006) 252–266.

[26] Hoa A. Nguyen & Hisham Al-mubaid "New ontology-based Semantic Similarity measure for the Biomedical domain" IEEE 2006.

[27] Astrid Duque-Ramos, et al. "Evaluation of the OQuaRE framework for ontology quality" Expert Systems with Applications 40 (2013) 2696–2703.

[28] Katrin Simone Zai, "Instance-Based Ontology Matching and the Evaluation of Matching Systems" November 2010.

[29] DraganGasevic, et al. "Model Driven Architecture and Ontology Development" Springer Berlin Heidelberg New York.pp 93-109.

[30] Karin K. Breitman, et. al. "Semantic Web, concept, technologies, and applications" Springer –Verlage London Limited 2007.

[31] MICHAEL J. GROVE. "Development of an Ontology for Rehabilitation: Traumatic Brain Injury" PhD thesis. SEPTEMBER 2013.

[32] ArashShaban-Nejad, "Design and Development of an Integrated Formal Ontology for Fungal Genomics Master of Computer Science", March 2005.

[33] Roberto Poli, Dr. Michael Healy, Achilles Kameas, “Theory and Applications of Ontology: Computer Applications” © Springer Science Business Media B.V. 2010.

[34] Raul Garcia Castro “Benchmarking Semantic Web technology”, PhD Thesis, July, 2008.

[35] Jonathan Yu, “Requirements-Oriented Methodology forEvaluating Ontologies” PhD thesis, July 2008.

[36] https://dkm.fbk.eu/technologies/icd-10-ontology, Authors: Elena Cardillo, Andrei Tamilin, Claudio Eccher, and Luciano Serafini, March 2008.

[37] Sure et al, “Why evaluate ontology technologies? because it works!” IEEE Intelligent Systems, pp 74 – 81, July 2004.

[38] VipulKashyap et al, “The Semantic Web: Semantics for Data and Services on the Web” Springer-Verlag Berlin Heidelberg -2008, pp 79-84.

[39] KlaokanlayaSilachan, PanjaiTantatsanawongPh.D, “Domain Ontology Health Informatics Service From Text Medical Data Classification” IEEE Annual SRII Global Conference, 2011, pp 357-362

[40] Abdelrahman, A.M.B. and Kayed, A. (2015) A Survey on Semantic Similarity Measures between Concepts in Health Domain. American Journal of Computational Mathematics, 5, 204-214.

[41] Hoa A. Nguyen, “New Semantic Similarity Techniques of Concepts Applied in the Biomedical Domain and WordNet” Master Thesis, Dec, 2006.

[42] World Health Organization., “ICD-10, International Statistical Classification of Diseases and Related Health Problems.” 10th Revision.5th ed. Vol.2 instruction manual (2016).

[43] NamiraMohammadiDinani, “Ontology Development for Drug-Disease Knowledge Management” Master Thesis, June 2012.

[44] Guarino, “Formal ontology and information systems” 1998.

[45] J. Brank, et al. “A survey of ontology evaluation techniques, Proceedings of the Conference on Data Mining and Data Warehouses”, 2005.

[46] J. Brank, et al. “Gold standard based ontology evaluation using instance assignment”, Edinburgh, UK, 2006.

[47] KarimKamounet al. “A Novel Global Measure Approach based on OntologySpectrum to Evaluate Ontology Enrichment” International Journal of Computer Applications (0975 – 8887), Volume 39– No.17, February 2012.

[48] SteffenStaab “Why Evaluate Ontology Technologies? Because It Works!”IEEE INTELLIGENT SYSTEMS, JULY/AUGUST 2004.

[49] Hoehndorf Ret. al. “Evaluation of research in biomedical” Briefings in Bioinformatics. September, 2012.

[50] Brank J, et. al. “A survey of ontology evaluation techniques”Proceedings of the Conference on Data Mining and Data Warehouses. 2005. p. 166–70.

[51] Kaifeng, et. al. “AN IMPROVED METHOD FOR MEASURING CONCEPT SEMANTIC SIMILARITY COMBINING MULTIPLE METRICS”Proceedings of IEEE IC-BNMT2013.

[52] Samiullah Khan, et. al. “A Framework for Evaluation of OWL Biomedical Ontologies based on Properties Coverage”2015 13th International Conference on Frontiers of Information Technology.

[53] Basili et al, “Experimentation in software engineering” IEEE Transactions on Software Engineering, January 1986.

[54] Park et al., ”Goal-Driven Software Measurement-Guidebook” Software Engineering Institute,August 1996.

[55] Gedigaet al, “Evaluation of Software System”, pages 166–192. 2002.

[56] Tiffani JeDawn Bright, “Development and Evaluation of an Ontology for Guiding Appropriate Antibiotic Prescribing” PhD Thesis COLUMBIA UNIVERSITY 2009.

[57] Roxana Dogaru, et. al, “Searching for Taxonomy-based Similarity Measures for Medical Data”BCI September 2015.

[58] David Sánchez, “Semantic variance: An intuitive measure for ontology accuracy evaluation”, Engineering Applications of Artificial Intelligence 39 (2015) 89–99

[59] M. Sabou, J. Garcia, S. Angeletou, M. d'Aquin, E. Motta, Evaluating the Semantic web: a task-based approach, In: Proc. of 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference2007, pp. 423-437.

[60] Pedersen,T. et al, “Measures of Semantic Similarity andRelatedness in the Medical Domain”, University of Minnesota Digital Technology Center Research Report, Journal of Biomedical Informatics April 2006.

[61] http://apps.who.int/classifications/icd10/browse/2010/en

[62] http://protege.stanford.edu/

[63] Caviedes, J. and Cimino, J. “Towards the development of a conceptual distance metric for the UMLS”. Journal of Biomedical Informatics 37,77-85, 2004.

[64] Rada, et. al.“Development and Application of a Metric on Semantic Net”.IEEE Transactions on Systems, Man and Cybernetics,19,1(1989),17-30.

[65] Resnik, P. “Using information content to evaluate semantic similarity in ontology”. Joint Conference on Artificial Intelligence,pp448-453,1995.

[66] UMLSKS. Available:

http://umlsks.nlm.nih.gov

[67] MeSH Browser. Available:

http://www.nlm.nih.gov/mesh/MBrowser.html

[68] https://www.ebi.ac.uk/ols/ontologies.

[69] V. Jain and M. Singh, "Ontology Development and Query Retrieval using Protégé Tool"
International Journal of Intelligent Systems and Applications (IJISA), vol. 5, pp. 67, (2013).

https://raw.githubusercontent.com/monarch-initiative/monarch-disease
ontology/master/src/mondo/mondo.owl

[70] Jain, et. al. “Data Clustering: A Survey. ACM Computing Surveys”, September 1999.

[71] Farley and Raftery “How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis”, Department of Statistics University of Washington, 1998.

[72] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[73] Everitt and Brian, “Dictionary of Statistics” Cambridge University, (1998).

[74] Sneath, P., and Sokal, R. “Numerical Taxonomy” San Francisco, 1973.

[75] King, B. Step-wise Clustering Procedures, pp. 86-101, 1967.

[76] Murtagh, F. “A survey of recent advances in hierarchical clustering algorithms which use cluster centers”, 1984.

[77] Ward, J. H. “Hierarchical grouping to optimize an objective function”. Journal of the American Statistical Association, 1963.

[78] JAIN, A. K. AND DUBES, “Algorithms for Clustering Data” 1988.

[79] M. Venkat Reddy et. al “Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering” International Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 5, Sep – Oct 2017

[80] Wu, Z., and Palmer, M. Verb “semantics and lexical selection” 133-138, 1994.

[81] https://www.nlm.nih.gov/research/umls/quickstart.html

Appendix A

ID	Class Name (Concept)	ICD-10Codes	Notes
1	*Cholera due to Vibrio cholerae 01, biovarcholera*	*A00.0*	
2	*Cholera due to Vibrio cholerae 01, biovareltor*	*A00.1*	
3	*Cholera, unspecified*	*A00.9*	
4	*Tuberculosis of lung, confirmed by sputum microscopy with or without culture*	*A15.0*	
5	*Tuberculosis of lung, confirmed by culture only*	*A15.1*	
6	*Tuberculosis of lung, confirmed his tologically*	*A15.2*	
7	*Tuberculosis of lung, confirmed by unspecified means*	*A15.3*	
8	*Tuberculosis of intrathoracic lymph nodes, confirmed bacteriologically and histologically*	*A15.4*	
9	*Tuberculosis of larynx, trachea and bronchus, confirmed bacteriologically and histologically*	*A15.5*	
10	*Tuberculous pleurisy, confirmed bacteriologically and histologically*	*A15.6*	
11	*Primary respiratory tuberculosis, confirmed bacteriologically and histologically*	*A15.7*	
12	*Other respiratory tuberculosis, confirmed bacteriologically and histologically*	*A15.8*	
13	*Respiratory tuberculosis unspecified, confirmed bacteriologically and histologically*	*A15.9*	
14	*Bubonic plague*	*A20.0*	
15	*Cellulocutaneous plague*	*A20.1*	
16	*Pneumonic plague*	*A20.2*	
17	*Plague meningitis*	*A20.3*	
18	*Septicaemic plague*	*A20.7*	

19	*Other forms of plague*	*A20.8*	
20	*Plague, unspecified*	*A20.9*	
21	*Indeterminate leprosy*	*A30.0*	
22	*Tuberculoid leprosy*	*A30.1*	
23	*Borderline tuberculoid leprosy*	*A30.2*	
24	*Borderline leprosy*	*A30.3*	
25	*Borderline lepromatous leprosy*	*A30.4*	
26	*Lepromatous leprosy*	*A30.5*	
27	*Other forms of leprosy*	*A30.8*	
28	*Leprosy, unspecified*	*A30.9*	
29	*Early congenital syphilis, symptomatic*	*A50.0*	
30	*Early congenital syphilis, latent*	*A50.1*	
31	*Early congenital syphilis, unspecified*	*A50.2*	
32	*Late congenital syphilitic oculopathy*	*A50.3*	
33	*Late congenital neurosyphilis [juvenile neurosyphilis]*	*A50.4*	
34	*Other late congenital syphilis, symptomatic*	*A50.5*	
35	*Late congenital syphilis, latent*	*A50.6*	
36	*Late congenital syphilis, unspecified*	*A50.7*	
37	*Congenital syphilis, unspecified*	*A50.9*	
38	*Initial lesions of yaws*	*A66.0*	
39	*Multiple papillomata and wet crab yaws*	*A66.1*	
40	*Other early skin lesions of yaws*	*A66.2*	
41	*Hyperkeratosis of yaws*	*A66.3*	
42	*Gummata and ulcers of yaws*	*A66.4*	
43	*Gangosa*	*A66.5*	
44	*Bone and joint lesions of yaws*	*A66.6*	
45	*Other manifestations of yaws*	*A66.7*	
46	*Latent yaws*	*A66.8*	
47	*Yaws, unspecified*	*A66.9*	

48	Initial stage of trachoma	A71.0	
49	Active stage of trachoma	A71.1	
50	Trachoma, unspecified	A71.9	
51	Epidemic louse-borne typhus fever due to Rickettsia prowazekii	A75.0	
52	Recrudescent typhus [Brill's disease]	A75.1	
53	Typhus fever due to Rickettsia typhi	A75.2	
54	Typhus fever due to Rickettsia tsutsugamushi	A75.3	
55	Typhus fever, unspecified	A75.9	
56	Acute paralytic poliomyelitis, vaccine-associated	A80.0	
57	Acute paralytic poliomyelitis, wild virus, imported	A80.1	
58	Acute paralytic poliomyelitis, wild virus, indigenous	A80.2	
59	Acute paralytic poliomyelitis, other and unspecified	A80.3	
60	Acute nonparalytic poliomyelitis	A80.4	
61	Acute poliomyelitis, unspecified	A80.9	
62	Chikungunya virus disease	A92.0	
63	O'nyong-nyong fever	A92.1	
64	Venezuelan equine fever	A92.2	
65	West Nile fever	A92.3	
66	Rift Valley fever	A92.4	
67	Other specified mosquito-borne viral fevers	A92.8	
68	Mosquito-borne viral fever, unspecified	A92.9	
69	Eczema herpeticum	B00.0	
70	Herpesviral vesicular dermatitis	B00.1	
71	Herpesviralgingivostomatitis and pharyngotonsillitis	B00.2	
72	Herpesviral meningitis	B00.3	
73	Herpesviral encephalitis	B00.4	
74	Herpesviral ocular disease	B00.5	
75	Disseminated herpesviral disease	B00.7	

76	*Other forms of herpesviral infection*	*B00.8*	
77	*Herpesviral infection, unspecified*	*B00.9*	
78	*Hepatitis A with hepatic coma*	*B15.0*	
79	*Herpesviral infection, unspecified*	*B15.9*	
80	*HIV disease resulting in mycobacterial infection*	*B20.0*	
81	*HIV disease resulting in other bacterial infections*	*B20.1*	
82	*HIV disease resulting in cytomegaloviral disease*	*B20.2*	
83	*HIV disease resulting in other viral infections*	*B20.3*	
84	*HIV disease resulting in candidiasis*	*B20.4*	
85	*HIV disease resulting in other mycoses*	*B20.5*	
86	*HIV disease resulting in Pneumocystis carinii pneumonia*	*B20.6*	
87	*HIV disease resulting in multiple infections*	*B20.7*	
88	*HIV disease resulting in other infectious and parasitic diseases*	*B20.8*	
89	*HIV disease resulting in unspecified infectious or parasitic disease*	*B20.9*	
90	*Cytomegaloviral pneumonitis*	*B25.0*	
91	*Cytomegaloviral hepatitis*	*B25.1*	
92	*Cytomegaloviral pancreatitis*	*B25.2*	
93	*Other cytomegaloviral diseases*	*B25.8*	
94	*Cytomegaloviral disease, unspecified*	*B25.9*	
95	*Tineabarbae and tineacapitis*	*B35.0*	
96	*Tineaunguium*	*B35.1*	
97	*Tineamanuum*	*B35.2*	
98	*Tineapedis*	*B35.3*	
99	*Tineacorporis*	*B35.4*	
100	*Tinea imbricate*	*B35.5*	
10	*Tineacruris*	*B35.6*	

1			
102	Other dermatophytoses	B35.8	
103	Dermatophytosis, unspecified	B35.9	
104	Plasmodium falciparum malaria with cerebral complications	B50.0	
105	Other severe and complicated Plasmodium falciparum malaria	B50.8	
106	Plasmodium falciparum malaria, unspecified	B50.9	
107	Schistosomiasis due to Schistosomahaematobium [urinary schistosomiasis]	B65.0	
108	Schistosomiasis due to Schistosomamansoni [intestinal schistosomiasis]	B65.1	
109	Schistosomiasis due to Schistosomajaponicum	B65.2	
110	Cercarial dermatitis	B65.3	
111	Other schistosomiases	B65.8	
112	Schistosomiasis, unspecified	B65.9	
113	Pediculosis due to Pediculushumanuscapitis	B85.0	
114	Pediculosis due to Pediculushumanuscorporis	B85.1	
115	Pediculosis, unspecified	B85.2	

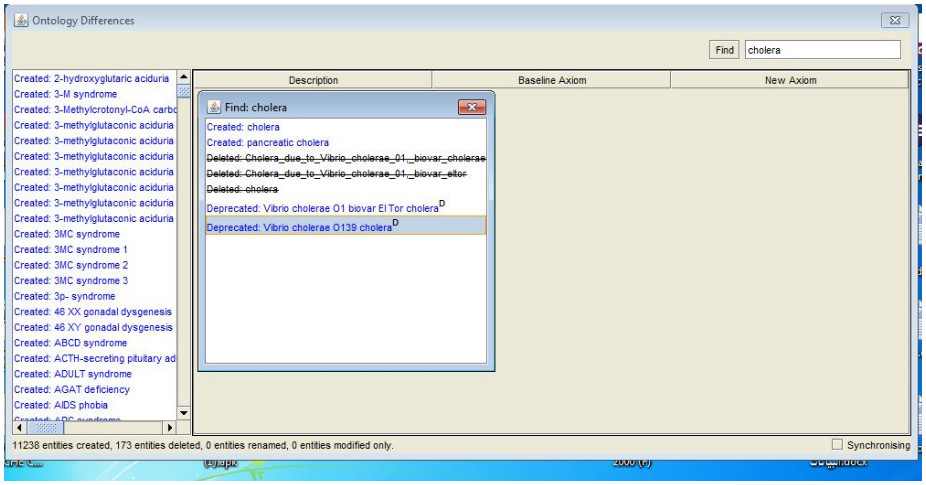
116	*Phthiriasis*	*B85.3*	
117	*Mixed pediculosis and phthiriasis*	*B85.4*	
118	*Sequelae of central nervous system tuberculosis*	*B90.0*	
119	*Sequelae of genito-urinary tuberculosis*	*B90.1*	
120	*Sequelae of tuberculosis of bones and joints*	*B90.2*	
121	*Sequelae of tuberculosis of other organs*	*B90.8*	
122	*Sequelae of respiratory and unspecified tuberculos*	*B90.9*	
123	*Streptococcus, group A, as the cause of diseases classified to other chapters*	*B95.0*	
124	*Streptococcus, group B, as the cause of diseases classified to other chapters*	*B95.1*	
125	*Streptococcus, group D, as the cause of diseases classified to other chapters*	*B95.2*	
126	*Streptococcus pneumoniae as the cause of diseases classified to other chapters*	*B95.3*	
127	*Other streptococcus as the cause of diseases classified to other chapters*	*B95.4*	
128	*Unspecified streptococcus as the cause of diseases classified to other chapters*	*B95.5*	
129	*Staphylococcus aureus as the cause of diseases classified to other chapters*	*B95.6*	
13	*Other staphylococcus as the cause of diseases classified to*	*B95.7*	

0	*other chapters*		
131	*Unspecified staphylococcus as the cause of diseases classified to other chapters*	*B95.8*	

ID	**Class Name**	**Code in ICD-10**	**Note**
1	*Intestinal infectious diseases*	A00_A09	
2	*Tuberculosis*	A15_A19	
3	*Certain zoonotic bacterial diseases*	A20_A28	
4	*Other bacterial diseases*	A30_A49	
5	*Infections with a predominantly sexual mode of transmission*	A50_A64	
6	*Other spirochaetal diseases*	A65_A69	
7	*Other diseases caused by chlamydiae*	A70_A74	
8	*Rickettsioses*	A75_A79	
9	*Viral diseases of the central nervous system*	A80_A89	
10	*Arthropod-borne viral fevers and viral haemorrhagic fevers*	A90_A99	
11	*Viral infections characterised by skin and mucous membrane lesions*	B00_B09	
12	*Viral hepatitis*	B15_B19	
13	*Human immunodeficiency virus [HIV] disease*	B20_B24	
14	*Other viral diseases*	B25_B34	
15	*Mycoses*	B35_B49	
16	*Protozoal diseases*	B50_B64	
17	*Helminthiases,*	B65_B83	
18	*Pediculosis, acariasis and other infestations,*	B85_B89	
19	*Sequelae of infectious and parasitic diseases,*	B90_B94	
20	*Bacterial, viral and other infectious agents*	B95_B97	

Appendix B

We compare between two ontologies (*diod* ontology and our reference dataset)using protégé tool.

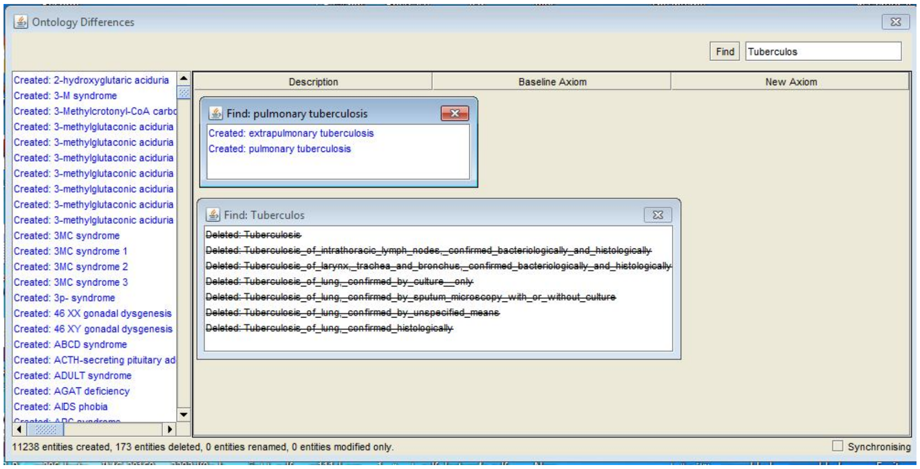


Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia

Created: cholera
Created: pancreatic cholera
~~Deleted: Cholera_due_to_Vibrio_cholerae_01,_biovar_cholerae~~
~~Deleted: Cholera_due_to_Vibrio_cholerae_01,_biovar_eltor~~
~~Deleted: cholera~~
Deprecated: Vibrio cholerae O1 biovar El Tor cholera D
Deprecated: Vibrio cholerae O139 cholera D

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

	Concept1(Class) doid Ontology	**Code ICD-10**	**Concept2 (Class) Our Dataset**	**CodeICD-10**
1	Vibrio Cholera O139, Cholera		Cholera due to Vibrio cholerae 01, biovarcholerae	*A00.0*
2	Vibrio Cholera choleraeO1, biovareltor Cholera		Cholera due to Vibrio cholerae 01, biovareltor	A00.1
3	Cholera		Cholera, unspecified	*A00.9*



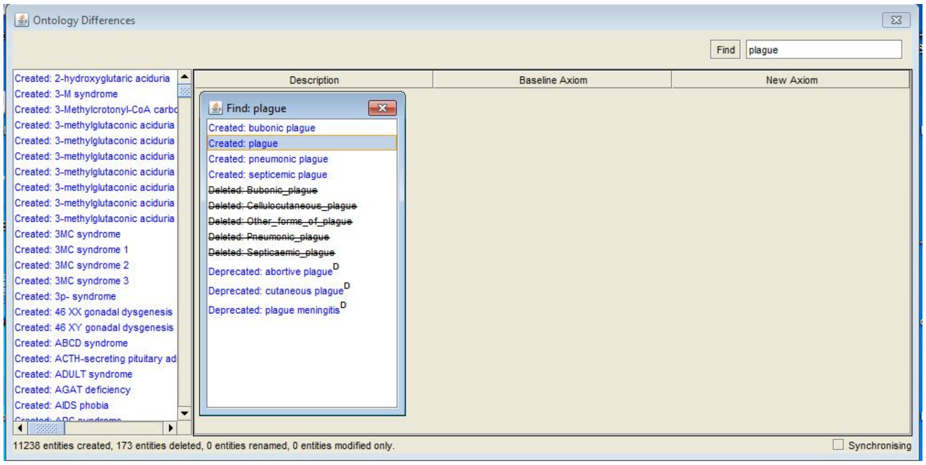
Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia

Created: extrapulmonary tuberculosis
Created: pulmonary tuberculosis

~~Deleted: Tuberculosis~~
~~Deleted: Tuberculosis_of_intrathoracic_lymph_nodes,_confirmed_bacteriologically_and_histologically~~
~~Deleted: Tuberculosis_of_larynx,_trachea_and_bronchus,_confirmed_bacteriologically_and_histologically~~
~~Deleted: Tuberculosis_of_lung,_confirmed_by_culture__only~~
~~Deleted: Tuberculosis_of_lung,_confirmed_by_sputum_microscopy_with_or_without_culture~~
~~Deleted: Tuberculosis_of_lung,_confirmed_by_unspecified_means~~
~~Deleted: Tuberculosis_of_lung,_confirmed_histologically~~

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

ID	***Concept1(Class) doid Ontology***		***Concept1(Class)*** reference dataset ***(Our dataset)***	***ICD10-Code***
4	pulmonary tuberculosis		Tuberculosis of lung, confirmed by sputum microscopy with or without culture	*A15.0*
5			***Tuberculosis of lung, confirmed by culture only***	***A15.1***
6			***Tuberculosis of lung, confirmed histologically***	***A15.2***
7			***Tuberculosis of lung, confirmed by unspecified means***	***A15.3***
8			***Tuberculosis of intrathoracic lymph nodes, confirmed bacteriologically and histologically***	***A15.4***
9			***Tuberculosis of larynx, trachea and bronchus, confirmed bacteriologically and histologically***	***A15.5***
10	pulmonary tuberculosis		*Tuberculous pleurisy, confirmed bacteriologically and histologically*	*A15.6*
11			***Primary respiratory tuberculosis, confirmed bacteriologically and histologically***	***A15.7***
12			***Other respiratory tuberculosis, confirmed bacteriologically and histologically***	***A15.8***
13			***Respiratory tuberculosis unspecified, confirmed bacteriologically and histologically***	***A15.9***

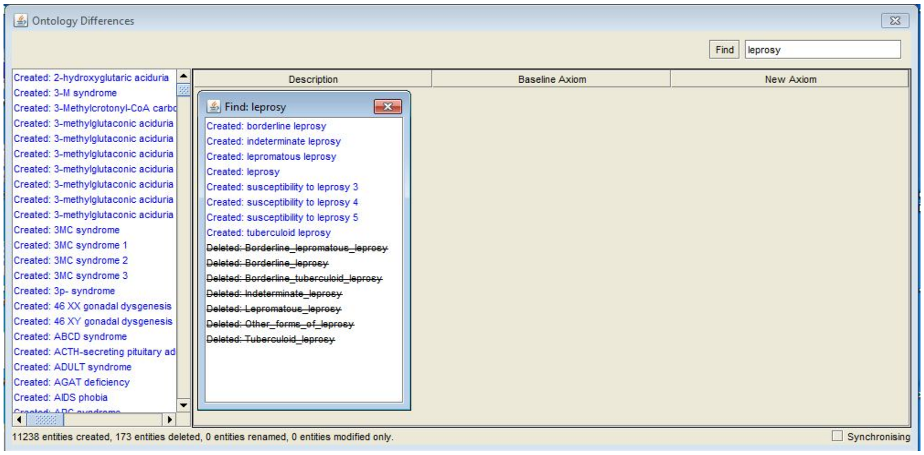


Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia

Created: bubonic plague
Created: plague
Created: pneumonic plague
Created: septicemic plague
~~Deleted: Bubonic_plague~~
~~Deleted: Cellulocutaneous_plague~~
~~Deleted: Other_forms_of_plague~~
~~Deleted: Pneumonic_plague~~
~~Deleted: Septicaemic_plague~~
Deprecated: abortive plague^D
Deprecated: cutaneous plague^D
Deprecated: plague meningitis^D

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

ID	***Concept1(Class) doid Ontology***		***Concept2(Class)*** reference dataset ***(Our dataset)***	***ICD10-Code***
14	bubonic plague		Bubonic plague	A20.0
15			Cellulocutaneous plague	A20.1
16	pneumonic plague		Pneumonic plague	A20.2
17			Plague meningitis	A20.3
18	septicaemic plague		Septicaemic plague	A20.7
19			Other forms of plague	A20.8
20	Plague		Plague, unspecified	A20.9

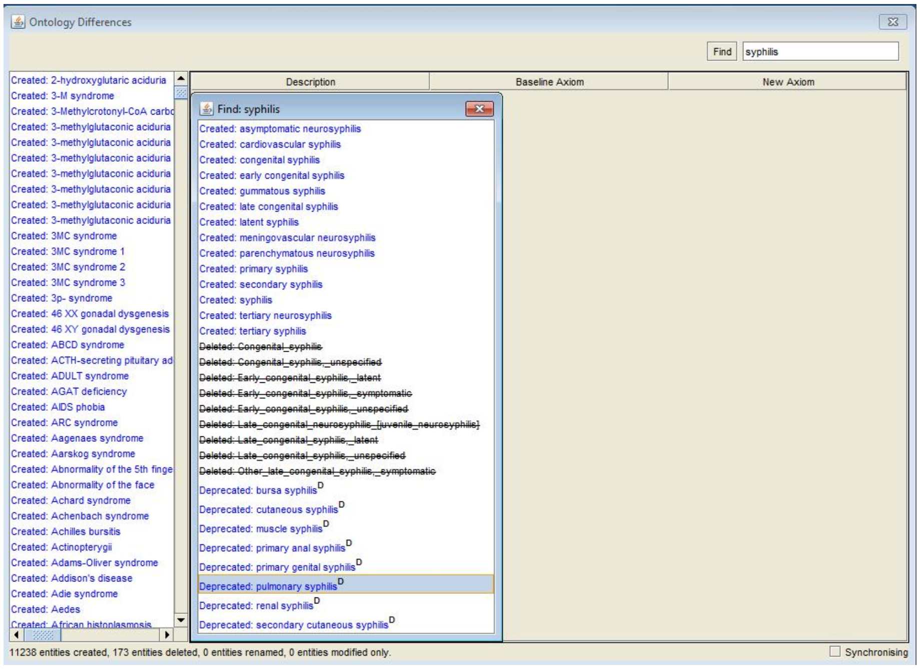


Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia

Created: borderline leprosy
Created: indeterminate leprosy
Created: lepromatous leprosy
Created: leprosy
Created: susceptibility to leprosy 3
Created: susceptibility to leprosy 4
Created: susceptibility to leprosy 5
Created: tuberculoid leprosy
~~Deleted: Borderline_lepromatous_leprosy~~
~~Deleted: Borderline_leprosy~~
~~Deleted: Borderline_tuberculoid_leprosy~~
~~Deleted: Indeterminate_leprosy~~
~~Deleted: Lepromatous_leprosy~~
~~Deleted: Other_forms_of_leprosy~~
~~Deleted: Tuberculoid_leprosy~~

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

ID	**Concept1(Class) doid Ontology**		**Concept2(Class)** reference dataset **(Our dataset)**	**ICD10-Code**
21	Indeterminate leprosy		Indeterminate leprosy	A30.0
22	Tuberculoid leprosy		Tuberculoid leprosy	A30.1
23			Borderline tuberculoid leprosy	A30.2
24	borderline leprosy		Borderline leprosy	A30.3
25			Borderline lepromatous leprosy	A30.4
26	Lepromatous leprosy		Lepromatous leprosy	A30.5
27			Other forms of leprosy	A30.8
28	Leprosy		Leprosy, unspecified	A30.9

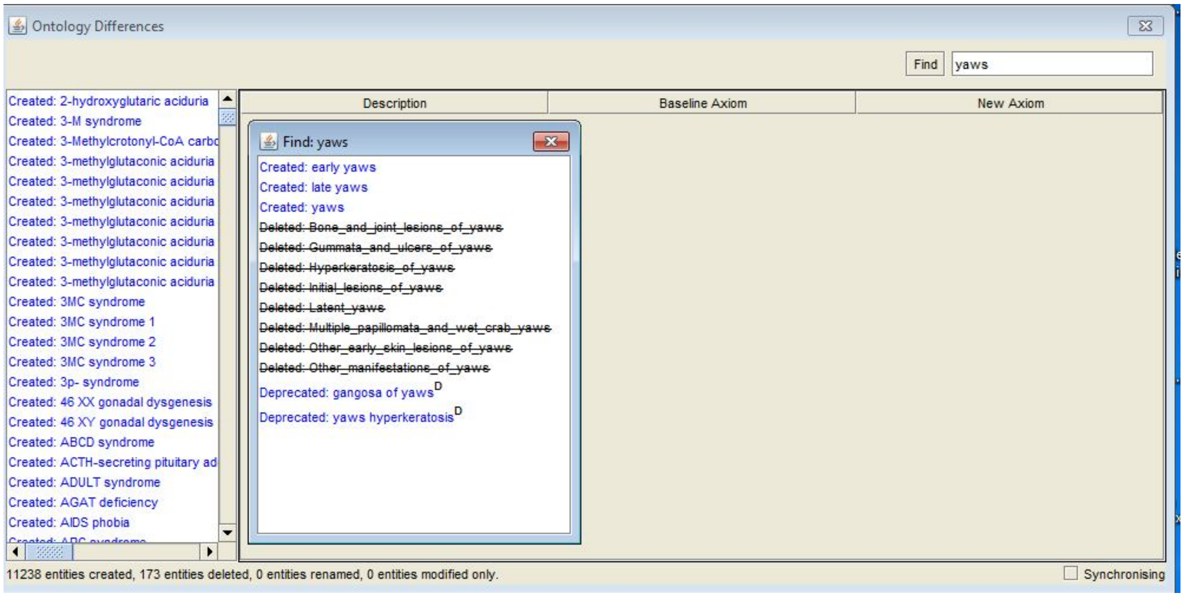


Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome
Created: Aarskog syndrome
Created: Abnormality of the 5th finge
Created: Abnormality of the face
Created: Achard syndrome
Created: Achenbach syndrome
Created: Achilles bursitis
Created: Actinopterygii
Created: Adams-Oliver syndrome
Created: Addison's disease
Created: Adie syndrome
Created: Aedes
Created: African histoplasmosis

Created: asymptomatic neurosyphilis
Created: cardiovascular syphilis
Created: congenital syphilis
Created: early congenital syphilis
Created: gummatous syphilis
Created: late congenital syphilis
Created: latent syphilis
Created: meningovascular neurosyphilis
Created: parenchymatous neurosyphilis
Created: primary syphilis
Created: secondary syphilis
Created: syphilis
Created: tertiary neurosyphilis
Created: tertiary syphilis
~~Deleted: Congenital_syphilis~~
~~Deleted: Congenital_syphilis,_unspecified~~
~~Deleted: Early_congenital_syphilis,_latent~~
~~Deleted: Early_congenital_syphilis,_symptomatic~~
~~Deleted: Early_congenital_syphilis,_unspecified~~
~~Deleted: Late_congenital_neurosyphilis_[juvenile_neurosyphilis]~~
~~Deleted: Late_congenital_syphilis,_latent~~
~~Deleted: Late_congenital_syphilis,_unspecified~~
~~Deleted: Other_late_congenital_syphilis,_symptomatic~~
Deprecated: bursa syphilis^{D}
Deprecated: cutaneous syphilis^{D}
Deprecated: muscle syphilis^{D}
Deprecated: primary anal syphilis^{D}
Deprecated: primary genital syphilis^{D}
Deprecated: pulmonary syphilis^{D}
Deprecated: renal syphilis^{D}
Deprecated: secondary cutaneous syphilis^{D}

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

ID	**Concept1(Class) doid Ontology**		**Concept2(Class)** reference dataset **(Our dataset)**	**ICD10-Code**
29	Early congenital syphilis		Early congenital syphilis, symptomatic	A50.0
30			Early congenital syphilis, latent	A50.1
31			Early congenital syphilis, unspecified	A50.2
32			Late congenital syphilitic oculopathy	A50.3
33	late congenital syphilis		Late congenital neurosyphilis [juvenile neurosyphilis]	A50.4
34	late congenital syphilis		Other late congenital syphilis, symptomatic	A50.5
35	Late congenital syphilis		Late congenital syphilis, latent	A50.6
36			Late congenital syphilis, unspecified	A50.7
37	Congenital syphilis		Congenital syphilis, unspecified	A50.9



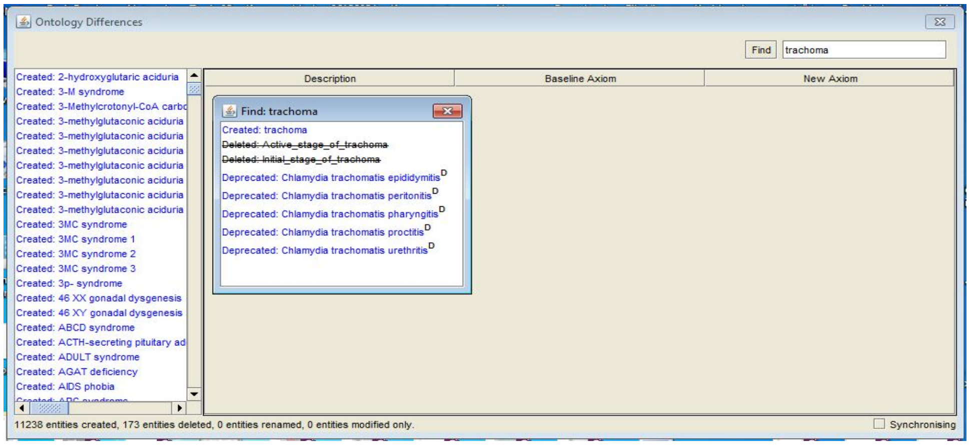
Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ADC syndrome

Created: early yaws
Created: late yaws
Created: yaws
~~Deleted: Bone_and_joint_lesions_of_yaws~~
~~Deleted: Gummata_and_ulcers_of_yaws~~
~~Deleted: Hyperkeratosis_of_yaws~~
~~Deleted: Initial_lesions_of_yaws~~
~~Deleted: Latent_yaws~~
~~Deleted: Multiple_papillomata_and_wet_crab_yaws~~
~~Deleted: Other_early_skin_lesions_of_yaws~~
~~Deleted: Other_manifestations_of_yaws~~
Deprecated: gangosa of yaws^{D}
Deprecated: yaws hyperkeratosis^{D}

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

ID	**Concept1(Class) doid Ontology**		**Concept2(Class)** reference dataset **(Our dataset)**	**ICD10-Code**
38	*early yaws*		*Initial lesions of yaws*	*A66.0*
39	*Late yaws*		*Multiple papillomata and wet crab yaws*	*A66.1*
40			***Other early skin lesions of yaws***	***A66.2***
41			***Hyperkeratosis of yaws***	***A66.3***
42	late yaws		*Gummata and ulcers of yaws*	*A66.4*
43	gangosa of yaws		*Gangosa*	*A66.5*
44	early yaws		*Bone and joint lesions of yaws*	*A66.6*
45			***Other manifestations of yaws***	***A66.7***
46	Late yaws		Latent yaws	A66.8
47			*Yaws, unspecified*	*A66.9*

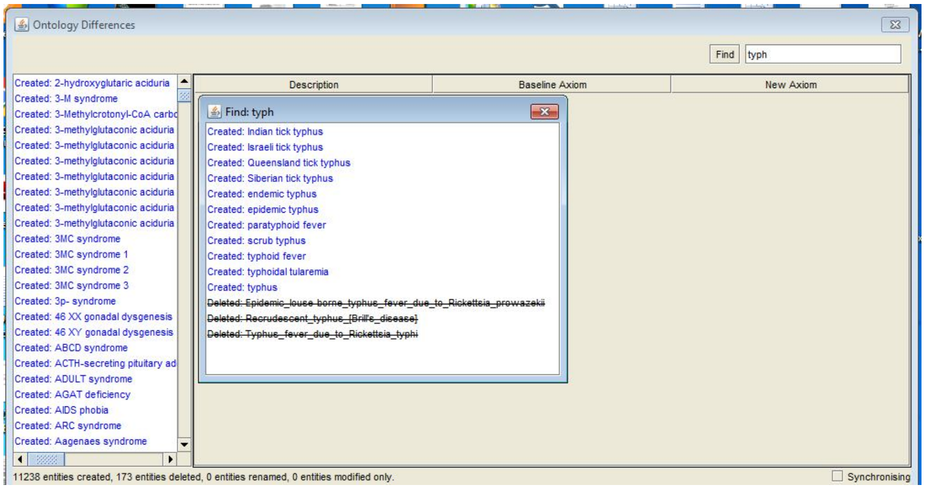
48			*Initial stage of trachoma*	*A71.0*
49			*Active stage of trachoma*	*A71.1*
50	*Trachoma*		*Trachoma, unspecified*	*A71.9*



Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ABC syndrome

Created: trachoma
~~Deleted: Active_stage_of_trachoma~~
~~Deleted: Initial_stage_of_trachoma~~
Deprecated: Chlamydia trachomatis epididymitis^{D}
Deprecated: Chlamydia trachomatis peritonitis^{D}
Deprecated: Chlamydia trachomatis pharyngitis^{D}
Deprecated: Chlamydia trachomatis proctitis^{D}
Deprecated: Chlamydia trachomatis urethritis^{D}

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

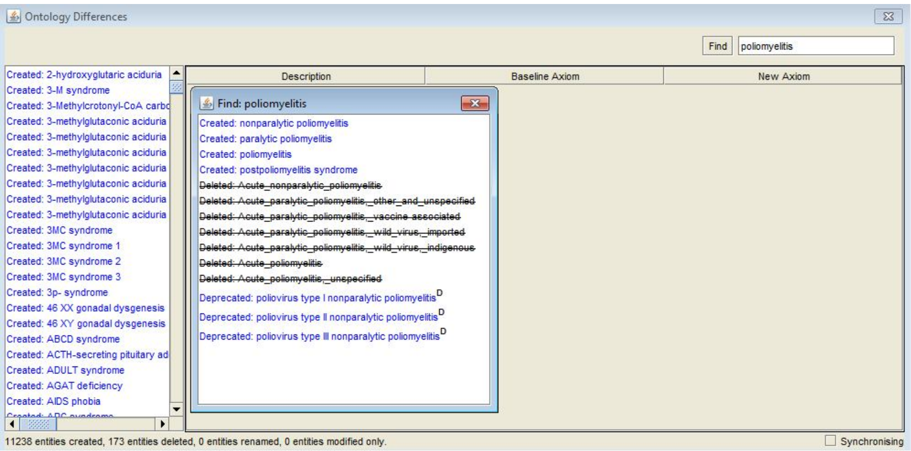


Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome

Created: Indian tick typhus
Created: Israeli tick typhus
Created: Queensland tick typhus
Created: Siberian tick typhus
Created: endemic typhus
Created: epidemic typhus
Created: paratyphoid fever
Created: scrub typhus
Created: typhoid fever
Created: typhoidal tularemia
Created: typhus
~~Deleted: Epidemic_louse-borne_typhus_fever_due_to_Rickettsia_prowazekii~~
~~Deleted: Recrudescent_typhus_[Brill's_disease]~~
~~Deleted: Typhus_fever_due_to_Rickettsia_typhi~~

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

51	*Typhus*		*Epidemic louse-borne typhus fever due to Rickettsia prowazekii*	*A75.0*
52	Brill-Zinsser disease		*Recrudescent typhus [Brill's disease]*	*A75.1*
53	Typhus		*Typhus fever due to Rickettsia typhi*	*A75.2*
54	scrub typhus		*Typhus fever due to Rickettsia tsutsugamushi*	*A75.3*
55	Typhus		*Typhus fever, unspecified*	*A75.9*



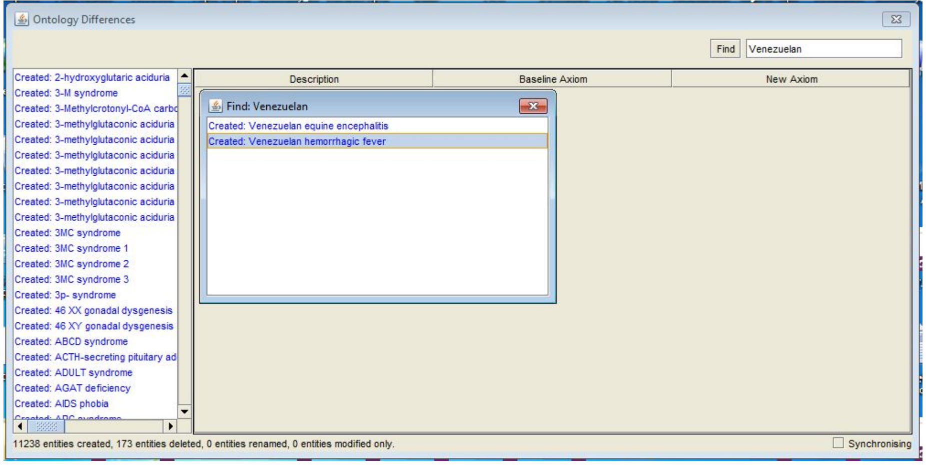
Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ABC syndrome

Find: poliomyelitis

Created: nonparalytic poliomyelitis
Created: paralytic poliomyelitis
Created: poliomyelitis
Created: postpoliomyelitis syndrome
~~Deleted: Acute_nonparalytic_poliomyelitis~~
~~Deleted: Acute_paralytic_poliomyelitis,_other_and_unspecified~~
~~Deleted: Acute_paralytic_poliomyelitis,_vaccine associated~~
~~Deleted: Acute_paralytic_poliomyelitis,_wild_virus,_imported~~
~~Deleted: Acute_paralytic_poliomyelitis,_wild_virus,_indigenous~~
~~Deleted: Acute_poliomyelitis~~
~~Deleted: Acute_poliomyelitis,_unspecified~~
Deprecated: poliovirus type I nonparalytic poliomyelitis^{D}
Deprecated: poliovirus type II nonparalytic poliomyelitis^{D}
Deprecated: poliovirus type III nonparalytic poliomyelitis^{D}

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

56			Acute paralytic poliomyelitis, vaccine-associated	A80.0
57			Acute paralytic poliomyelitis, wild virus, imported	A80.1
58			Acute paralytic poliomyelitis, wild virus, indigenous	A80.2
59			Acute paralytic poliomyelitis, other and unspecified	A80.3
60	nonparalytic poliomyelitis		Acute nonparalytic poliomyelitis	A80.4
61	Poliomyelitis		Acute poliomyelitis, unspecified	A80.9



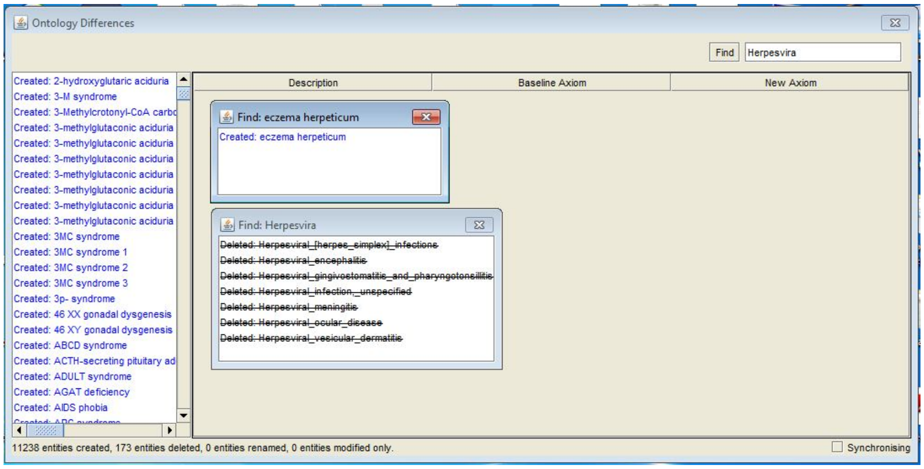
Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ABC syndrome

Find: Venezuelan

Created: Venezuelan equine encephalitis
Created: Venezuelan hemorrhagic fever

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

62	*Chikungunya*		*Chikungunya virus disease*	*A92.0*
63	O'nyong'nyong fever		*O'nyong-nyong fever*	*A92.1*
64	Venezuelan equine encephalitis		*Venezuelan equine fever*	*A92.2*
65	West Nile fever		*West Nile fever*	*A92.3*
66	Rift Valley fever		*Rift Valley fever*	*A92.4*
67	Zika fever		*Other specified mosquito-borne viral fevers*	*A92.8*
68			***Mosquito-borne viral fever, unspecified***	***A92.9***



Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ABC syndrome

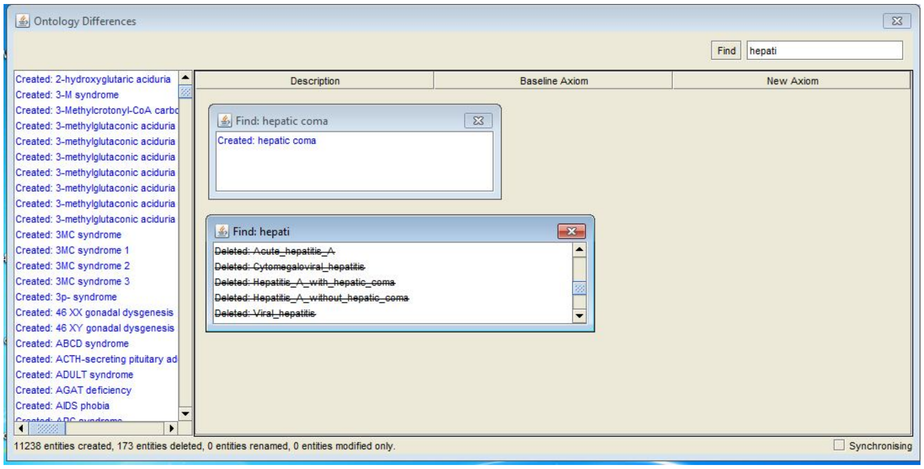
Created: eczema herpeticum

~~Deleted: Herpesviral_[herpes_simplex]_infections~~
~~Deleted: Herpesviral_encephalitis~~
~~Deleted: Herpesviral_gingivostomatitis_and_pharyngotonsillitis~~
~~Deleted: Herpesviral_infection,_unspecified~~
~~Deleted: Herpesviral_meningitis~~
~~Deleted: Herpesviral_ocular_disease~~
~~Deleted: Herpesviral_vesicular_dermatitis~~

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

69	*eczema herpeticum*		*Eczema herpeticum*	*B00.0*
70			***Herpesviral vesicular dermatitis***	***B00.1***
71			***Herpesviralgingivostomatitis and pharyngotonsillitis***	***B00.2***
72			***Herpesviral meningitis***	***B00.3***

73			***Herpesviral encephalitis***	***B00.4***
74			***Herpesviral ocular disease***	***B00.5***
75			***Disseminated herpesviral disease***	***B00.7***
76			***Other forms of herpesviral infection***	***B00.8***
77	herpes simplex		*Herpesviral infection, unspecified*	*B00.9*



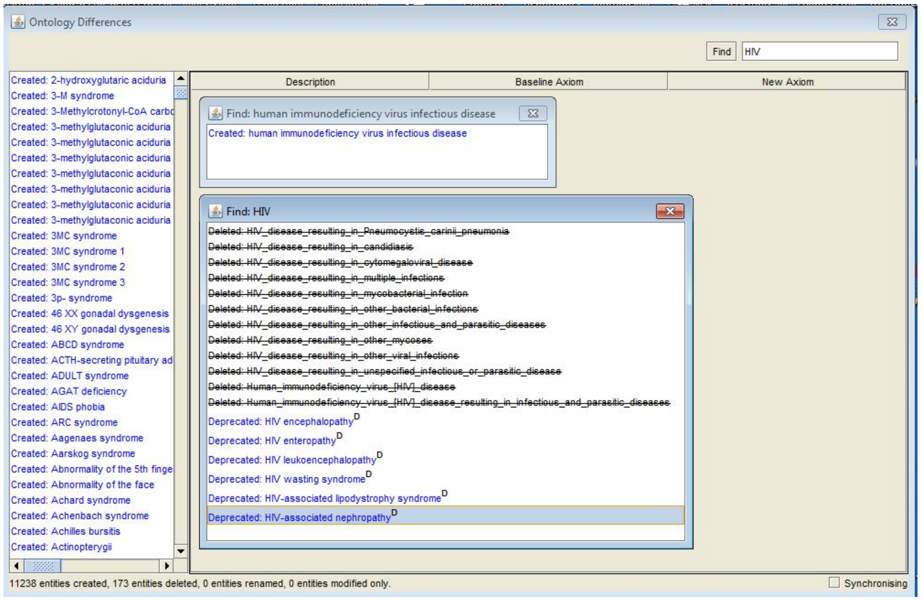
Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ADC syndrome

Created: hepatic coma

~~Deleted: Acute_hepatitis_A~~
~~Deleted: Cytomegaloviral_hepatitis~~
~~Deleted: Hepatitis_A_with_hepatic_coma~~
~~Deleted: Hepatitis_A_without_hepatic_coma~~
~~Deleted: Viral_hepatitis~~

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

78	*hepatic coma*		*Hepatitis A with hepatic coma*	*B15.0*
79			***Hepatitis A without hepatic coma***	***B15.9***



Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome
Created: Aarskog syndrome
Created: Abnormality of the 5th finge
Created: Abnormality of the face
Created: Achard syndrome
Created: Achenbach syndrome
Created: Achilles bursitis
Created: Actinopterygii

Find: human immunodeficiency virus infectious disease

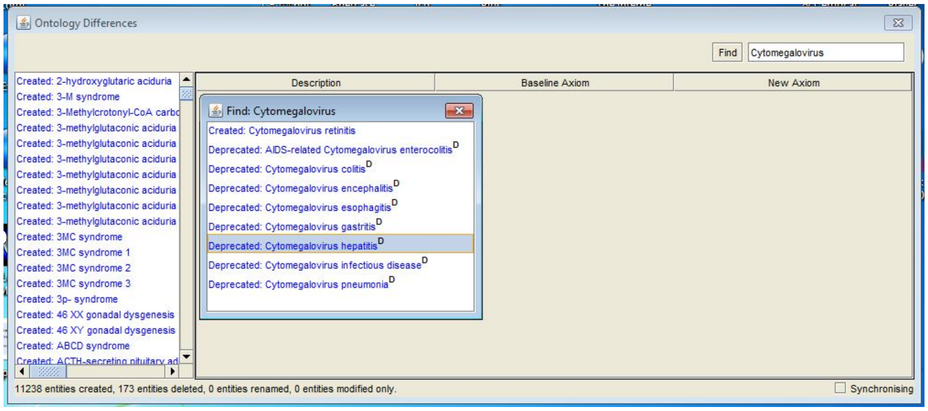
Created: human immunodeficiency virus infectious disease

Find: HIV

~~Deleted: HIV_disease_resulting_in_Pneumocystis_carinii_pneumonia~~
~~Deleted: HIV_disease_resulting_in_candidiasis~~
~~Deleted: HIV_disease_resulting_in_cytomegaloviral_disease~~
~~Deleted: HIV_disease_resulting_in_multiple_infections~~
~~Deleted: HIV_disease_resulting_in_mycobacterial_infection~~
~~Deleted: HIV_disease_resulting_in_other_bacterial_infections~~
~~Deleted: HIV_disease_resulting_in_other_infectious_and_parasitic_diseases~~
~~Deleted: HIV_disease_resulting_in_other_mycoses~~
~~Deleted: HIV_disease_resulting_in_other_viral_infections~~
~~Deleted: HIV_disease_resulting_in_unspecified_infectious_or_parasitic_disease~~
~~Deleted: Human_immunodeficiency_virus_[HIV]_disease~~
~~Deleted: Human_immunodeficiency_virus_[HIV]_disease_resulting_in_infectious_and_parasitic_diseases~~
Deprecated: HIV encephalopathy^{D}
Deprecated: HIV enteropathy^{D}
Deprecated: HIV leukoencephalopathy^{D}
Deprecated: HIV wasting syndrome^{D}
Deprecated: HIV-associated lipodystrophy syndrome^{D}
Deprecated: HIV-associated nephropathy^{D}

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

80	*human immunodeficiency virus infectious disease*		*HIV disease resulting in mycobacterial infection*	*B20.0*
81			***HIV disease resulting in other bacterial infections***	***B20.1***
82			***HIV disease resulting in cytomegaloviral disease***	***B20.2***
83			***HIV disease resulting in other viral infections***	***B20.3***
84			***HIV disease resulting in candidiasis***	***B20.4***
85			***HIV disease resulting in other mycoses***	***B20.5***
86			***HIV disease resulting in Pneumocystis carinii pneumonia***	***B20.6***
87			***HIV disease resulting in multiple infections***	***B20.7***
88			***HIV disease resulting in other infectious and parasitic diseases***	***B20.8***
89			***HIV disease resulting in unspecified infectious or parasitic disease***	***B20.9***

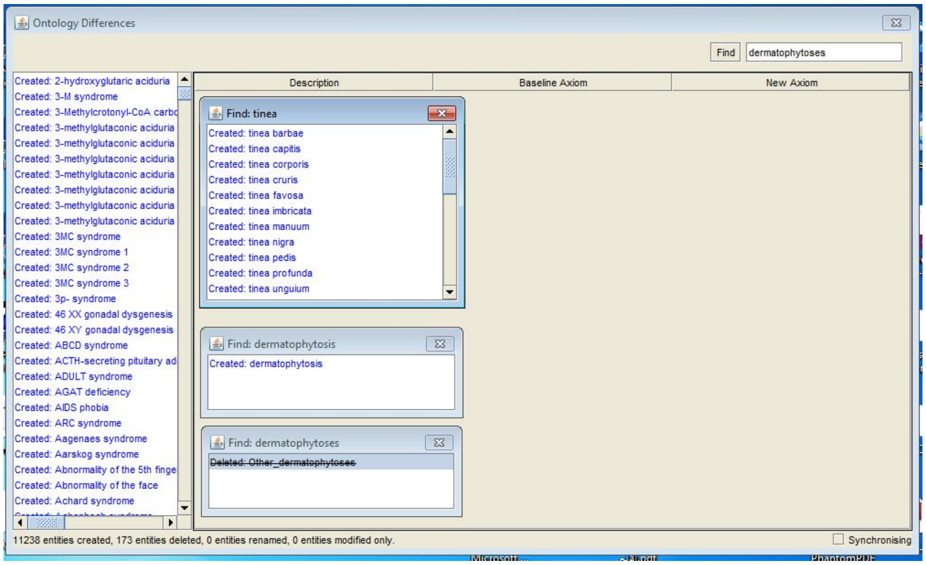


Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad

Created: Cytomegalovirus retinitis
Deprecated: AIDS-related Cytomegalovirus enterocolitis^{D}
Deprecated: Cytomegalovirus colitis^{D}
Deprecated: Cytomegalovirus encephalitis^{D}
Deprecated: Cytomegalovirus esophagitis^{D}
Deprecated: Cytomegalovirus gastritis^{D}
Deprecated: Cytomegalovirus hepatitis^{D}
Deprecated: Cytomegalovirus infectious disease^{D}
Deprecated: Cytomegalovirus pneumonia^{D}

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

90	*Cytomegalovirus pneumonia*	-	*Cytomegaloviral pneumonitis*	*B25.0*
91	*Cytomegalovirus hepatitis*	-	*Cytomegaloviral hepatitis*	*B25.1*
92			***Cytomegaloviral pancreatitis***	***B25.2***
93			***Other cytomegaloviral diseases***	***B25.8***
94			***Cytomegaloviral disease, unspecified***	***B25.9***



Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbc
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome
Created: Aarskog syndrome
Created: Abnormality of the 5th finge
Created: Abnormality of the face
Created: Achard syndrome

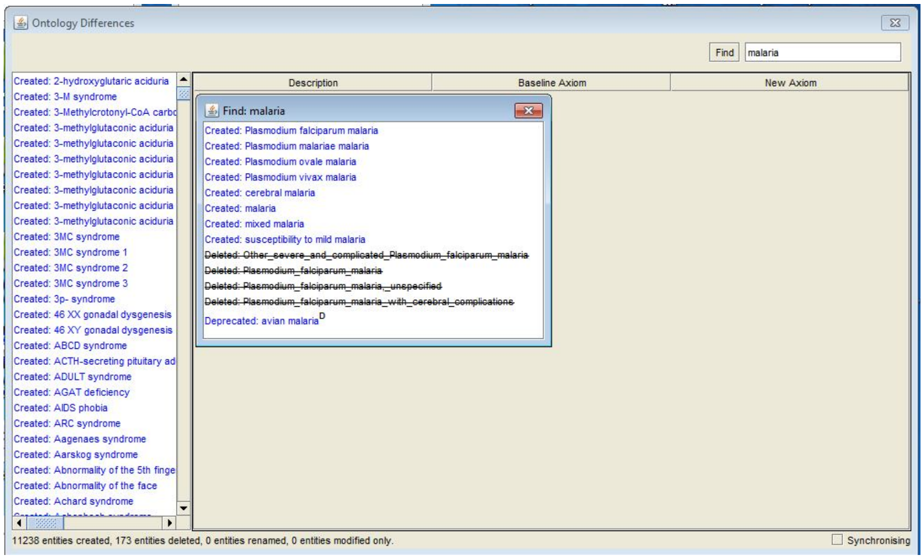
Created: tinea barbae
Created: tinea capitis
Created: tinea corporis
Created: tinea cruris
Created: tinea favosa
Created: tinea imbricata
Created: tinea manuum
Created: tinea nigra
Created: tinea pedis
Created: tinea profunda
Created: tinea unguium

Created: dermatophytosis

~~Deleted: Other_dermatophytoses~~

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

95	*tineabarbae, tineacapitis*	-	*Tineabarbae and tineacapitis*	*B35.0*
96	Tineaunguium	-	*Tineaunguium*	*B35.1*
97	Tineamanuum	*B35.2*	*Tineamanuum*	*B35.2*
98	Tineapedis	*B35.3*	*Tineapedis*	*B35.3*
99	Tineacorporis	-	*Tineacorporis*	*B35.4*
100	*Tineaimbricate*	-	*Tineaimbricate*	*B35.5*
101	Tineacruris	-	*Tineacruris*	*B35.6*
102			***Other dermatophytoses***	***B35.8***
103	Dermatophytosis	*B35,* *B35.9*	*Dermatophytosis, unspecified*	*B35.9*



Ontology Differences

Find malaria

Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome
Created: Aarskog syndrome
Created: Abnormality of the 5th finge
Created: Abnormality of the face
Created: Achard syndrome

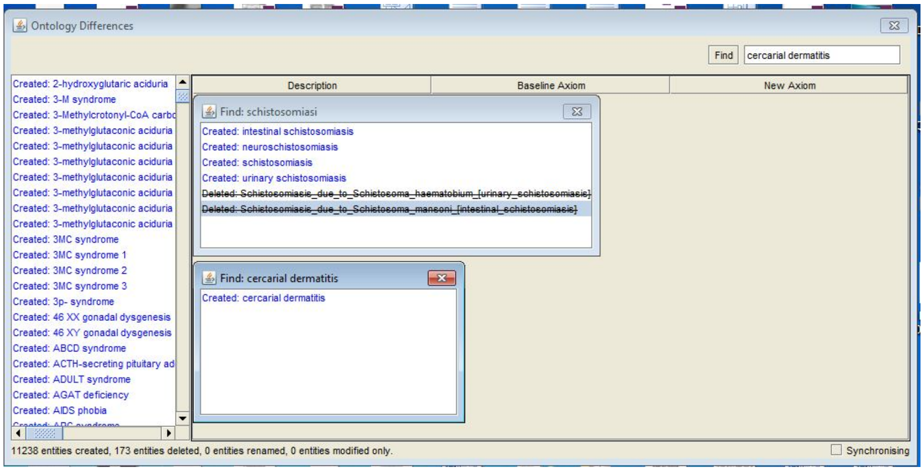
Description	Baseline Axiom	New Axiom

Find: malaria

Created: Plasmodium falciparum malaria
Created: Plasmodium malariae malaria
Created: Plasmodium ovale malaria
Created: Plasmodium vivax malaria
Created: cerebral malaria
Created: malaria
Created: mixed malaria
Created: susceptibility to mild malaria
~~Deleted: Other_severe_and_complicated_Plasmodium_falciparum_malaria~~
~~Deleted: Plasmodium_falciparum_malaria~~
~~Deleted: Plasmodium_falciparum_malaria,_unspecified~~
~~Deleted: Plasmodium_falciparum_malaria_with_cerebral_complications~~
Deprecated: avian malaria^D

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

104	*cerebral malaria*	*B50.0*	*Plasmodium falciparum malaria with cerebral complications*	*B50.0*
105			***Other severe and complicated Plasmodium falciparum malaria***	***B50.8***
106	Plasmodium falciparum malaria	*B50.9*	*Plasmodium falciparum malaria, unspecified*	*B50.9*



Ontology Differences

Find cercarial dermatitis

Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia

Description	Baseline Axiom	New Axiom

Find: schistosomiasi

Created: intestinal schistosomiasis
Created: neuroschistosomiasis
Created: schistosomiasis
Created: urinary schistosomiasis
~~Deleted: Schistosomiasis_due_to_Schistosoma_haematobium_[urinary_schistosomiasis]~~
~~Deleted: Schistosomiasis_due_to_Schistosoma_mansoni_[intestinal_schistosomiasis]~~

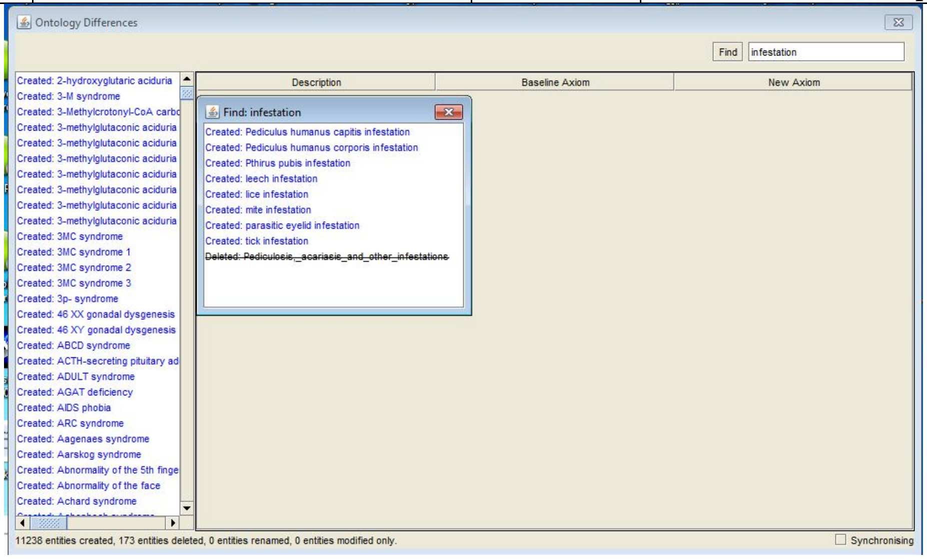
Find: cercarial dermatitis

Created: cercarial dermatitis

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

107	*urinary schistosomiasis*	-	*Schistosomiasis due to Schistosomahaematobium [urinary schistosomiasis]*	*B65.0*
108	intestinal schistosomiasis	*B65.1*	*Schistosomiasis due to Schistosomamansoni [intestinal*	*B65.1*

			schistosomiasis]	
109	intestinal schistosomiasis	*B65.2*	*Schistosomiasis due to Schistosomajaponicum*	*B65.2*
110	cercarial dermatitis	*B65.3*	*Cercarial dermatitis*	*B65.3*
111			***Other schistosomiases***	***B65.8***
112	*Schistosomiases*	-	*Schistosomiasis, unspecified*	*B65.9*

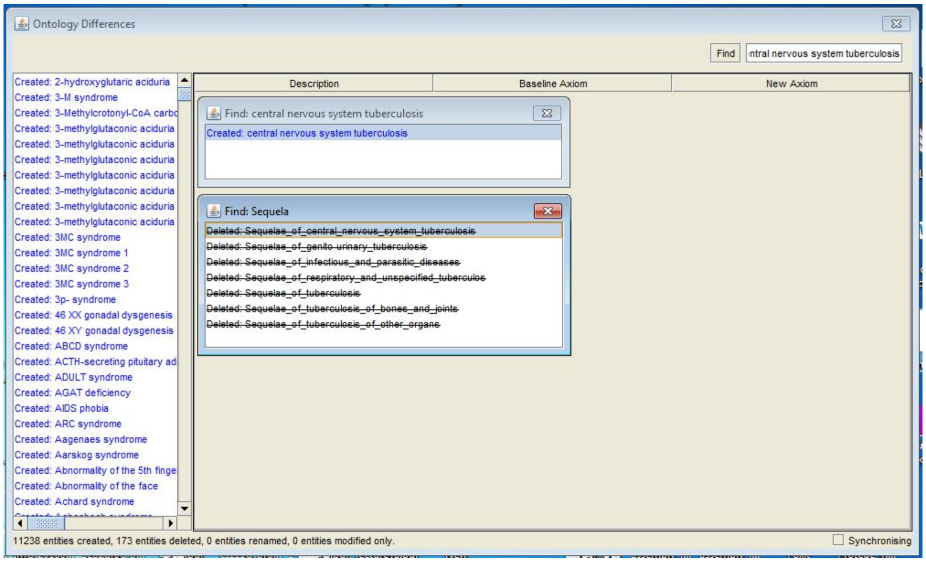


Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome
Created: Aarskog syndrome
Created: Abnormality of the 5th finge
Created: Abnormality of the face
Created: Achard syndrome

Created: Pediculus humanus capitis infestation
Created: Pediculus humanus corporis infestation
Created: Pthirus pubis infestation
Created: leech infestation
Created: lice infestation
Created: mite infestation
Created: parasitic eyelid infestation
Created: tick infestation
~~Deleted: Pediculosis,_acariasis_and_other_infestations~~

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

113	*Pediculushumanuscapitis infestation*	*B85.0*	*Pediculosis due to Pediculushumanuscapitis*	*B85.0*
114	*Pediculushumanuscapitis infestation*	*B85.1*	*Pediculosis due to Pediculushumanuscorporis*	*B85.1*
115	*lice infestation*	*B85.2*	*Pediculosis, unspecified*	*B85.2*
116	*Pthirus pubis infestation*	*B85.3*	*Phthiriasis*	*B85.3*
117			***Mixed pediculosis and phthiriasis***	***B85.4***



Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome
Created: Aarskog syndrome
Created: Abnormality of the 5th finge
Created: Abnormality of the face
Created: Achard syndrome

Find: central nervous system tuberculosis

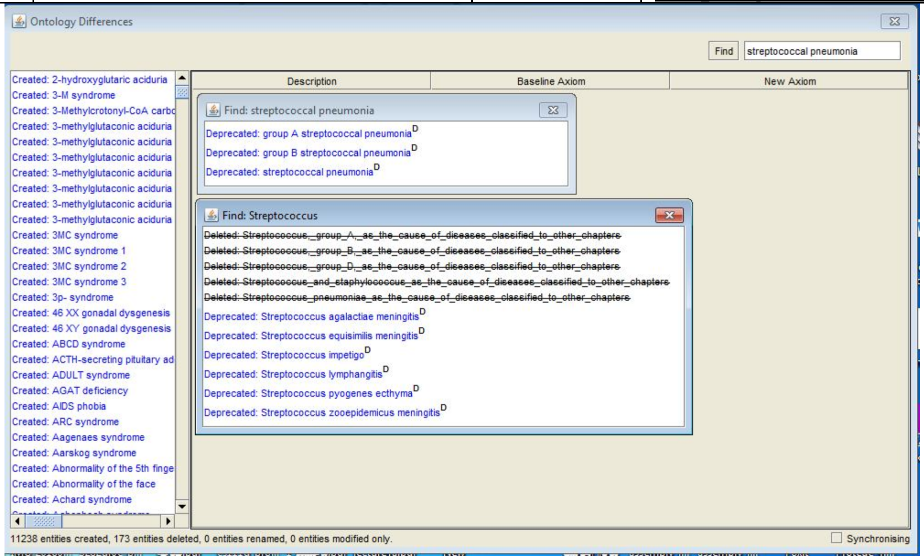
Created: central nervous system tuberculosis

Find: Sequela

~~Deleted: Sequelae_of_central_nervous_system_tuberculosis~~
~~Deleted: Sequelae_of_genito-urinary_tuberculosis~~
~~Deleted: Sequelae_of_infectious_and_parasitic_diseases~~
~~Deleted: Sequelae_of_respiratory_and_unspecified_tuberculos~~
~~Deleted: Sequelae_of_tuberculosis~~
~~Deleted: Sequelae_of_tuberculosis_of_bones_and_joints~~
~~Deleted: Sequelae_of_tuberculosis_of_other_organs~~

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

118	*central nervous system tuberculosis*	-	*Sequelae of central nervous system tuberculosis*	*B90.0*
119			***Sequelae of genito-urinary tuberculosis***	***B90.1***
120			***Sequelae of tuberculosis of bones and joints***	***B90.2***
121			***Sequelae of tuberculosis of other organs***	***B90.8***
122			***Sequelae of respiratory and unspecified tuberculos***	***B90.9***



Created: 2-hydroxyglutaric aciduria
Created: 3-M syndrome
Created: 3-Methylcrotonyl-CoA carbo
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3-methylglutaconic aciduria
Created: 3MC syndrome
Created: 3MC syndrome 1
Created: 3MC syndrome 2
Created: 3MC syndrome 3
Created: 3p- syndrome
Created: 46 XX gonadal dysgenesis
Created: 46 XY gonadal dysgenesis
Created: ABCD syndrome
Created: ACTH-secreting pituitary ad
Created: ADULT syndrome
Created: AGAT deficiency
Created: AIDS phobia
Created: ARC syndrome
Created: Aagenaes syndrome
Created: Aarskog syndrome
Created: Abnormality of the 5th finge
Created: Abnormality of the face
Created: Achard syndrome

Find: streptococcal pneumonia

Deprecated: group A streptococcal pneumonia D
Deprecated: group B streptococcal pneumonia D
Deprecated: streptococcal pneumonia D

Find: Streptococcus

~~Deleted: Streptococcus,_group_A,_as_the_cause_of_diseases_classified_to_other_chapters~~
~~Deleted: Streptococcus,_group_B,_as_the_cause_of_diseases_classified_to_other_chapters~~
~~Deleted: Streptococcus,_group_D,_as_the_cause_of_diseases_classified_to_other_chapters~~
~~Deleted: Streptococcus_and_staphylococcus_as_the_cause_of_diseases_classified_to_other_chapters~~
~~Deleted: Streptococcus_pneumoniae_as_the_cause_of_diseases_classified_to_other_chapters~~
Deprecated: Streptococcus agalactiae meningitis D
Deprecated: Streptococcus equisimilis meningitis D
Deprecated: Streptococcus impetigo D
Deprecated: Streptococcus lymphangitis D
Deprecated: Streptococcus pyogenes ecthyma D
Deprecated: Streptococcus zooepidemicus meningitis D

11238 entities created, 173 entities deleted, 0 entities renamed, 0 entities modified only.

123	*group A streptococcal pneumonia*	-	*Streptococcus, group A, as the cause of diseases classified to other chapters*	*B95.0*
124	*group B streptococcal pneumonia*	-	*Streptococcus, group B, as the cause of diseases classified to other chapters*	*B95.1*
125			***Streptococcus, group D, as the cause of diseases classified to other chapters***	***B95.2***
126			***Streptococcus pneumoniae as the cause of diseases classified to other chapters***	***B95.3***
127			***Other streptococcus as the cause of diseases classified to other chapters***	***B95.4***
128			***Unspecified streptococcus as the cause of diseases classified to other chapters***	***B95.5***
129			***Staphylococcus aureus as the cause of diseases classified to other chapters***	***B95.6***
130			***Other staphylococcus as the cause of diseases classified to other chapters***	***B95.7***
131			***Unspecified staphylococcus as the cause of diseases classified to other chapters***	***B95.8***

Appendix C

Step one: Using Chapter One, which contains 738 concepts, we compare all concepts with each other. For example, compare the concept *”A00.0”* with all the concepts in the chapter I, using SemDist Measure and then we select the minimum semantic similarity values and collect them in one group and call them (Cluster One). As shown in Table 3.1. After that, we delete 58 concepts from our experiment.

ID	**Concept1**	**Concept2**	**LCS (c1, c2)**	**Length(c1, c2)**	**CSPec(c1, c2)**	**SemDist(c1, c2)**
1	A00.0	A00.0	A00.0	1	0	0
2	A00.0	A00.1	A00	3	1	1.6
3	A00.0	A00.9	A00	3	1	1.6
4	A00.0	A01.0	A00_A09	5	2	3.2
5	A00.0	A01.1	A00_A09	5	2	3.2
6	A00.0	A01.2	A00_A09	5	2	3.2
7	A00.0	A01.3	A00_A09	5	2	3.2
8	A00.0	A01.4	A00_A09	5	2	3.2
9	A00.0	A02.0	A00_A09	5	2	3.2
10	A00.0	A02.1	A00_A09	5	2	3.2
11	A00.0	A02.2	A00_A09	5	2	3.2
12	A00.0	A02.8	A00_A09	5	2	3.2
13	A00.0	A02.9	A00_A09	5	2	3.2
14	A00.0	A03.0	A00_A09	5	2	3.2
15	A00.0	A03.1	A00_A09	5	2	3.2
16	A00.0	A03.2	A00_A09	5	2	3.2
17	A00.0	A03.3	A00_A09	5	2	3.2
18	A00.0	A03.8	A00_A09	5	2	3.2
19	A00.0	A03.9	A00_A09	5	2	3.2
20	A00.0	A04.0	A00_A09	5	2	3.2
21	A00.0	A04.1	A00_A09	5	2	3.2
22	A00.0	A04.2	A00_A09	5	2	3.2
23	A00.0	A04.3	A00_A09	5	2	3.2
24	A00.0	A04.4	A00_A09	5	2	3.2
25	A00.0	A04.5	A00_A09	5	2	3.2
26	A00.0	A04.6	A00_A09	5	2	3.2
27	A00.0	A04.7	A00_A09	5	2	3.2
28	A00.0	A04.8	A00_A09	5	2	3.2
29	A00.0	A04.9	A00_A09	5	2	3.2

30	A00.0	A05.0	A00_A09	5	2	3.2
31	A00.0	A05.1	A00_A09	5	2	3.2
32	A00.0	A05.2	A00_A09	5	2	3.2
33	A00.0	A05.3	A00_A09	5	2	3.2
34	A00.0	A05.4	A00_A09	5	2	3.2
35	A00.0	A05.8	A00_A09	5	2	3.2
36	A00.0	A05.9	A00_A09	5	2	3.2
37	A00.0	A06.0	A00_A09	5	2	3.2
38	A00.0	A06.1	A00_A09	5	2	3.2
39	A00.0	A06.2	A00_A09	5	2	3.2
40	A00.0	A06.3	A00_A09	5	2	3.2
41	A00.0	A06.4	A00_A09	5	2	3.2
42	A00.0	A06.5	A00_A09	5	2	3.2
43	A00.0	A06.6	A00_A09	5	2	3.2
44	A00.0	A06.7	A00_A09	5	2	3.2
45	A00.0	A06.8	A00_A09	5	2	3.2
46	A00.0	A06.9	A00_A09	5	2	3.2
47	A00.0	A07.0	A00_A09	5	2	3.2
48	A00.0	A07.1	A00_A09	5	2	3.2
49	A00.0	A07.2	A00_A09	5	2	3.2
50	A00.0	A07.3	A00_A09	5	2	3.2
51	A00.0	A07.8	A00_A09	5	2	3.2
52	A00.0	A07.9	A00_A09	5	2	3.2
53	A00.0	A08.0	A00_A09	5	2	3.2
54	A00.0	A08.1	A00_A09	5	2	3.2
55	A00.0	A08.2	A00_A09	5	2	3.2
56	A00.0	A08.3	A00_A09	5	2	3.2
57	A00.0	A08.4	A00_A09	5	2	3.2
58	A00.0	A08.5	A00_A09	5	2	3.2

Table 3.1: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster One)

Note: LCS (c_1, c_2) = Lowest node in hierarchy that is a hypernym of both c1, c2.

Step two: we select another concept (class) *"A15.0"* as first class node and compare it with all remaining leaf nodes (680 concepts) in *"chapter I"* using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (they are more similar, and share more information) and put them all together to create our second cluster (cluster two). As shown in Table 3.2. After that we delete 37 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec(c1, c2)	SemDist(c1, c2)
1	A15.0	A15.0	A15.0	1	0	0
2	A15.0	A15.1	A15	3	1	1.6
3	A15.0	A15.2	A15	3	1	1.6
4	A15.0	A15.3	A15	3	1	1.6
5	A15.0	A15.4	A15	3	1	1.6
6	A15.0	A15.5	A15	3	1	1.6
7	A15.0	A15.6	A15	3	1	1.6
8	A15.0	A15.7	A15	3	1	1.6
9	A15.0	A15.8	A15	3	1	1.6
10	A15.0	A15.9	A15	3	1	1.6
11	A15.0	A16.0	A15_A19	5	2	3.2
12	A15.0	A16.1	A15_A19	5	2	3.2
13	A15.0	A16.2	A15_A19	5	2	3.2
14	A15.0	A16.3	A15_A19	5	2	3.2
15	A15.0	A16.4	A15_A19	5	2	3.2
16	A15.0	A16.5	A15_A19	5	2	3.2
17	A15.0	A16.7	A15_A19	5	2	3.2
18	A15.0	A16.8	A15_A19	5	2	3.2
19	A15.0	A16.9	A15_A19	5	2	3.2
20	A15.0	A17.0	A15_A19	5	2	3.2
21	A15.0	A17.1	A15_A19	5	2	3.2
22	A15.0	A17.8	A15_A19	5	2	3.2
23	A15.0	A17.9	A15_A19	5	2	3.2
24	A15.0	A18.0	A15_A19	5	2	3.2
25	A15.0	A18.1	A15_A19	5	2	3.2
26	A15.0	A18.2	A15_A19	5	2	3.2
27	A15.0	A18.3	A15_A19	5	2	3.2
28	A15.0	A18.4	A15_A19	5	2	3.2
29	A15.0	A18.5	A15_A19	5	2	3.2
30	A15.0	A18.6	A15_A19	5	2	3.2
31	A15.0	A18.7	A15_A19	5	2	3.2
32	A15.0	A18.8	A15_A19	5	2	3.2
33	A15.0	A19..0	A15_A19	5	2	3.2
34	A15.0	A19..1	A15_A19	5	2	3.2
35	A15.0	A19..2	A15_A19	5	2	3.2
36	A15.0	A19..8	A15_A19	5	2	3.2
37	A15.0	A19..9	A15_A19	5	2	3.2

Table 3.2: Compare between concepts (classes) using SemDist (C_1, C_2)   (Cluster Two)

Step three: we select another concept (class) *"A20.0"* as third class node and compare it with all remaining leaf nodes (643 concepts) in *"chapter I"* using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our third cluster (cluster three). As shown in Table 3.3. After that we delete 46 concepts from our experiment.

ID	Concept1	Concept2	LCS(c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	A20.0	A20.0	A20.0	1	0	0
2	A20.0	A20.1	A20	3	1	1.6
3	A20.0	A20.2	A20	3	1	1.6
4	A20.0	A20.3	A20	3	1	1.6
5	A20.0	A20.7	A20	3	1	1.6
6	A20.0	A20.8	A20	3	1	1.6
7	A20.0	A20.9	A20	3	1	1.6
8	A20.0	A21.0	A20_A28	5	2	3.2
9	A20.0	A21.1	A20_A28	5	2	3.2
10	A20.0	A21.2	A20_A28	5	2	3.2
11	A20.0	A21.3	A20_A28	5	2	3.2
12	A20.0	A21.7	A20_A28	5	2	3.2
13	A20.0	A21.8	A20_A28	5	2	3.2
14	A20.0	A21.9	A20_A28	5	2	3.2
15	A20.0	A22.0	A20_A28	5	2	3.2
16	A20.0	A22.1	A20_A28	5	2	3.2
17	A20.0	A22.2	A20_A28	5	2	3.2
18	A20.0	A22.7	A20_A28	5	2	3.2
19	A20.0	A22.8	A20_A28	5	2	3.2
20	A20.0	A22.9	A20_A28	5	2	3.2
21	A20.0	A23.0	A20_A28	5	2	3.2
22	A20.0	A23.1	A20_A28	5	2	3.2
23	A20.0	A23.2	A20_A28	5	2	3.2
24	A20.0	A23.3	A20_A28	5	2	3.2
25	A20.0	A23.8	A20_A28	5	2	3.2
26	A20.0	A23.9	A20_A28	5	2	3.2
27	A20.0	A24.0	A20_A28	5	2	3.2

28	A20.0	A24.1	A20_A28	5	2	3.2
29	A20.0	A24.2	A20_A28	5	2	3.2
30	A20.0	A24.3	A20_A28	5	2	3.2
31	A20.0	A24.4	A20_A28	5	2	3.2
32	A20.0	A25.0	A20_A28	5	2	3.2
33	A20.0	A25.1	A20_A28	5	2	3.2
34	A20.0	A25.9	A20_A28	5	2	3.2
35	A20.0	A26.0	A20_A28	5	2	3.2
36	A20.0	A26.7	A20_A28	5	2	3.2
37	A20.0	A26.8	A20_A28	5	2	3.2
38	A20.0	A26.9	A20_A28	5	2	3.2
39	A20.0	A27.0	A20_A28	5	2	3.2
40	A20.0	A27.8	A20_A28	5	2	3.2
41	A20.0	A27.9	A20_A28	5	2	3.2
42	A20.0	A28.0	A20_A28	5	2	3.2
43	A20.0	A28.1	A20_A28	5	2	3.2
44	A20.0	A28.2	A20_A28	5	2	3.2
45	A20.0	A28.8	A20_A28	5	2	3.2
46	A20.0	A28.9	A20_A28	5	2	3.2

Table 3.3: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Three)

Step Four: we select another concept (class) *"A30.0"* as third class node and compare it with all remaining leaf nodes (597 concepts) in *"chapter I"* using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our fourth cluster (cluster four). As shown in Table 3.4. After that we delete 75 concepts from our experiment.

ID	**Concept1**	**Concept2**	**LCS(c1, c2)**	**Length(c1, c2)**	**CSPec (c1, c2)**	**SemDist(c1, c2)**
1	A30.0	A30.0	A30.0	1	0	0
2	A30.0	A30.1	A30	3	1	1.6
3	A30.0	A30.2	A30	3	1	1.6
4	A30.0	A30.3	A30	3	1	1.6
5	A30.0	A30.4	A30	3	1	1.6
6	A30.0	A30.5	A30	3	1	1.6
7	A30.0	A30.8	A30	3	1	1.6

8	A30.0	A30.9	A30	3	1	1.6
9	A30.0	A31.0	A30_A49	5	2	3.2
10	A30.0	A31.1	A30_A49	5	2	3.2
11	A30.0	A31.8	A30_A49	5	2	3.2
12	A30.0	A31.9	A30_A49	5	2	3.2
13	A30.0	A32.0	A30_A49	5	2	3.2
14	A30.0	A32.1	A30_A49	5	2	3.2
15	A30.0	A32.7	A30_A49	5	2	3.2
16	A30.0	A32.8	A30_A49	5	2	3.2
17	A30.0	A32.9	A30_A49	5	2	3.2
18	A30.0	A36.0	A30_A49	5	2	3.2
19	A30.0	A36.1	A30_A49	5	2	3.2
20	A30.0	A36.2	A30_A49	5	2	3.2
21	A30.0	A36.3	A30_A49	5	2	3.2
22	A30.0	A36.8	A30_A49	5	2	3.2
23	A30.0	A36.9	A30_A49	5	2	3.2
24	A30.0	A37.0	A30_A49	5	2	3.2
25	A30.0	A37.1	A30_A49	5	2	3.2
26	A30.0	A37.8	A30_A49	5	2	3.2
27	A30.0	A37.9	A30_A49	5	2	3.2
28	A30.0	A39.0	A30_A49	5	2	3.2
29	A30.0	A39.1	A30_A49	5	2	3.2
30	A30.0	A39.2	A30_A49	5	2	3.2
31	A30.0	A39.3	A30_A49	5	2	3.2
32	A30.0	A39.4	A30_A49	5	2	3.2
33	A30.0	A39.5	A30_A49	5	2	3.2
34	A30.0	A39.8	A30_A49	5	2	3.2
35	A30.0	A39.9	A30_A49	5	2	3.2
36	A30.0	A40.0	A30_A49	5	2	3.2
37	A30.0	A40.1	A30_A49	5	2	3.2
38	A30.0	A40.2	A30_A49	5	2	3.2
39	A30.0	A40.3	A30_A49	5	2	3.2
40	A30.0	A40.8	A30_A49	5	2	3.2
41	A30.0	A40.9	A30_A49	5	2	3.2
42	A30.0	A41.0	A30_A49	5	2	3.2
43	A30.0	A41.1	A30_A49	5	2	3.2
44	A30.0	A41.2	A30_A49	5	2	3.2
45	A30.0	A41.3	A30_A49	5	2	3.2
46	A30.0	A41.4	A30_A49	5	2	3.2

47	A30.0	A41.5	A30_A49	5	2	3.2
48	A30.0	A41.8	A30_A49	5	2	3.2
49	A30.0	A41.9	A30_A49	5	2	3.2
50	A30.0	A42.0	A30_A49	5	2	3.2
51	A30.0	A42.1	A30_A49	5	2	3.2
52	A30.0	A42.2	A30_A49	5	2	3.2
53	A30.0	A42.7	A30_A49	5	2	3.2
54	A30.0	A42.8	A30_A49	5	2	3.2
55	A30.0	A42.9	A30_A49	5	2	3.2
56	A30.0	A43.0	A30_A49	5	2	3.2
57	A30.0	A43.1	A30_A49	5	2	3.2
58	A30.0	A43.8	A30_A49	5	2	3.2
59	A30.0	A43.9	A30_A49	5	2	3.2
60	A30.0	A44.0	A30_A49	5	2	3.2
61	A30.0	A44.1	A30_A49	5	2	3.2
62	A30.0	A44.8	A30_A49	5	2	3.2
63	A30.0	A44.9	A30_A49	5	2	3.2
64	A30.0	A48.0	A30_A49	5	2	3.2
65	A30.0	A48.1	A30_A49	5	2	3.2
66	A30.0	A48.2	A30_A49	5	2	3.2
67	A30.0	A48.3	A30_A49	5	2	3.2
68	A30.0	A48.4	A30_A49	5	2	3.2
69	A30.0	A48.8	A30_A49	5	2	3.2
70	A30.0	A49.0	A30_A49	5	2	3.2
71	A30.0	A49.1	A30_A49	5	2	3.2
72	A30.0	A49.2	A30_A49	5	2	3.2
73	A30.0	A49.3	A30_A49	5	2	3.2
74	A30.0	A49.8	A30_A49	5	2	3.2
75	A30.0	A49.9	A30_A49	5	2	3.2

Table 3.4: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Four)

Step Five: we select another concept (class) *"A50.0"* as class node and compare it with all remaining leaf nodes (522 concepts) in *"chapter I"* using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our fifth cluster (cluster five). As shown in Table 3.5. After that we delete 48 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	A50.0	A50.0	A50.0	1	0	0
2	A50.0	A50.1	A50	3	1	1.6
3	A50.0	A50.2	A50	3	1	1.6
4	A50.0	A50.3	A50	3	1	1.6
5	A50.0	A50.4	A50	3	1	1.6
6	A50.0	A50.5	A50	3	1	1.6
7	A50.0	A50.6	A50	3	1	1.6
8	A50.0	A50.7	A50	3	1	1.6
9	A50.0	A50.9	A50	3	1	1.6
10	A50.0	A51.0	A50_A64	5	2	3.2
11	A50.0	A51.1	A50_A64	5	2	3.2
12	A50.0	A51.2	A50_A64	5	2	3.2
13	A50.0	A51.3	A50_A64	5	2	3.2
14	A50.0	A51.4	A50_A64	5	2	3.2
15	A50.0	A51.5	A50_A64	5	2	3.2
16	A50.0	A51.9	A50_A64	5	2	3.2
17	A50.0	A52.0	A50_A64	5	2	3.2
18	A50.0	A52.1	A50_A64	5	2	3.2
19	A50.0	A52.2	A50_A64	5	2	3.2
20	A50.0	A52.3	A50_A64	5	2	3.2
21	A50.0	A52.7	A50_A64	5	2	3.2
22	A50.0	A52.8	A50_A64	5	2	3.2
23	A50.0	A52.9	A50_A64	5	2	3.2
24	A50.0	A53.0	A50_A64	5	2	3.2
25	A50.0	A53.9	A50_A64	5	2	3.2
26	A50.0	A54.0	A50_A64	5	2	3.2
27	A50.0	A54.1	A50_A64	5	2	3.2
28	A50.0	A54.2	A50_A64	5	2	3.2
29	A50.0	A54.3	A50_A64	5	2	3.2
30	A50.0	A54.4	A50_A64	5	2	3.2
31	A50.0	A54.5	A50_A64	5	2	3.2
32	A50.0	A54.6	A50_A64	5	2	3.2
33	A50.0	A54.8	A50_A64	5	2	3.2
34	A50.0	A54.9	A50_A64	5	2	3.2
35	A50.0	A56.0	A50_A64	5	2	3.2
36	A50.0	A56.1	A50_A64	5	2	3.2
37	A50.0	A56.2	A50_A64	5	2	3.2
38	A50.0	A56.3	A50_A64	5	2	3.2

39	A50.0	A56.4	A50_A64	5	2	3.2
40	A50.0	A56.8	A50_A64	5	2	3.2
41	A50.0	A59.0	A50_A64	5	2	3.2
42	A50.0	A59.8	A50_A64	5	2	3.2
43	A50.0	A59.9	A50_A64	5	2	3.2
44	A50.0	A60.0	A50_A64	5	2	3.2
45	A50.0	A60.1	A50_A64	5	2	3.2
46	A50.0	A60.9	A50_A64	5	2	3.2
47	A50.0	A63.0	A50_A64	5	2	3.2
48	A50.0	A63.8	A50_A64	5	2	3.2

Table 3.5: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Five)

Step six: we select another concept (class) *"A66.0"* as class node and compare it with all remaining leaf nodes (474 concepts) in *"chapter I"* using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our sixth cluster (cluster six). As shown in Table 3.6. After that we delete 23 concepts from our experiment.

ID	**Concept1**	**Concept2**	**LCS (c1, c2)**	**Length(c1, c2)**	**CSPec (c1, c2)**	**SemDist(c1, c2)**
1	A66.0	A66.0	A66.0	1	0	0
2	A66.0	A66.1	A66	3	1	1.6
3	A66.0	A66.2	A66	3	1	1.6
4	A66.0	A66.3	A66	3	1	1.6
5	A66.0	A66.4	A66	3	1	1.6
6	A66.0	A66.5	A66	3	1	1.6
7	A66.0	A66.6	A66	3	1	1.6
8	A66.0	A66.7	A66	3	1	1.6
9	A66.0	A66.8	A66	3	1	1.6
10	A66.0	A66.9	A66	3	1	1.6
11	A66.0	A67.0	A65_A69	5	2	3.2
12	A66.0	A67.1	A65_A69	5	2	3.2
13	A66.0	A67.2	A65_A69	5	2	3.2
14	A66.0	A67.3	A65_A69	5	2	3.2
15	A66.0	A67.9	A65_A69	5	2	3.2
16	A66.0	A68.0	A65_A69	5	2	3.2
17	A66.0	A68.1	A65_A69	5	2	3.2
18	A66.0	A68.9	A65_A69	5	2	3.2
19	A66.0	A69.0	A65_A69	5	2	3.2

20	A66.0	A69.1	A65_A69	5	2	3.2
21	A66.0	A69.2	A65_A69	5	2	3.2
22	A66.0	A69.8	A65_A69	5	2	3.2
23	A66.0	A69.9	A65_A69	5	2	3.2

Table 3.6: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Six)

Step seven: we select another concept (class) *“A71.0”* as class node and compare it with all remaining leaf nodes (451 concepts) in “chapter I” using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our seventh cluster (cluster seven). As shown in Table 3.7. After that we delete 6 concepts from our experiment.

ID	**Concept1**	**Concept2**	**LCS (c1, c2)**	**Length(c1, c2)**	**CSPec (c1, c2)**	**SemDist(c1, c2)**
1	A71.0	A71.0	A71.0	1	0	0
2	A71.0	A71.1	A71	3	1	1.6
3	A71.0	A71.9	A71	3	1	1.6
4	A71.0	A74.0	A70_A74	5	2	3.2
5	A71.0	A74.8	A70_A74	5	2	3.2
6	A71.0	A74.9	A70_A74	5	2	3.2

Table 3.7: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Seven)

Eighth Step: we select another concept (class) *“A75.0”* as class node and compare it with all remaining leaf nodes (445 concepts) in “chapter I” using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our eighth cluster (cluster eight). As shown in Table 3.8. After that we delete 15 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	A75.0	A75.0	A75.0	1	0	0
2	A75.0	A75.1	A75	3	1	1.6
3	A75.0	A75.2	A75	3	1	1.6
4	A75.0	A75.3	A75	3	1	1.6
5	A75.0	A75.9	A75	3	1	1.6
6	A75.0	A77.0	A75_A79	5	3	3.2
7	A75.0	A77.1	A75_A79	5	3	3.2
8	A75.0	A77.2	A75_A79	5	3	3.2
9	A75.0	A77.3	A75_A79	5	3	3.2
10	A75.0	A77.8	A75_A79	5	3	3.2
11	A75.0	A77.9	A75_A79	5	3	3.2
12	A75.0	A79.0	A75_A79	5	3	3.2
13	A75.0	A79.1	A75_A79	5	3	3.2
14	A75.0	A79.8	A75_A79	5	3	3.2
15	A75.0	A79.9	A75_A79	5	3	3.2

Table 3.8: Compare between concepts (classes) using SemDist (C_1, C_2)   (Cluster Eigth)

Step eight : we select another concept (class) “A75.0” as class node and compare it with all remaining leaf nodes (430 concepts) in “chapter I” using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our ninth cluster (cluster nine). As shown in Table 3.9. After that we delete 39 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length (C1, C2)	CSPec (c1, c2)	SemDist(c1, c2)
1	A80.0	A80.0	A80.0	1	0	0
2	A80.0	A80.1	A80	3	1	1.6
3	A80.0	A80.2	A80	3	1	1.6
4	A80.0	A80.3	A80	3	1	1.6
5	A80.0	A80.4	A80	3	1	1.6
6	A80.0	A80.9	A80	3	1	1.6
7	A80.0	A81.0	A80_A89	5	2	3.2
8	A80.0	A81.1	A80_A89	5	2	3.2
9	A80.0	A81.2	A80_A89	5	2	3.2
10	A80.0	A81.8	A80_A89	5	2	3.2

11	A80.0	A81.9	A80_A89	5	2	3.2
12	A80.0	A82.0	A80_A89	5	2	3.2
13	A80.0	A82.1	A80_A89	5	2	3.2
14	A80.0	A82.9	A80_A89	5	2	3.2
15	A80.0	A83.0	A80_A89	5	2	3.2
16	A80.0	A83.1	A80_A89	5	2	3.2
17	A80.0	A83.2	A80_A89	5	2	3.2
18	A80.0	A83.3	A80_A89	5	2	3.2
19	A80.0	A83.4	A80_A89	5	2	3.2
20	A80.0	A83.5	A80_A89	5	2	3.2
21	A80.0	A83.6	A80_A89	5	2	3.2
22	A80.0	A83.8	A80_A89	5	2	3.2
23	A80.0	A83.9	A80_A89	5	2	3.2
24	A80.0	A84.0	A80_A89	5	2	3.2
25	A80.0	A84.1	A80_A89	5	2	3.2
26	A80.0	A84.8	A80_A89	5	2	3.2
27	A80.0	A84.9	A80_A89	5	2	3.2
28	A80.0	A85.0	A80_A89	5	2	3.2
29	A80.0	A85.1	A80_A89	5	2	3.2
30	A80.0	A85.2	A80_A89	5	2	3.2
31	A80.0	A85.8	A80_A89	5	2	3.2
32	A80.0	A87.0	A80_A89	5	2	3.2
33	A80.0	A87.1	A80_A89	5	2	3.2
34	A80.0	A87.2	A80_A89	5	2	3.2
35	A80.0	A87.8	A80_A89	5	2	3.2
36	A80.0	A87.9	A80_A89	5	2	3.2
37	A80.0	A88.0	A80_A89	5	2	3.2
38	A80.0	A88.1	A80_A89	5	2	3.2
39	A80.0	A88.8	A80_A89	5	2	3.2

Table 3.9: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Nine)

Step nine: we select another concept (class) "A80.0" as class node and compare it with all remaining leaf nodes (391 concepts) in "*chapter I*" using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our tenth cluster (cluster ten). As shown in Table 3.10. After that we delete 26 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	A92.0	A92.0	A92.0	1	0	0
2	A92.0	A92.1	A92	3	1	1.6
3	A92.0	A92.2	A92	3	1	1.6
4	A92.0	A92.3	A92	3	1	1.6
5	A92.0	A92.4	A92	3	1	1.6
6	A92.0	A92.8	A92	3	1	1.6
7	A92.0	A92.9	A92	3	1	1.6
8	A92.0	A93.0	A90_A99	5	2	3.2
9	A92.0	A93.1	A90_A99	5	2	3.2
10	A92.0	A93.2	A90_A99	5	2	3.2
11	A92.0	A93.8	A90_A99	5	2	3.2
12	A92.0	A95.0	A90_A99	5	2	3.2
13	A92.0	A95.1	A90_A99	5	2	3.2
14	A92.0	A95.9	A90_A99	5	2	3.2
15	A92.0	A96.0	A90_A99	5	2	3.2
16	A92.0	A96.1	A90_A99	5	2	3.2
17	A92.0	A96.2	A90_A99	5	2	3.2
18	A92.0	A96.8	A90_A99	5	2	3.2
19	A92.0	A96.9	A90_A99	5	2	3.2
20	A92.0	A98.0	A90_A99	5	2	3.2
21	A92.0	A98.1	A90_A99	5	2	3.2
22	A92.0	A98.2	A90_A99	5	2	3.2
23	A92.0	A98.3	A90_A99	5	2	3.2
24	A92.0	A98.4	A90_A99	5	2	3.2
25	A92.0	A98.5	A90_A99	5	2	3.2
26	A92.0	A98.8	A90_A99	5	2	3.2

Table 3.10: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Ten)

Step ten: we select another concept (class) “B00.0” as class node and compare it with all remaining leaf nodes (365 concepts) in *“chapter I’* using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our eleventh cluster (cluster eleven). As shown in Table 3.11. After that we delete 38 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B00.0	B00.0	B00.0	1	0	0
2	B00.0	B00.1	B00	3	1	1.6
3	B00.0	B00.2	B00	3	1	1.6
4	B00.0	B00.3	B00	3	1	1.6
5	B00.0	B00.4	B00	3	1	1.6
6	B00.0	B00.5	B00	3	1	1.6
7	B00.0	B00.7	B00	3	1	1.6
8	B00.0	B00.8	B00	3	1	1.6
9	B00.0	B00.9	B00	3	1	1.6
10	B00.0	B01.0	B00_B09	5	2	3.2
11	B00.0	B01.1	B00_B09	5	2	3.2
12	B00.0	B01.2	B00_B09	5	2	3.2
13	B00.0	B01.8	B00_B09	5	2	3.2
14	B00.0	B01.9	B00_B09	5	2	3.2
15	B00.0	B02.0	B00_B09	5	2	3.2
16	B00.0	B02.1	B00_B09	5	2	3.2
17	B00.0	B02.2	B00_B09	5	2	3.2
18	B00.0	B02.3	B00_B09	5	2	3.2
19	B00.0	B02.7	B00_B09	5	2	3.2
20	B00.0	B02.8	B00_B09	5	2	3.2
21	B00.0	B02.9	B00_B09	5	2	3.2
22	B00.0	B05.0	B00_B09	5	2	3.2
23	B00.0	B05.1	B00_B09	5	2	3.2
24	B00.0	B05.2	B00_B09	5	2	3.2
25	B00.0	B05.3	B00_B09	5	2	3.2
26	B00.0	B05.4	B00_B09	5	2	3.2
27	B00.0	B05.8	B00_B09	5	2	3.2
28	B00.0	B05.9	B00_B09	5	2	3.2
29	B00.0	B06.0	B00_B09	5	2	3.2
30	B00.0	B06.8	B00_B09	5	2	3.2
31	B00.0	B06.9	B00_B09	5	2	3.2
32	B00.0	B08.0	B00_B09	5	2	3.2
33	B00.0	B08.1	B00_B09	5	2	3.2
34	B00.0	B08.2	B00_B09	5	2	3.2
35	B00.0	B08.3	B00_B09	5	2	3.2
36	B00.0	B08.4	B00_B09	5	2	3.2
37	B00.0	B08.5	B00_B09	5	2	3.2
38	B00.0	B08.8	B00_B09	5	2	3.2

Table 3.11: Compare between concepts (classes) using SemDist (C_1, C_2 (Cluster Eleven)

Step eleven: we select another concept (class) "B15.0" as class node and compare it with all remaining leaf nodes (327 concepts) in "chapter I" using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our Twelfth cluster (cluster Twelve). As shown in Table 3.12. After that we delete 17 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B15.0	B15.0	B15.0	1	0	0
2	B15.0	B15.9	B15	3	1	1.6
3	B15.0	B16.0	B15_B19	5	2	3.2
4	B15.0	B16.1	B15_B19	5	2	3.2
5	B15.0	B16.2	B15_B19	5	2	3.2
6	B15.0	B16.9	B15_B19	5	2	3.2
7	B15.0	B17.0	B15_B19	5	2	3.2
8	B15.0	B17.1	B15_B19	5	2	3.2
9	B15.0	B17.2	B15_B19	5	2	3.2
10	B15.0	B17.8	B15_B19	5	2	3.2
11	B15.0	B18.0	B15_B19	5	2	3.2
12	B15.0	B18.1	B15_B19	5	2	3.2
13	B15.0	B18.2	B15_B19	5	2	3.2
14	B15.0	B18.8	B15_B19	5	2	3.2
15	B15.0	B18.9	B15_B19	5	2	3.2
16	B15.0	B19.0	B15_B19	5	2	3.2
17	B15.0	B19.9	B15_B19	5	2	3.2

Table4.12: Compare between concepts (classes) using SemDist (C_1, C_2)(Cluster Twelve)

Step twelve: we select another concept (class) "B20.0" as class node and compare it with all remaining leaf nodes (310 concepts) in "*chapter I*" using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our Thirteenth cluster (cluster Thirteen). As shown in Table 3.13. After that we delete 25 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B20.0	B20.0	B20.0	1	0	0
2	B20.0	B20.1	B20	3	1	1.6
3	B20.0	B20.2	B20	3	1	1.6
4	B20.0	B20.3	B20	3	1	1.6
5	B20.0	B20.4	B20	3	1	1.6
6	B20.0	B20.5	B20	3	1	1.6
7	B20.0	B20.6	B20	3	1	1.6
8	B20.0	B20.7	B20	3	1	1.6
9	B20.0	B20.8	B20	3	1	1.6
10	B20.0	B20.9	B20	3	1	1.6
11	B20.0	B21.0	B20_B24	5	2	3.2
12	B20.0	B21.1	B20_B24	5	2	3.2
13	B20.0	B21.2	B20_B24	5	2	3.2
14	B20.0	B21.3	B20_B24	5	2	3.2
15	B20.0	B21.7	B20_B24	5	2	3.2
16	B20.0	B21.8	B20_B24	5	2	3.2
17	B20.0	B21.9	B20_B24	5	2	3.2
18	B20.0	B22.0	B20_B24	5	2	3.2
19	B20.0	B22.1	B20_B24	5	2	3.2
20	B20.0	B22.2	B20_B24	5	2	3.2
21	B20.0	B22.7	B20_B24	5	2	3.2
22	B20.0	B23.0	B20_B24	5	2	3.2
23	B20.0	B23.1	B20_B24	5	2	3.2
24	B20.0	B23.2	B20_B24	5	2	3.2
25	B20.0	B23.8	B20_B24	5	2	3.2

Table4.13: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Thirteen)

Step thirteen: we select another concept (class) “B25.0” as class node and compare it with all remaining leaf nodes (285 concepts) in “chapter I” using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our Fourteenth cluster (cluster Fourteen). As shown in Table 3.14. After that we delete 33 concepts from our experiment.

ID	**Concept1**	**Concept2**	**LCS (c1, c2)**	**Length(c1, c2)**	**CSPec (c1, c2)**	**SemDist(c1, c2)**
1	B25.0	B25.0	B25.0	1	0	0
2	B25.0	B25.1	B25	3	1	1.6
3	B25.0	B25.2	B25	3	1	1.6
4	B25.0	B25.8	B25	3	1	1.6
5	B25.0	B25.9	B25	3	1	1.6
6	B25.0	B26.0	B25_B34	5	2	3.2
7	B25.0	B26.1	B25_B34	5	2	3.2
8	B25.0	B26.2	B25_B34	5	2	3.2
9	B25.0	B26.3	B25_B34	5	2	3.2
10	B25.0	B26.8	B25_B34	5	2	3.2
11	B25.0	B26.9	B25_B34	5	2	3.2
12	B25.0	B27.0	B25_B34	5	2	3.2
13	B25.0	B27.1	B25_B34	5	2	3.2
14	B25.0	B27.8	B25_B34	5	2	3.2
15	B25.0	B27.9	B25_B34	5	2	3.2
16	B25.0	B30.0	B25_B34	5	2	3.2
17	B25.0	B30.1	B25_B34	5	2	3.2
18	B25.0	B30.2	B25_B34	5	2	3.2
19	B25.0	B30.3	B25_B34	5	2	3.2
20	B25.0	B30.8	B25_B34	5	2	3.2
21	B25.0	B30.9	B25_B34	5	2	3.2
22	B25.0	B33.0	B25_B34	5	2	3.2
23	B25.0	B33.1	B25_B34	5	2	3.2
24	B25.0	B33.2	B25_B34	5	2	3.2
25	B25.0	B33.3	B25_B34	5	2	3.2
26	B25.0	B33.8	B25_B34	5	2	3.2
27	B25.0	B34.0	B25_B34	5	2	3.2
28	B25.0	B34.1	B25_B34	5	2	3.2
29	B25.0	B34.2	B25_B34	5	2	3.2
30	B25.0	B34.3	B25_B34	5	2	3.2
31	B25.0	B34.4	B25_B34	5	2	3.2
32	B25.0	B34.8	B25_B34	5	2	3.2
33	B25.0	B34.9	B25_B34	5	2	3.2

Table4.14: Compare between concepts (classes) using SemDist(C_1, C_2)(Cluster Fourteen)

Step fourteen: we select another concept (class) "B35.0" as class node and compare it with all remaining leaf nodes (252 concepts) in "chapter I" using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our Fifteenth cluster (cluster Fifteenth). As shown in Table 3.15. After that we delete 92 concepts from our experiment.

ID	**Concept1**	**Concept2**	**LCS (c1, c2)**	**Length(c1, c2)**	**CSPec (c1, c2)**	**SemDist(c1, c2)**
1	B35.0	B35.0	B35.0	1	0	0
2	B35.0	B35.1	B35	3	1	1.6
3	B35.0	B35.2	B35	3	1	1.6
4	B35.0	B35.3	B35	3	1	1.6
5	B35.0	B35.4	B35	3	1	1.6
6	B35.0	B35.5	B35	3	1	1.6
7	B35.0	B35.6	B35	3	1	1.6
8	B35.0	B35.8	B35	3	1	1.6
9	B35.0	B35.9	B35	3	1	1.6
10	B35.0	B36.0	B35_B49	5	2	3.2
11	B35.0	B36.1	B35_B49	5	2	3.2
12	B35.0	B36.2	B35_B49	5	2	3.2
13	B35.0	B36.3	B35_B49	5	2	3.2
14	B35.0	B36.8	B35_B49	5	2	3.2
15	B35.0	B36.9	B35_B49	5	2	3.2
16	B35.0	B37.0	B35_B49	5	2	3.2
17	B35.0	B37.1	B35_B49	5	2	3.2
18	B35.0	B37.2	B35_B49	5	2	3.2
19	B35.0	B37.3	B35_B49	5	2	3.2
20	B35.0	B37.4	B35_B49	5	2	3.2
21	B35.0	B37.5	B35_B49	5	2	3.2
22	B35.0	B37.6	B35_B49	5	2	3.2
23	B35.0	B37.7	B35_B49	5	2	3.2
24	B35.0	B37.8	B35_B49	5	2	3.2
25	B35.0	B37.9	B35_B49	5	2	3.2
26	B35.0	B38.0	B35_B49	5	2	3.2
27	B35.0	B38.1	B35_B49	5	2	3.2
28	B35.0	B38.2	B35_B49	5	2	3.2
29	B35.0	B38.3	B35_B49	5	2	3.2

30	B35.0	B38.4	B35_B49	5	2	3.2
31	B35.0	B38.7	B35_B49	5	2	3.2
32	B35.0	B38.8	B35_B49	5	2	3.2
33	B35.0	B38.9	B35_B49	5	2	3.2
34	B35.0	B39.0	B35_B49	5	2	3.2
35	B35.0	B39.1	B35_B49	5	2	3.2
36	B35.0	B39.2	B35_B49	5	2	3.2
37	B35.0	B39.3	B35_B49	5	2	3.2
38	B35.0	B39.4	B35_B49	5	2	3.2
39	B35.0	B39.5	B35_B49	5	2	3.2
40	B35.0	B39.9	B35_B49	5	2	3.2
41	B35.0	B40.0	B35_B49	5	2	3.2
42	B35.0	B40.1	B35_B49	5	2	3.2
43	B35.0	B40.2	B35_B49	5	2	3.2
44	B35.0	B40.3	B35_B49	5	2	3.2
45	B35.0	B40.7	B35_B49	5	2	3.2
46	B35.0	B40.8	B35_B49	5	2	3.2
47	B35.0	B40.9	B35_B49	5	2	3.2
48	B35.0	B41.0	B35_B49	5	2	3.2
49	B35.0	B41.7	B35_B49	5	2	3.2
50	B35.0	B41.8	B35_B49	5	2	3.2
51	B35.0	B41.9	B35_B49	5	2	3.2
52	B35.0	B42.0	B35_B49	5	2	3.2
53	B35.0	B42.1	B35_B49	5	2	3.2
54	B35.0	B42.7	B35_B49	5	2	3.2
55	B35.0	B42.8	B35_B49	5	2	3.2
56	B35.0	B42.9	B35_B49	5	2	3.2
57	B35.0	B43.0	B35_B49	5	2	3.2
58	B35.0	B43.1	B35_B49	5	2	3.2
59	B35.0	B43.2	B35_B49	5	2	3.2
60	B35.0	B43.8	B35_B49	5	2	3.2
61	B35.0	B43.9	B35_B49	5	2	3.2
62	B35.0	B44.0	B35_B49	5	2	3.2
63	B35.0	B44.1	B35_B49	5	2	3.2
64	B35.0	B44.2	B35_B49	5	2	3.2
65	B35.0	B44.7	B35_B49	5	2	3.2
66	B35.0	B44.8	B35_B49	5	2	3.2
67	B35.0	B44.9	B35_B49	5	2	3.2
68	B35.0	B45.0	B35_B49	5	2	3.2

69	B35.0	B45.1	B35_B49	5	2	3.2
70	B35.0	B45.2	B35_B49	5	2	3.2
71	B35.0	B45.3	B35_B49	5	2	3.2
72	B35.0	B45.7	B35_B49	5	2	3.2
73	B35.0	B45.8	B35_B49	5	2	3.2
74	B35.0	B45.9	B35_B49	5	2	3.2
75	B35.0	B46.0	B35_B49	5	2	3.2
76	B35.0	B46.1	B35_B49	5	2	3.2
77	B35.0	B46.2	B35_B49	5	2	3.2
78	B35.0	B46.3	B35_B49	5	2	3.2
79	B35.0	B46.4	B35_B49	5	2	3.2
80	B35.0	B46.5	B35_B49	5	2	3.2
81	B35.0	B46.8	B35_B49	5	2	3.2
82	B35.0	B46.9	B35_B49	5	2	3.2
83	B35.0	B47.0	B35_B49	5	2	3.2
84	B35.0	B47.1	B35_B49	5	2	3.2
85	B35.0	B47.9	B35_B49	5	2	3.2
86	B35.0	B48.0	B35_B49	5	2	3.2
87	B35.0	B48.1	B35_B49	5	2	3.2
88	B35.0	B48.2	B35_B49	5	2	3.2
89	B35.0	B48.3	B35_B49	5	2	3.2
90	B35.0	B48.4	B35_B49	5	2	3.2
91	B35.0	B48.7	B35_B49	5	2	3.2
92	B35.0	B48.8	B35_B49	5	2	3.2

Table4.15: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Fifteen)

Step fifteen: we select another concept (class) “B50.0” as class node and compare it with all remaining leaf nodes (160 concepts) in chapter I using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our Sixteenth cluster (cluster Sixteen). As shown in Table 3.16. After that we delete 35 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B50.0	B50.0	B50.0	1	0	0
2	B50.0	B50.8	B50	3	1	1.6
3	B50.0	B50.9	B50	3	1	1.6
4	B50.0	B51.0	B50_B64	5	2	3.2
5	B50.0	B51.8	B50_B64	5	2	3.2
6	B50.0	B51.9	B50_B64	5	2	3.2
7	B50.0	B52.0	B50_B64	5	2	3.2
8	B50.0	B52.8	B50_B64	5	2	3.2
9	B50.0	B52.9	B50_B64	5	2	3.2
10	B50.0	B53.0	B50_B64	5	2	3.2
11	B50.0	B53.1	B50_B64	5	2	3.2
12	B50.0	B53.8	B50_B64	5	2	3.2
13	B50.0	B55.0	B50_B64	5	2	3.2
14	B50.0	B55.1	B50_B64	5	2	3.2
15	B50.0	B55.2	B50_B64	5	2	3.2
16	B50.0	B55.9	B50_B64	5	2	3.2
17	B50.0	B56.0	B50_B64	5	2	3.2
18	B50.0	B56.1	B50_B64	5	2	3.2
19	B50.0	B56.9	B50_B64	5	2	3.2
20	B50.0	B57.0	B50_B64	5	2	3.2
21	B50.0	B57.1	B50_B64	5	2	3.2
22	B50.0	B57.2	B50_B64	5	2	3.2
23	B50.0	B57.3	B50_B64	5	2	3.2
24	B50.0	B57.4	B50_B64	5	2	3.2
25	B50.0	B57.5	B50_B64	5	2	3.2
26	B50.0	B58.0	B50_B64	5	2	3.2
27	B50.0	B58.1	B50_B64	5	2	3.2
28	B50.0	B58.2	B50_B64	5	2	3.2
29	B50.0	B58.3	B50_B64	5	2	3.2
30	B50.0	B58.8	B50_B64	5	2	3.2
31	B50.0	B58.9	B50_B64	5	2	3.2
32	B50.0	B60.0	B50_B64	5	2	3.2
33	B50.0	B60.1	B50_B64	5	2	3.2
34	B50.0	B60.2	B50_B64	5	2	3.2
35	B50.0	B60.8	B50_B64	5	2	3.2

Table 3.16: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Sixteen)

Step sixteen: we select another concept (class) "B65.0" as class node and compare it with all remaining leaf nodes (125 concepts) in chapter I using Al-Mubaid and Nguyen's measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our Seventeenth cluster (cluster Seventeen). As shown in Table 3.17. After that we delete 71 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B65.0	B65.0	B65.0	1	0	0
2	B65.0	B65.1	B65	3	1	1.6
3	B65.0	B65.2	B65	3	1	1.6
4	B65.0	B65.3	B65	3	1	1.6
5	B65.0	B65.8	B65	3	1	1.6
6	B65.0	B65.9	B65	3	1	1.6
7	B65.0	B66.0	B65_B83	5	2	3.2
8	B65.0	B66.1	B65_B83	5	2	3.2
9	B65.0	B66.2	B65_B83	5	2	3.2
10	B65.0	B66.3	B65_B83	5	2	3.2
11	B65.0	B66.4	B65_B83	5	2	3.2
12	B65.0	B66.5	B65_B83	5	2	3.2
13	B65.0	B66.8	B65_B83	5	2	3.2
14	B65.0	B66.9	B65_B83	5	2	3.2
15	B65.0	B67.0	B65_B83	5	2	3.2
16	B65.0	B67.1	B65_B83	5	2	3.2
17	B65.0	B67.2	B65_B83	5	2	3.2
18	B65.0	B67.3	B65_B83	5	2	3.2
19	B65.0	B67.4	B65_B83	5	2	3.2
20	B65.0	B67.5	B65_B83	5	2	3.2
21	B65.0	B67.6	B65_B83	5	2	3.2
22	B65.0	B67.7	B65_B83	5	2	3.2
23	B65.0	B67.8	B65_B83	5	2	3.2
24	B65.0	B67.9	B65_B83	5	2	3.2
25	B65.0	B68.0	B65_B83	5	2	3.2
26	B65.0	B68.1	B65_B83	5	2	3.2
27	B65.0	B68.9	B65_B83	5	2	3.2
28	B65.0	B69.0	B65_B83	5	2	3.2
29	B65.0	B69.1	B65_B83	5	2	3.2

30	B65.0	B69.8	B65_B83	5	2	3.2
31	B65.0	B69.9	B65_B83	5	2	3.2
32	B65.0	B70.0	B65_B83	5	2	3.2
33	B65.0	B70.1	B65_B83	5	2	3.2
34	B65.0	B71.0	B65_B83	5	2	3.2
35	B65.0	B71.1	B65_B83	5	2	3.2
36	B65.0	B71.8	B65_B83	5	2	3.2
37	B65.0	B71.9	B65_B83	5	2	3.2
38	B65.0	B74.0	B65_B83	5	2	3.2
39	B65.0	B74.1	B65_B83	5	2	3.2
40	B65.0	B74.2	B65_B83	5	2	3.2
41	B65.0	B74.3	B65_B83	5	2	3.2
42	B65.0	B74.4	B65_B83	5	2	3.2
43	B65.0	B74.8	B65_B83	5	2	3.2
44	B65.0	B74.9	B65_B83	5	2	3.2
45	B65.0	B76.0	B65_B83	5	2	3.2
46	B65.0	B76.1	B65_B83	5	2	3.2
47	B65.0	B76.8	B65_B83	5	2	3.2
48	B65.0	B76.9	B65_B83	5	2	3.2
49	B65.0	B77.0	B65_B83	5	2	3.2
50	B65.0	B77.8	B65_B83	5	2	3.2
51	B65.0	B77.9	B65_B83	5	2	3.2
52	B65.0	B78.0	B65_B83	5	2	3.2
53	B65.0	B78.1	B65_B83	5	2	3.2
54	B65.0	B78.7	B65_B83	5	2	3.2
56	B65.0	B78.9	B65_B83	5	2	3.2
57	B65.0	B81.0	B65_B83	5	2	3.2
58	B65.0	B81.1	B65_B83	5	2	3.2
59	B65.0	B81.2	B65_B83	5	2	3.2
60	B65.0	B81.3	B65_B83	5	2	3.2
61	B65.0	B81.4	B65_B83	5	2	3.2
62	B65.0	B81.8	B65_B83	5	2	3.2
63	B65.0	B82.0	B65_B83	5	2	3.2
64	B65.0	B82.9	B65_B83	5	2	3.2
65	B65.0	B83.0	B65_B83	5	2	3.2
66	B65.0	B83.1	B65_B83	5	2	3.2
67	B65.0	B83.2	B65_B83	5	2	3.2
68	B65.0	B83.3	B65_B83	5	2	3.2
69	B65.0	B83.4	B65_B83	5	2	3.2

70	B65.0	B83.8	B65_B83	5	2	3.2
71	B65.0	B83.9	B65_B83	5	2	3.2

Table 3.17: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Seventeen)

Step seventeen: we select another concept (class) “B85.0” as class node and compare it with all remaining leaf nodes (54 concepts) in chapter I using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our eighteenth cluster (cluster eighteen). As shown in Table 3.18. After that we delete 18 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B85.0	B85.0	B85.0	1	0	0
2	B85.0	B85.1	B85	3	1	1.6
3	B85.0	B85.2	B85	3	1	1.6
4	B85.0	B85.3	B85	3	1	1.6
5	B85.0	B85.4	B85	3	1	1.6
6	B85.0	B87.0	B85_B89	5	2	3.2
7	B85.0	B87.1	B85_B89	5	2	3.2
8	B85.0	B87.2	B85_B89	5	2	3.2
9	B85.0	B87.3	B85_B89	5	2	3.2
10	B85.0	B87.4	B85_B89	5	2	3.2
11	B85.0	B87.8	B85_B89	5	2	3.2
12	B85.0	B87.9	B85_B89	5	2	3.2
13	B85.0	B88.0	B85_B89	5	2	3.2
14	B85.0	B88.1	B85_B89	5	2	3.2
15	B85.0	B88.2	B85_B89	5	2	3.2
16	B85.0	B88.3	B85_B89	5	2	3.2
17	B85.0	B88.8	B85_B89	5	2	3.2
18	B85.0	B88.9	B85_B89	5	2	3.2

Table 3.18: Compare between concepts (classes) using SemDist(C_1, C_2)(Cluster Seventeen)

Step eighteen: we select another concept (class) “B90.0” as class node and compare it with all remaining leaf nodes (36 concepts) in chapter I using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our nineteenth cluster (cluster nineteenth). As shown in Table 3.19. After that we delete 10 concepts from our experiment.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B90.0	B90.0	B90.0	1	0	0
2	B90.0	B90.1	B90	3	1	1.6
3	B90.0	B90.2	B90	3	1	1.6
4	B90.0	B90.8	B90	3	1	1.6
5	B90.0	B90.9	B90	3	1	1.6
6	B90.0	B94.0	B90_B94	5	2	3.2
7	B90.0	B94.1	B90_B94	5	2	3.2
8	B90.0	B94.2	B90_B94	5	2	3.2
9	B90.0	B94.8	B90_B94	5	2	3.2
10	B90.0	B94.9	B90_B94	5	2	3.2

Table 3.19: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Nineteen)

Step nineteen: we select another concept (class) “95.0” as class node and compare it with all remaining leaf nodes (26 concepts) in “chapter I” using Al-Mubaid and Nguyen’s measure (SemDist), and then we select the minimum semantic similarity values from all pairs (more similar, and share more information) and put them all together to create our twenty cluster (cluster twenty). As shown in Table 3.20.

ID	Concept1	Concept2	LCS (c1, c2)	Length(c1, c2)	CSPec (c1, c2)	SemDist(c1, c2)
1	B95.0	B95.0	B95.0	1	0	0
2	B95.0	B95.1	B95	3	1	1.6
3	B95.0	B95.2	B95	3	1	1.6
4	B95.0	B95.3	B95	3	1	1.6
5	B95.0	B95.4	B95	3	1	1.6
6	B95.0	B95.5	B95	3	1	1.6
7	B95.0	B95.6	B95	3	1	1.6
8	B95.0	B95.7	B95	3	1	1.6
9	B95.0	B95.8	B95	3	1	1.6
10	B95.0	B96.0	B95_B97	5	2	3.2
11	B95.0	B96.1	B95_B97	5	2	3.2
12	B95.0	B96.2	B95_B97	5	2	3.2

13	B95.0	B96.3	B95_B97	5	2	3.2
14	B95.0	B96.4	B95_B97	5	2	3.2
15	B95.0	B96.5	B95_B97	5	2	3.2
16	B95.0	B96.6	B95_B97	5	2	3.2
17	B95.0	B96.7	B95_B97	5	2	3.2
18	B95.0	B96.8	B95_B97	5	2	3.2
19	B95.0	B97.0	B95_B97	5	2	3.2
20	B95.0	B97.1	B95_B97	5	2	3.2
21	B95.0	B97.2	B95_B97	5	2	3.2
22	B95.0	B97.3	B95_B97	5	2	3.2
23	B95.0	B97.4	B95_B97	5	2	3.2
24	B95.0	B97.5	B95_B97	5	2	3.2
25	B95.0	B97.6	B95_B97	5	2	3.2
26	B95.0	B97.7	B95_B97	5	2	3.2

Table 3.20: Compare between concepts (classes) using SemDist (C_1, C_2) (Cluster Nineteen)

From all clusters we take small similarity value between two concepts nodes, and we create our reference dataset dataset.

ID	**Concept1(Class)**	**ICD-10 Code**	**Concept2 (Class)**	**ICD-10 Code**	**SemDist**
1	Cholera due to Vibrio cholerae 01, biovar cholera	A00.0	Cholera due to Vibrio cholerae 01, biovar cholera	A00.0	0
2	Cholera due to Vibrio cholerae 01, biovar cholera	A00.0	Cholera due to Vibrio cholerae 01, biovareltor	A00.1	1.6
3	Cholera due to Vibrio cholerae 01, biovar cholera	A00.0	Cholera, unspecified	A00.9	1.6
4	Tuberculosis of lung, confirmed by sputum microscopy with or without culture	A15.0	Tuberculosis of lung, confirmed by sputum microscopy with or without culture	A15.0	0
5	Tuberculosis of lung, confirmed by sputum microscopy with or without culture	A15.0	Tuberculosis of lung, confirmed by culture only	A15.1	1.6
131	. . . Streptococcus, group A, as the cause of diseases classified to other chapters	. . B95.0	Unspecified staphylococcus as the cause of diseases classified to other chapters	B95.8	1.6

Table 3.3: Infectious and Parasitic DO-Bench concepts (classes)

LIST OF PUBLICATIONS

1. Paper title: **“Survey on Semantic Similarity Measures between Terms in the Biomedical Domain within frame work Unified Medical Language System (UMLS)”** Scientific Research An Academic Publisher, June 26, 2015

   address: http://file.scirp.org/pdf/AJCM_2015062613262891.pdf

2. Research paper title: “**Evaluating Semantic Similarity between Biomedical Concepts/Classes through Single Ontology”**

   address: http://ijcat.com/archieve/volume7/issue8/ijcatr07081009.pdf