JIMMA UNIVERSITY

INSTITUTE OF TECHNOLOGY.

DEPARTMENT OF INFORMATION SCIENCE

MSC.IN ELECTRONIC AND DIGITAL RESOURCE MANAGEMENT

ANALYSIS OF SEMANTIC WEB TECHNOLOGY AND WEB ONTOLOGY LANGUAGE FOR DIGITAL LIBRARY SERVICE

A thesis research submitted to Jimma University in partial fulfillment of the requirement for degree of Masters of Science in Information Science (Electronic and Digital Resource Management).

TIJANI AWOL

Principal Advisor: - Dr.Amanuel Ayde

Co-advisor: Elsabet Wodajo (Msc)

JIMMA, ETHIOPHIA

April, 2022

JIMMA UNIVERSITY

POST GRADUATE PROGRAM DIRECTORATE

I hereby certify that I have read and evaluated this Thesis research entitled "**Analysis of Semantic Web Technology and Web Ontology for Digital Library Service**" prepared under my guidance by Tijani Awol. I recommend that it be submitted as fulfilling the thesis proposal requirement.

_____ _____
_____

Major Advisor Signature
Date

_____ _____
_____

Co-Advisor Signature Date

ACKNOWLEDGMENT

First and foremost, I would like to give a special gratitude to Allah who provided me everything to write this thesis proposal. Next, I would like to acknowledge the department of information science for giving the chance to prepare this thesis. My grateful thanks go to my principal advisor Dr.Amanuel Ayde (PHD) who provided unreserved precious advices, comments, important directions at each step of the research, starting from the scratch. I gratefully thank my co-Advisor Mr. Elsabet wodajo (M.Sc.) for her constructive comments, encouragements and advices towards my thesis research during my studies.

ACRONOMY AND ABBRIVIATION

DL digital library

HTML hypertext markup language

IEEE Institute of Electrical and Electronics Engineers

IFLA International Federation of Library Association

OPAC Online Public Access Catalogue

OWL web ontology language

RDF Resource Description Frame Work

SKOS Simple Knowledge Organization System

SW Semantic Web

W3C world wide web consortium

XML Extensible Markup Language

Table of Contents

List of tables

List of Figures

Abstract

The broad concept of digital libraries ideally represents the needs of heterogeneous information resources combining the development of complex systems issues such as interoperability among existing data providers, distributed retrieval, and long-term preservation. These days most of the existing web-based digital libraries offer search as well as navigation services. Search services are mainly based on a set of metadata such as domain related keywords, author, title, or journal name that carries no semantic information for the search engine. Besides, digital libraries may support graphical user interfaces to aid the formation of queries in order to search documents. Queries can contain one or more domain keywords to be searched in the full text document, or in one of the metadata fields. This study aimed to harness the collective knowledge within communities in digital libraries, improving the discovery and dissemination of knowledge through ontologies. The study is intended to analyze the semantic web technology and web ontology–based language application for sharing knowledge through digital library service. The design science and exploratory research methodology is employed in this study to explore and analyze semantic web and ontology. In this study the researcher have evaluated the effect of semantic web technology and web ontology language for digital library service. The study analyzed various digital library services with the support of semantic web technologies. Mainly, digital libraries have been benefited from ontology language, semantic web and syntactical features in order to enhance the searching capabilities of their systems. To do so, in this study the researcher have discussed all the digital libraries searching strategies and come up with a new integrated robust framework with the help of domain specific ontologies and semantic sentence embedding approach. The study has used an existing ontology for infectious disease portal in order to extract dictionaries of terms, concept relationships and properties. On the other hand, the researcher used the capabilities of sentence embedding techniques on the top of rich set of pre trained vocabularies and pre trained models In order to evaluate the propose framework the researcher prepared a small size dataset on domain specific communicable diseases. The dataset consists of relevant metadata of articles such as author, title, Uniform Resource Locater, abstract, Internatiol Standard of Book Number, year of publication, and disease term or vocabularies. The Semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. In addition, the semantic web has played a significant role to improve the efficiency of different search engines including digital library systems.

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the study

The semantic web is a collaborative movement led by international standards body the World Wide Web consortium. According to the W3C, The Semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries, the main purpose of the Semantic web is driving the evolution of the current web by enabling users to find. Share and combine in formation more easily. The Semantic web, as originally envisioned, is a system that enables machines to understand and respond to complex human requests based on their meaning. Such an understanding requires that the relevant information sources be semantically structured (S. A. Khan, 2015)

Following the advent of modern technology into our lives, mankind leaped to a new level. The internet has played a great role in the evolution of modern lifestyles, which subsequently became a valuable invention and which has made internet accessible and more convenient to its users. Such innovation formed the 'Web'. Nowadays, the Web is a massive information exchange platform that was introduced by Tim Burners-Lee; basically, the idea was to link documents over the internet. For information society, information is economic resources and it affects social life, economics, news, politics, and many other normal life aspects. Now days, with the evolution of the technologies, not only can we connect documents, but we can also understand documents. Currently, the World Wide Web is primarily composed of documents written in HTML (Hyper Text Markup Language), a language that is useful for publishing information. HTML is a set of "markup" symbols contained in a Web page intended for display on a Web browse (Castro, L.J., Giraldo, O.L., & Castro, A.G, 2010)

According to the IFLA/UNESCO Manifesto for Digital Libraries, "a digital library is an online collection of digital objects, of assured quality, that are created or collected and managed according to internationally accepted principles for collection development and made ac- accessible in a coherent and sustainable manner, supported by services necessary to allow users to retrieve and exploit the resources" on the other hand digital libraries contain electronic copies of valuable books, periodicals, documents, maps. According to (Patel-Schneider et al., 2002) digital library can be defined as a quadruple consisting of a repository, a set of metadata catalogues, a set of services, and a society that allow a community of users

to access and re-sue the digital objects. As the digital library universe evolves, its content, data modeling to support, identification, description, and discovery of digital objects are explored to lay the foundations of semantic digital library (Meghini, Spyratos& Yang, 2010).

Digital libraries ideally represent the needs of heterogeneous information resources combining the development of complex systems issues such as interoperability among existing data providers, distributed retrieval, and long-term preservation as well as new issues viz., social network models, large-scale computing, micro information and embedded semantics (Candela, 2010).

Recently, most of the existing web-based digital libraries offer search as well as navigation services. Search services are mainly based on a set of metadata (domain keywords, author, title, or journal name) that carries no semantic information for the search engine. Digital libraries may support graphical user interfaces to aid the formation of queries in order to search documents. Queries can contain one or more domain keywords to be searched in the full text document, or in one of the metadata fields. Users can also formulate multi-criteria queries, by allowing more than one search condition. Navigation services supported by digital libraries are mainly based on catalogues of journals or series names, or on subject directories (Kruk., 2014).

As it described by (Berners-Lee et al, 2001) the vision of the Semantic Web is that it will "bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" to achieve this vision the combination of metadata schemes, the use of the Resource Description Framework (RDF), and the creation of ontologies is essential.

The ultimate goal of the Semantic Web is to allow machines to understand the meaning of digital objects, rather than just the key words used to describe them. This will revolutionize the search and retrieval of digital objects, a key function for digital libraries. One of the main benefits of the Semantic Web is that it operates within the current Web environment to add logic and meaning to digital objects (Jiang, X., & Tan, A, 2009).

As per (Jiang, X., & Tan, A, 2009) describe the core of Semantic Web is the use of ontologies. Ontologies "serve as the backbone of the Semantic Web by providing vocabularies and formal conceptualization of a given domain to facilitate information sharing

and exchange Ontology's are basically sets of logical rules that define the relationships between sets of concepts.

The research intended to explore the challenge of digital library community and information professionals in collecting, organizing, classifying and managing electronic resources for easy access. For the information societies, information is economic resources, thus why peoples are using and conducting digital library collection for satisfying information need. But, the means of searching and accessing information is still do not seem to be sufficient to obtain quality information as per user's queries. Today with classical information overload, searching information, everywhere is becoming a habit also in digital libraries. However, finding exactly what it is intended to search is remains a very hard job. To overcome this bottleneck, study of semantic web technology and its tools is very essential for digital library community and information professionals. In general, using semantic web and incorporating metadata into data sources to enable the end-user to find items and contextually relevant information become important.

1.2. Statement of the problem

According to (Senthil et al, 2006), Semantic Web is the medium to interchange data from one system to another system in a well-defined manner. It is a mesh of information linked up to be easily process able by machines, on a global scale and efficient way of representing data on the World Wide Web, or as a globally linked database. This powerful tool can be used by Digital Library applications for data processing and efficient information retrieval in addition to interchange the data globally.

Today, more libraries rely on electronic sources for collecting, organizing and distributing information. However, the challenge is that it is difficult to use this vast data in the ways that one might want to do or use for their routine work, Because of Digital libraries offer access to large amounts of content in the form of digital documents. Many of them have evolved from traditional libraries and concentrated on making their information sources available to a wider audience, e.g., by scanning journals and books, thereby only taking limited advantage of the modern computing technologies offer (Kamlesh et al, 2016).

Initially artificial intelligence techniques were applied for knowledge re-engineering to achieve the semantics, but maintaining the knowledge base is not an easy task either. Even more the manifestation of digital libraries (DLs) as one of the important knowledge-based

system has thrown up potential challenges for knowledge acquisition as well as integrating with intelligent applications (d'Aquin et al., 2008).

Moreover, Web need knowledge organization tools to keep it less noisy and more precise in the light of semantic heterogeneity of information resources and semantic complexity of knowledge representation (Shapiro, 2010).However, Data interoperability and sharing is another issue that must be faced when developing tools concerning digital library: often, contents and archives should be shared across different platforms and applications, usually by means of a Web-based infrastructure.

As noted by (Kamlesh et al, 2016) typical digital libraries usually focus on categorizing and cataloguing resources. Information retrieval in such libraries relies primarily on text search engines and frees browsing. This approach proved to be useful however, it suffers from ambiguity of natural language, neglecting the importance of metadata; it also does not engage users in the process of sharing knowledge. Simple searching still returns too many results which have to be filtered somehow for easy access. Page ranking algorithm helps with websites but cannot be easily applied to books or e-learning objects. On the other hand, having a look on a friend's bookshelf can give us much clearer view on what is worth reading in a particular domain than digging through a thousand books or websites published this month. The Semantic digital library is an attempt to restore the collaborative approach for sharing knowledge.

In this digital age, there is a growing need of libraries and cultural institutions to cooperate with each other and to expose cultural artifacts and digital contents to a broader audience using a common digital library system. Metadata serve as basis for search and discovery services, it is essential to establish uniform access to metadata provided by the various involved institutions and to provide machine-support for the end users in their search for information (Sebastian R et al, 2014).

Paradigm shift from traditional digital library to social semantic digital library opens innovative and fresh possibilities to define digital library landscape. Contemporary digital library is not merely considered as a digitized collection with information management tools rather than digital library creates an environment to bring together collections, services, and people in support of the full life cycle of creation, dissemination, use, and preservation of data, information, and knowledge (Kalinichenko, 2015). But, ensuring resource discovery

and navigation across heterogeneous resources become challenges for information professionals to enable digital libraries more interactive, relevant and social by facilitating "improved navigation and retrieval within heterogeneous document environments, user profiling, personalization and contextualization, improved user interfaces and human-computer interaction" (Macgregor, 2008).

Semantic web is an evolving extension of the WWW in which web content can be expressed not only in natural language, but also in a form that can be read and used by software agents, thus permitting them to find, share and integrate information more easily (Berners-Lee et al, 2001). It is a vision on imposing intelligence on computes in finding, sharing and combining data so as to arrive at relevant solution sets. The aims of the semantic web is Organization of knowledge in conceptual spaces according to its meaning, usage of automated tools to support maintenance by checking for inconsistencies and extracting new knowledge, replacement of keyword-based search by query answering: requested knowledge will be retrieved, extracted, and presented in a human friendly way and query answering over several documents will be supported, Thus, semantic web technologies sponsors not only use of raw information, but also constellations of information for knowledge processing and creation of new set of knowledge based on existing one. It also asks for representation of knowledge in a machine process able format for making possible commitments lacking in the current web. Semantic web and Ontology is widely viewed as the backbone to support various types of information management including information retrieval, storage, and sharing on web (Abdulelah et al, 2017).

Digital library users are more demanding than web search engine users. In many cases digital libraries are considered as an outreach of classical libraries. In order to reach the status of trust and reliability classical "bricks and mortar" libraries have digital libraries that requires effective information access facilities (Kruk., 2014) So that, integrating semantic web technology standard and ontology in to digital library services is gaining momentum as it facilitates better data exchange and access to information.

To this end, this study tried to explore and answer the following research questions based on the problem stated on the semantic web technology and web ontology language for digital library:

- What is the role of semantic web technology and web ontology language for better digital library services?
- How digital library content management can benefit from the semantic web technology?
- How semantic descriptions of resource and ontology improve the usability of digital library Collections?

What are the semantic web technology standards used in digital library to solve challenges of Large Volume Knowledge Management?

1.3. Objective of the Study

1.3.1 General Objective

The general objective of this study is to analyze the semantic web technology and web ontology-based language application for sharing knowledge through digital library service.

1.3.2 Specific Objectives

The specific objectives of the study was include the following

- To identify the role of semantic web technology and web ontology language in digital library service
- To investigate the semantic descriptions of resource and users' semantic search in improving the usability of digital library service.
- To select the best algorithms and techniques to integrate semantic web technology standards and tools to digital library.
- To design a framework for a semantic web and ontology languages use in digital library service
- To develop a prototype semantic web and ontology language for communicable disease that can serve as a model for digital library.
- To evaluate the performance of the developed prototype system for digital library service.

1.4 Scope and limitation of the Study

This study focused on analyzing the importance of semantic web technology and ontology for better digital library service using resource description framework (RDF) and web ontology language (OWL). Therefore, this study only deal with electronic resource collections in text

format by excluding other types or format, such as images, videos, and gestures. To gain familiarity and understanding with the concept this study deals with existing digital library ontology. Regarding the design and implementation of any study, there are certain inherent limitations, once the domain of study is embracing and the study need to focus on one part of particular area. The methodology chosen for the purpose of this research allows the representation of digital library ontology in particular domain, although the choice of the methodology not be the most appropriate to solve digital library searching challenges

In addition to this according to (Ján Hreňo et al, 2010), metadata is the core of digital library service to provide patrons information about the documents and supporting users for better access of the information. Digital library deals with heterogeneous metadata such as syntactic, structural and semantic metadata. However, this study focused on semantic web and same metadata elements.

1.4.1 Significance of the study

Libraries are a key component of the information infrastructure indispensable for education. They provide an essential resource for students and researchers for reference and for research. Metadata has been used in libraries for centuries. As traditional libraries are increasingly converting themselves to digital libraries, a new set of requirements has emerged (Jorge C, et al, 2014). So, the study on semantic web technology and ontology languages analysis is very important to overcome digital library service challenges. One important feature for digital libraries is the availability to efficiently browse electronic catalogues browsed. This requires the use of common metadata to describe the records of the catalogue (such as author, title, and publisher) and common controlled vocabularies to allow subject identifiers to be assigned to publications. This study also will have significant for initiating information professional and researchers from the domain of library and information science, especially those in digital libraries. In addition to this, the study is significant to save the time of users while information seeking, gathering and processing.

1.4.2 Definition of Terms

Concept - Concepts are the units of thought, ideas, meanings, of objects and events which

Underlie many knowledge organization systems. Concepts exist in the mind as abstract Entities which are independent of the terms used to label them («Simple Knowledge Organization System», 2015).

Controlled vocabulary - is an organized arrangement of words and phrases, to provide a way to organize knowledge used to index content and/or for subsequent retrieval through browsing or searching. It typically includes preferred and variant terms and has a limited scope or describes a specific domain. (Trust, 2009 and «Controlled vocabulary », 2015).

Digital Library (DL) - are organizations that provide the resources, including the Specialized staff, to select, structure and offer intellectual access to interpret, distribute, preserve the Integrity and ensure the persistence over time of collections of digital works (DFL,1998).

Digital Library Management System (DLMS) - is a generic software system that provides the appropriate software infrastructure both to produce and administer a Digital Library System, as well as to integrate additional Software Components offering more refined,Specialised or advanced functionality (Candela et al, 2011).

Digital Library System (DLS) - a software system based on a given architecture and providing all the functions required by a particular DL.

Extensible Markup Language (XML) - is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable. It is flexible way to create common information formats and share both the format and the data on The World Wide Web. XML is a formal recommendation from the W3C («Extensible Markup Language », 2014).

Knowledge Organization Systems (KOS) - is a generic term used in knowledge organization about authority lists, classification systems, thesauri, topic maps, ontologies. («Knowledge Organization Systems », 2015).

Linked data - linked data describes a method of publishing structured data so that it can be

Interlinked and become more useful through semantic queries. It builds upon standard Web Technologies such as HTTP, RDF and URIs, data from different sources to be connected and Queried («Linked Data », 2015).

Ontology - is an organized way of designing, categorizing, helping and explaining the Relationships between the various concepts in the same area of knowledge domain and research.

Resource Description Framework (RDF) - is a standard model for data interchange on the Web. It has features that facilitate data merging even if the underlying schemas differ, and it Specifically supports the evolution of schemas over time without requiring all the data Consumers to be changed (W3C, 2004).

Resource Description Framework schema (RDFs) - provides a data-modeling vocabulary for RDF data. RDF Schema is an extension of the basic RDF vocabulary Semantic Web - is a Web of Data. This data can be dates and titles and part numbers and Chemical properties and any other data one might conceive of. The collection of Semantic Web Technologies (RDF, OWL, SKOS, SPARQL, etc.) enables people to create data stores on the Web, build vocabularies, and writes rules for handling data and also, provides an environment Where application can query that data, draw inferences using vocabularies (W3C, 2014) .

Simple Knowledge Organization System (SKOS) - is a World Wide Web Consortium (W3C) Recommended standard for representing and publishing Knowledge Organization Systems (KOS) on the Web, using a vocabulary and data model expressing Knowledge Organization Systems (KOS's) such as thesauri and classification schemes for referencing and re-use in Semantic Web applications (W3C, 2009)

Taxonomy - in general mode it is the practice and science of classification of things or Concepts, including the principles that underlie such classification; basically it is a controlled Vocabulary organized in a hierarch («Taxonomy », 2015).

Thesaurus - is a reference work that lists words grouped together according to similarity of Meaning or in order words it is taxonomy with more information about each concept

including Preferred and alternative terms. Additionally a thesaurus may contain relationships to related Concepts ((«Thesaurus », 2015).

Web Ontology Language (OWL) - is a language for defining and instance ontologies in Web. This includes descriptions of classes and their properties and their relationships. OWL was designed for use by applications that need to process the content of information, instead of just presenting it to humans. It further facilitates the possibility for interpretation by machines of Web content by providing additional vocabulary with a formal semantics. OWL is a W3C Recommendation («Web Ontology Language », 2014).

World Wide Web Consortium (W3C) - is an international community that develops open

Standards to ensure the long-term growth of the Web.

1.4.3 The structure of this thesis

Chapter 1: **Introduction provides** the background of the research topic and statement of the problem. This is followed by a short introduction about DL, their needs and challenges and also about ontologies. It includes the aim and specific objective, the significant, limitation of the study.

Chapter 2: Literature Review is dedicated to exploring the DL and its challenges, and also explores the basic concept of ontologies. It is built in a form of a theoretical framework that should help conducting this study.

Chapter 3: Methodology describes the methodology chosen to implement this study and provides a justification for that choice this chapter presents proposed approach and architectures including ontology development process.

Chapter 4: experiment and evaluation, presents the process of Experiments and data set preparation for ontology of communicable desease including discussion on finding result and prototype developments.

Chapter 4: Conclusions and future work: offers conclusions to the research question presented at the beginning of the thesis. Finally, some suggestions are made for the directions that might be taken by future researchers focusing on this topic.

Chapter two

2.1. Literature Review

The trend and behavior of publishing and manipulating information have been evolved as more technological innovations which are developed lead by technology companies such as Google, YouTube, Yahoo, Face book, Twitter, and etc. similarly, accessing information also has changed, and more people are relying on the Web as a primary source of information. This information can be obtained from different places such as websites, blogs, online publications, social networks, and databases (Patel-Schneider et al., 2002)

Nowadays, with the evolution of the technologies not only documents connection understanding the document is also possible. Documents in general, come in three categories:-The first type is structured documents. In structured document the formation of the document and the inner data are structured in a way that for each piece of information, it is explicitly known how that piece of information fits with the other data. This leads to the retrieval of more relevant data. Examples of structured documents include databases or spreadsheets; the second type of documents is semi-structured documents. In this type of document, the data are structured, at least in part, based on semantics, but the underlying structure and the semantics are not explicitly given. This kind of document can produce relatively rich semantic information because the semantics is embedded in data and the data's structure. Examples of semi-structured documents are HTML documents, Word Net, and XML documents. The third type of documents is unstructured documents where the knowledge is available only in the data and not in the structure of the documents. for e.g., standard text files (Abdulelah et al, 2017).

According to Laney, Gandomi and Haider (2015), the Web, as an information source, is holding enormous amounts of information in variety document structures. To automate this process, absorbing this diversity is vital to understand the current trend of the improvement. for instance, the Big Data aims to effectively assist in efficient data processing that beyond trivial computing power especially with "Variety" and "Variability" characteristics.

2.1.1. Concept of Digitization

According to International Encyclopedia of Information & Library Science (2003), "Digitization of information material is the process of converting analogue information to a digital format" (p.138). Nowadays, libraries adopt digitization with the purpose of preserving information and dissemination knowledge. There are many reasons for libraries to go for digitization but the main profit is to preserve the rare and fragile objects; especially these items of high quality such as old manuscripts (Conway, 2010). As the material digitized can be easily accessed by anyone; libraries, institutions, individuals; from anywhere at any time without hindrances Fabunmi (2006) recognizes three reasons for digitization: (a) preservation of endangered library resources, (b)efficiency of information search mechanism, (c)improvement of access to library resources. Whilst, Mauyra (2011) adds to these: (a) the new generation needs, (b) reduction, and (c) the preservation of the virginity of the environment.

2.1.2. Academic Libraries in Digital Era

In the 21st century, academic libraries have a new role in sharing information. Libraries are not piles of books anymore; the general library environment has been changed from analogue to digital. Library automation systems have helped libraries to provide easy access to their collections through the use of computerized library catalogues (On-line Public Access Catalog – OPAC) which more recently led to digital libraries (IFLA,2013).

With rapid spread of electronic resources (i.e., E-resources) in the field of Information Science & Information Technology at the recent times, the sharing of information through digital sections has become an attractive idea for Librarians and the need for digital libraries to academic universities has been emerged so as to improve educational development and grewprovide online educational resources to students and scholars for effective learning. According to the UNESCO Institute for Information Technologies in Education (2003) digital libraries render educational resources for E- learning.

Dissemination of knowledge has always been one of academic libraries primary goals as long as they have served as learning institutions, cultural repositories and research centers but in this digital age of knowledge, they have to expand these roles and stop being passive repositories for printed material. To the contrary, academic libraries should upgrade their services and providing education of high quality by storing resources in various forms and maintaining easily accessible for online use among academic community. This would be

possible by implementing a digitization project. Hughes (2004) reported to concept of digitization as "the process by which analogue content is converted into sequence of 1s and 0s (these ones and zeros are called bits) and put into a binary code to be readable by a computer" (p.4). So, we could generally say that the digital material is every computer readable material.

According to Pandey and Mishra (2014) "Academic libraries are digitizing materials because they know the continuing value of library resources for learning, teaching, research, scholarship, documentation, and public accountability.

Academic Libraries are an important part of the National Educational System as long as they serve as information centers fulfilling basic library user's needs; efficiency, effectiveness, and utility. Evaluation plays a key role in the improvement of information services. Digital era has produced many changes in the society such as expansion of the served community; products and services; the need to break the space-time barrier in communication and further expectations from library users for high-quality and user-friendly online services. Therefore, the library manager has to impartially evaluate qualitative and quantitative value of the library resources and plan services for better functionality in order of making the invisible to be visible.

2.1.3. Semantic web

Semantic Web (SW) is other types of web concept that was first introduced by (Abdulelah et al, 2017) vision of the Semantic Web is that it will "bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" This vision is supported by the combination of metadata schemes, the use of the Resource Description Framework (RDF), and the creation of ontologies. The ultimate goal of the Semantic Web is to allow machines to understand the meaning of digital objects (Patel-Schneider et al., 2002)

Many cognitive scientists have worked earlier on how to structure the knowledge semantically and enable the automated agents to access the web more intelligently and perform the work of the users on their behalf. Accordingly, descriptive technologies such as XML, RDF, and OWL have been developed to address the limitations in using HTML. XML (Extensible Markup Language) provides a method for transmitting structured documents.

According to (Abdulelah et al, 2017), during the first decade of its existence, most of the information on the Web is designed only for human consumption. Humans can read Web pages and understand them, but their inherent meaning is not shown in a way that allows their interpretation by computers. The information on the Web can be defined in a way that it can be used by computers not only for display purposes, but also for interoperability and integration between systems and applications. One way to enable machine-to-machine exchange and automated processing is to provide the information in such a way that computers can understand it. This is precisely the objective of the semantic Web – to make possible the processing of Web information by computers.

As noted by Tim Berners Lee, the inventor of the Web, semantic web needs structured documents that a machine can process by querying or differencing to derive more precise information, the term Semantic Web is coined for a Web of meaning that makes it understandable to machines rather than just readable by machines. As such, the Web has been developing toward this vision by embedding huge quantities of machine-process able metadata, structure and different semantic Web technologies into the current Web. Basically, the Semantic Web tries to shift the thinking of published data in the form of Web Pages (i.e., HTML documents) to allow machines to understand the contents. The content of Web 1.0 and Web 2.0 suffered from a number of issues, including the amount of information and how to access it and enable delegation . They were provided for humans rather than for comprehension by machines. Therefore, it was not easy to automate data across the Web. To this end, the key idea behind the Semantic Web is to identify and link the content of the Web in a way that allows machines to understand and derive meaning from the data. Recognizing this vision requires new approaches, languages, technologies and data representation models to be built , for this reason, a variety of semantic languages and standards are maturing, and different applications, tools, and services are developing.

2.1.4. Semantic web standards and technology

As already stated, the Web was originally a vast set of static Web pages linked together. Due to the widespread importance of integration and interoperability for intra- and inter-business processes, the research community has tackled this problem and developed semantic standards such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL). RDF and OWL standards enable the Web to be a global infrastructure for

sharing both documents and data, which make searching and reusing information easier and more reliable as well. RDF is the W3C standard for creating descriptions of information, describing their semantics and reasoning, especially information available on the World Wide Web (Patel-Schneider et al., 2002)

According to (Patel-Schneider et al., 2002), the main purpose of using XML is for syntax, where RDF is for semantics. Both share a unified model and together provide a framework for developing Web applications that deal with data and semantics. As it described by Senthil et al, 2006) Relationships are at the heart of semantics, perhaps the most important characteristic of RDF is that it elevates relationships to first class object, providing the first representational basis for giving semantic description. Hence RDF is also well suited for representing metadata for Web resources. OWL provides a language for defining structured Web-based ontology's which allows a richer integration and interoperability of data among communities and domains.

.

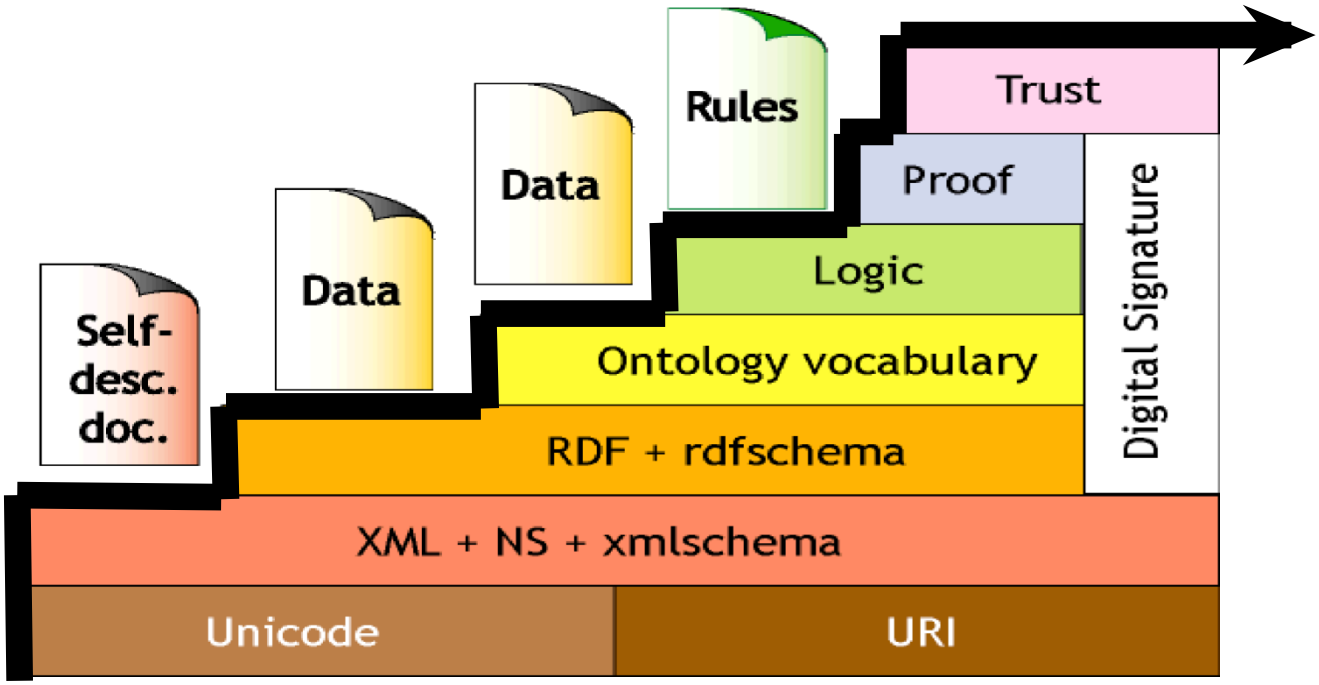


Figure 1 Semantic Web Layer Cake (Berners-Lee, 99; Swartz-Hendler, 2001).

- **Unicode and URI:** Unicode is used to represent of any character uniquely whatever this character was written by any language and uniform Resource Identifier (URI) is unique identifiers for resources of all.
- **XML**: It is a language that lets one write structured web documents with a user defined vocabulary XML is particularly suitable for sending documents across the web. XML has no built-in mechanism to convey the meaning of the user‘s new tags to other users.
- **RDF**: Resource Description Framework is a basic data model, like the entity relationship model, for writing simple statements about web object. RDF provides the technology for expressing the meaning of terms and concepts in a form that computers can readily process.
- **Logic Layer**: It is used to enhance the ontology languages further and to allow the writing of application-specific declarative knowledge.
- **Proof Layer**: It involves the actual deductive process as well as the representation of proofs in web languages and proof validation.
- **Trust Layer**: It will emerge through the use of digital signatures and other kinds of knowledge based on recommendations by trusted agents or on rating and certification agencies and consumer bodies. Semantic web is not limited to publish data on the web. It is about making links to connect related data.

2.2. Metadata

Metadata can be defined as “data about data.” Accordingly, the goal of incorporating metadata into data sources is to enable the end-user to find items and contextually relevant information. Data sources are generally heterogeneous and can be unstructured, semi-structured, and structured. In the semantic Web, a data source is typically a document, such as a Web page, containing textual content or data. Of course, other types of resources may also include metadata information, such as records from a digital library. The process of attaching semantic metadata to a document or any piece of content is called semantic. Semantic applications are created by exploiting metadata and ontology with associated knowledgebase. In essence, in the semantic Web, documents are marked up with semantic metadata which is machine-understandable about the human-readable content of documents (Fisher, M et al, 2004)

2.2.1. Syntactic Metadata

As it described in this study, the simplest form of metadata is syntactic metadata. It describes non-contextual information about content and provides very general information, such as the document's size, location, or date of creation. Syntactic metadata attaches labels or tags to data. The following example shows syntactic metadata describing a document:

<name> = "report.pdf"

<creation> = "30-09-2005"

<modified> = "15-10-2005"

<size> = 2048

According to (Chawner, 2008)Most documents have some degree of syntactic metadata. For example, E-mail headers provide author, recipient, date, and subject information. While these headers provide very little or no contextual understanding of what the document says or implies (assuming value of author is treated as a string or ordered sets of words, rather than its full semantics involving modeling of author as a person authoring a document, etc.), this information is useful for certain applications. For example, a mail client may constantly monitor incoming e-mail to find documents, related to a particular subject, with his/her interests.

2.2.2. Structural Metadata

Structural metadata provides information regarding the structure of content. It describes how items are put together or arranged. The amount and type of such metadata will vary widely with the type of document. For example, an HTML document may have a set of predefined tags, but these exist primarily for rendering purposes. Therefore, they are not very helpful in providing contextual information for content. Nevertheless, positional or structural placement of information within a document can be used to further embellish metadata (e.g., terms or concepts appear in a title may be give higher weight to that appearing in the body). On the other hand, XML gives the ability to enclose content within more meaningful tags. This is clearly more useful in determining context and relevance when compared to the limitations of syntactic metadata for providing information about the document itself. The following lines, extracted from a DTD, describe a set of valid XML documents:

<!ELEMENT contacts (contact*)>

<!ELEMENT contact (name, birthdate)>

<!ELEMENT name (#PCDATA)>

<! ELEMENT birthdate (#PCDATA)>

Structural metadata tell us how data are grouped and put in ordered arrangements with other data (Jorge Cardoso et al, 2006) (Jorge Cardoso et al, 2006).

2.2.3. Semantic Metadata

Semantic metadata adds relationships, rules, and constraints to syntactic and structural metadata. This metadata describes contextually relevant or domain-specific information about content based on a domain specific metadata model or ontology, providing a context for interpretation. In a sense, they capture a meaning associated with the content. If a formal ontology is used for describing and interpreting this type of metadata, then it lends itself to machine process ability and hence higher degrees of automation. Semantic data provides a means for high-precision searching, and, perhaps most importantly, it enables interoperability among heterogeneous data sources. Thus, why semantic web comes to core for digital library, Semantic metadata is used to give meaning to the elements described by the syntactic and structural metadata. These metadata elements allow applications to "understand" the actual meaning of the data (Senthil et al, 2006)

2.3. The Structure of Information and Sets of Digital Objects in Digital Library

Libraries are a key component of the information infrastructure indispensable for education. They provide an essential resource for students and researchers for reference and for research. Metadata has been used in libraries for centuries. For example, the two most common general classification systems, which use metadata, are the Dewey Decimal Classification (DDC) system and the Library of Congress Classification (LCC) system. As traditional libraries are increasingly converting themselves to digital libraries, a new set of requirements has emerged. One important feature for digital libraries is the availability to efficiently browse electronic catalogues browsed. This requires the use of common metadata to describe the records of the catalogue (such as author, title, and publisher) and common controlled vocabularies to allow subject identifiers to be assigned to publications. The purpose of the

information architecture is to represent the riches and variety of library information, using the building blocks of the digital library system (Jorge C, et al, 2014)

From a computing view, the digital library is built up from simple components, notably digital objects. A digital object is a way of structuring information in digital form, some of which may be metadata, and includes a unique identifier, called a handle. However, the information in the digital library is far from simple. A single work may have many parts, a complex internal structure, and one or more arbitrary relationships to other works. To represent the complexity of information in the digital library, several digital objects may be grouped together and this is called a set of digital objects. All digital objects have the same basic form, but the structure of a set of digital objects depends upon the information it represents. To enable the content to represent useful information, its type must be known. Thus part of the content may be of type text (perhaps encoded in a mark-up language), while another part may be of type audio. Semantic Web technologies, such as RDF and OWL, can be used as a common interchange format for catalogue metadata and shared vocabulary, which can be used by all libraries and search engines (Shum, Motta et al. 2000) across the Web.

2.3.1. How Semantic Web Relates to The Www on Dl.

Semantic web linked to digital library on World Wide Web. Currently, the World Wide Web is based primarily on documents written in HTML, is useful for describing, with an emphasis on visual presentation, a body of structured text intersperse with multimedia objects and interactive forms., HTML has limited ability to classify the blocks of text on a page, apart from the roles they play in a typical document's organization and in the desired visual layout. For example, with HTML and a tool to render it, one can create and present a page that lists the library catalog which proved only document-level assertions such as "this document's title is 'DSL Library Catalog' however, The Semantic Web addresses this shortcoming, using the descriptive technologies RDF and OWL, and the data-centric, customizable markup language XML. These technologies are combined in order to provide descriptions that supplement or replace the content of web documents. Thus, content may manifest as descriptive data stored in web-accessible databases, or as markup within. The machine-readable descriptions enable content managers to add meaning to the content, thereby facilitating automated information gathering and research by computers (Senthil et al, 2006).

<Bookstore>

Make up your own tags

<Book ID=“101”>

<Author>John Doe</Author>

<Title>Introduction to XML</Title>

<Date>12 June 2001</Date>

<ISBN>121232323</ISBN>

<Publisher>XYZ</Publisher>

Sub-elements

</Book>

<Book ID=“102”>

<Author>Foo Bar</Author>

<Title>Introduction to XSL</Title>

<Date>12 July,2019</Date>

<ISBN>12323573</ISBN>

<Publisher>ABC</Publisher>

</Book>

</Bookstore>

XML by itself is just hierarchically structured text

By publishing ontology’s, which can then be accessed by all users across the Web, library catalogues can use the same vocabularies for cataloguing, marking up items with the most relevant terms for the domain of interest. RDF and OWL provide a single and consistent encoding so implementers of digital library metadata systems will have their task simplified when interoperating with other digital library systems.

2.4. Web Ontology

Ontology is a formal description of knowledge as a set of concepts within a domain and the relationships that hold between them. To enable such a description, we need to formally specify components such as individuals (instances of objects), classes, attributes and relations as well as restrictions, rules and axioms. As a result, ontologies do not only introduce a sharable and reusable knowledge representation but can also add new knowledge about the domain. The ontology data model can be applied to a set of individual facts to create a knowledge graph – a collection of entities, where the types and the relationships between them are expressed by nodes and edges between these nodes, By describing the structure of the knowledge in a domain, the ontology sets the stage for the knowledge graph to capture the data in it. In recent years, there has been an uptake of expressing ontologies using ontology languages such as the Web Ontology Language (OWL). OWL is a semantic web computational logic-based language, designed to represent rich and complex knowledge about things and the relations between them. It also provides detailed, consistent and meaningful distinctions between classes, properties and relationships (S. A. Khan, 2015).

2.4.1. The Role of Ontology's in Semantic Digital Libraries

Ontologies "serve as the backbone of the Semantic Web by providing vocabularies and formal conceptualization of a given domain to facilitate information sharing and exchange" (Jiang, X., & Tan, A, 2009) Ontology is basically sets of logical rules that define the relationships between sets of concepts. Ontology can also be used to define the concepts and the XML codes used to designate them within Web pages (Berners-Lee et al, 2001). Ontology is usually domain specific, and attempt to exhaustively define concepts and map their relationships. The goal of ontology is to "reinforce the formal logic and tighten the 'meaning' to the point where it can no longer escape correct computer interpretation" .

The combination of XML, RDF, and ontology makes it possible for computer programs and agents to interpret digital objects based on their meaning and their relationship to other concepts. The benefits of using these technologies in digital libraries are myriad. Furthermore, digital librarians already possess a skill set that enables them to work with Semantic Web technology. Digital librarians are already working with XML to create metadata, and the creation of ontologies is akin to the development of a thesaurus (Sebastian R et al, 2014)

2.4.2. Ontology libraries

Similar to the idea of a digital library (Chowdhury, G.G., & Chowdhury, 2002) in the recent time, a new concept has emerged called an ontology library. Principally both a digital library and an ontology library have the similar kinds of purposes and objectives to achieve (e.g., store, organize and provide access to the digital objects), except the kind of materials they deal with. While a digital library deals with the documents of various types, such as text, audio, video, images, etc., an ontology library deals with the ontologies (where an ontology is an intelligent object, often referred as a digital artifact, consisted of representations of the entities in terms of their types, properties and the relationships of a domain of discourse. The ontology libraries have other similar terminologies such as ontology repository, ontology registry, and ontology directory. An ontology repository is a facility where ontologies and related information (e.g., ontology metadata including the information on who included the ontologies) are stored, managed and retrieved.

Ontology directory is a service that provides the information about ontologies that are available in a particular platform. Ontology directory contains the reference to the definition of ontology related language-based schema such as OWL, XML, and RDF Schema. Where the ontology registry provides a data storage interface where data, knowledge, metadata of a semantic object including the name of the person or any community and process of using data are registered.

2.4.3. Enhancing Digital Libraries with the Annotation Ontology

Semantic Digital Libraries (SDL) makes extensive use of meta-data in order to support information retrieval and classification tasks. Within the context of SDLs, ontologies can be used to: (i) organize bibliographic descriptions, (ii) represent and expose document contents, and (iii) share knowledge amongst users (Patel-Schneider et al., 2002). There have been some efforts aiming to make use of ontologies and Semantic Web technology in digital libraries that allows users to semantically annotate books, papers, and resources. Ontologies have shown to be useful for supporting the semantic annotation of scientific papers and thereby facilitating information retrieval tasks. However, as ontologies are often incomplete users should be able to provide additional metadata. This can be achieved by annotation of ontologies to structure and classify research articles. Annotations may be rooted in existing ontologies or provided by users; that support the tagging of atomic components within papers –e.g., words, tables, figures. The content of the paper and the corresponding tags are being

presented as linked data, this facilitates the interoperability between the paper and external resources –e.g., databases, repositories for experimental data, etc. Annotations are used to improve search and retrieval of papers; it also makes possible to find related papers and researchers. The search & retrieval module is based on that one usually provided by digital libraries; it uses clouds of annotations and annotators to facilitate navigation and filtering. In this regard annotating library resources using ontology could help to retrieve and search documents easily.

2.5. Related Work

In these section different attempts on different applications for digital library systems has been explored widely based on the previous studies conducted in the area.

(S. A. Khan, 2015) explores the Semantic Web and Ontology-based applications for future Semantic Digital Libraries. They have presented those digital libraries leverage various semantic web technologies including RDF schemas. They have also indicated that content analysis of interview responses showed that future Semantic Digital Libraries will provide precise and effective results and meet the user's need in effective way.

In (G. T. ,. T. A. ,. A. Norbert Fuhr, 2007) a detailed review of digital library metrics and methods has been explored. It's believed to be that digital library information varied based on structure and of institutions, type of information systems, new technologies used, as collections, or as new services. The evaluation has been made based on users, contents and information access. The main reason for developing digital library is the content because it clearly addresses the user's information needs. The relation between the user and the content strongly depends on the informational need of the user. Standard metrics used for digital library system is similar to other information extraction systems such as precision, recall and f-measure. The major categories of evaluation wings such as usability, usefulness and performance have been used.

(C. M. Owusu-Ansah, 2019) Explains conceptions of digital libraries in the context of Africa from the perspective of historical literature. They have reviewed publicly-available literature on the theme of digital libraries from both the Western and African perspectives. The analysis revealed that most of the initial digital library initiatives in Africa emanated from the west with African countries benefiting from international initiatives to expand access to information resources to bridge the global digital divide.

In (Senthil et al, 2006) has proposed a new ontology alignment approach based on the combination of word embedding and the radius measure. The alignment process is based on a set of rules exploiting the word embedding similarity to discover the alignment. They started by extracting two types of information from inputs: lexical information (e.g., labels of concepts) and structural information (e.g., to associate the labels of all child entities to their parent entities). Then the second step of our approach is to compute the vector representations of concepts. Every concept has been matched between source ontology O1 with the similar concept in the target ontology O2 using the cosine similarity between vector representations of concept and cluster.

(L. Guo, Y. Zhou, 2010) Introduce an ontology-based visualization model to integrate ontology in historical knowledge searching. Their ultimate goal was assisting knowledge representation and knowledge inference for the purpose of knowledge management. A skeletal approach has been used to construct the ontology which contains some guidelines without detail steps. In their work a total six ontologies has been defined including people, place, organizations, events, and resources and times following ontology database has built. During the ontology construction a single sentence is treated as a piece of text with its demarcation and grouped based on predefined principles to annotate each word. Apache Lucene, full text-featured- tool has been used as a searching tool.

(Sreeja, 2012) had conducted a study on "Agent based semantic web", the purpose of the study was to examine how semantic web provide machines better access to information resources so that they can be information intermediaries in support of humans and to build a network of content stored on the web and making it possible for machines to understand data and to satisfy requests from people and other machines. According to (Sreeja, 2012) in order to carry out and accomplish the goal of semantic web tasks intelligent agents must communicate and understand meaning of the data. The agent-based method for semantic analysis enables computers to understand documents written in natural language.

As it explained in the study, the major semantic web services are automatic web service discovery, automatic web service execution, and automatic web service composition and interoperability. According to (Senthil et al, 2006) semantic data provides a means for high-precision searching and perhaps it enables interoperability among heterogeneous data sources. Thus, why semantic web comes to core for digital library and information services? Semantic metadata is used to give meaning to the elements described by the syntactic and

structural metadata. These metadata elements allow applications to "understand" the actual meaning of the data and the study concluded that the integration of agent technology and ontologies has made significant impact on the use of semantic web services.

Digitization's have led in the last decades to a huge evolution in the way digital libraries and archives are conceived, designed, and used. Both the transition of library materials from traditional to digital formats and the large (and continuously growing) availability of digital content pose new challenges. Lynch & Garcia-Molina's (1995) article summarizes the key challenges facing digital libraries, Based on a digital libraries workshop. They identify five key challenges facing digital libraries: Interoperability; description of objects and repositories; collection management and organization; user Interfaces and human-computer interaction; and economic, social, and legal issues they conclude that Digital libraries should explore using Semantic Web technologies to meet their organizational Challenges In this study, the authors of the paper argued that more sophisticated software tools are needed to meet the expectations of users, which are often high due to the classical information overload problem. Searching everything everywhere is becoming a habit also in digital libraries however, finding exactly what is needed remains a very hard job. The flexibility of Semantic Web technology allows for greater interoperability between digital Libraries and collections.

According to research by Morales-del-Castillo, Pedraza-Jimenez, Ruiz, Peis&Hererra-Viedma (2009) "The Semantic Web has a common data model and syntax that guarantee the Interoperability of resources (independently of the platform), thus making easier the establishment of Exchange and collaborative networks between digital libraries" .This flexibility also enhances the Ability to describe objects, to manage collections, to create interactive user interfaces, and to manage access to copyrighted materials.

2.5.1. The Research Gap

Over the last decade the World Wide Web (WWW) has emerged to an important part of our everyday life. Companies, organizations, and people use the WWW as a powerful tool to share information: Companies offer Web pages to advertise and sell their products. Educational institutions present teaching material and online training services on the Web. Public authorities provide e-government services to make administrative tasks more efficient and citizen-friendly. User groups maintain Web portals to exchange information within their community. The popularity of the WWW leads to an exponential growth in the number of

Web pages available in the global information space. This success, however, leads to several problems:

The huge number of available Web pages makes it increasingly difficult for users to find and access required information. In searching the Web for specific information, one gets lost in the huge amount of irrelevant search results and may miss the relevant material. Electronic commerce is currently hampered by the way information is presented. Since the semantics of the Web pages in not directly accessible for machines, for example, shopping agents have to use wrappers and heuristics to extract relevant product information from weakly structured HTML documents to compile a market overview. In addition to this the Use of digital information tools and processes are changing the library environment and library services. Libraries and librarians are required to become more relative with technological environment (Hendrix, 2010).

The technological advancements in libraries are changing the ways for assessing, storing and disseminating information. The new technological environment will transform the present library into new dynamic informatics institution and this system shall provide required information to its users when and where they need it. However, optimizing the management processes of information resources, has becoming increasingly complex, since the task of developing and managing digital libraries, has become too complex, particularly because of their elusiveness and multiplicity of related technologies and information resources available in digital environment. Nevertheless, the increasing number of data sources available hampers the retrieval of information. Therefore, the study of semantic web and ontology for digital library is very important regarding to improving the digital library services.

CHAPTER THREE

3.1. METHODOLOGY

In this chapter the details of methods that has proposed to be employed in order to conduct this study is described. Along with the generic method the details of pipelines used in this study also discussed.

3.1.1 Design science research

In this study design science research is used, because it is a method for describing, explaining, and predicting the individual, organizational, or social effects of technology use. It is concerned with the systematic creation of knowledge about and with design as an intentional, intellectual, and creative activity for problem solving. According to (Peffers, K et al, 2008) , design science research is concerned with the artificial intelligence, i.e. information technology artifacts, can use experiments to thoroughly evaluate design alternatives and identify superior manifestations to bring improvements.

According to Fernandez et al.,(1997), methodologies is generally set out guidelines specifying how the researcher should carry out the activities identified in the development of ontologies, what kinds are the most appropriate techniques for each activity and what products each producing process. There are the steps and processes to follow in the design science research methods. These processes includes: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication (Peffers, K et al, 2008). A number of author such as Fernandez et al. (1997); Lopez, (1999); Corcho (2003); Gonzalez (2005) Yun et al. (2013) and Sawsaa& Lu (2014), recommend that for developing ontologies design science method, because ontologies are part of software products. Therefore, ontologies should be developed according to the standards proposed for software generally, which should be adapted to the special characteristics of ontologies. To conduct the experiment of this study the researcher has setup an experimental environment on googles colab to utilize high performance computing machines (GPU) and freely available limited storage capabilities of Google drives. Besides, in order to construct vectors of documents related to infectious disease vocabularies we could use large vocabularies of existing pre trained models. The researcher used the default settings of google colab in order to run our experiment. In this study, a semantic search technique has been employed in order to evaluate

the effectiveness of proposed method semantic web supported digital library semantic search systems. To do so, the researcher has used a semantic robust sentence embedding techniques along with rich set of dictionaries of infectious disease terms that has been already extracted from domain specific ontologies

3.1.1. Ontology Development process for digital library

In computer science one of the most commonly used ontology definition is from Gruber, ontology is an explicit specification of a conceptualization (Owusu-Ansah, 2019). Explicit in this context means that type of concepts and constraints are explicitly defined and conceptualization refers to an abstract model of some domain with identified relevant concepts of that particular. Ontologies aim at capturing domain knowledge in a generic way and provide a commonly agreed and shared understanding of domain. Ontology Learning aims at learning and extending ontologies on the basis of textual data. In digital library systems the ontologies have been constructed for different domains fundamentally considering major concepts such as; documents and their metadata. In this study, the researcher has followed existing steps to construct ontology for communicable disease domains that aimed to foster the digital library searching trends. The reason we have chosen this particular domain for our study is because of few resources availability to adapt the ontologies on the existing concepts. Since, our objective in this study is to analyze the effect of semantic web technologies mainly ontologies in digital library systems; we are particularly interested on the enhancement of searching strategies in digital libraries for particular domain specific healthcare resources by cascading different approaches along with concept development and matching strategies.

3.2. Semantic search in digital library

Semantic search is a data searching technique in a which a search query aims to not only find keywords, but to determine the intent and contextual meaning of the words a person is using for search. Ontology based semantic search will lead to new generation of search based on the meaning of keyword rather than keyword and helps in finding correct information on the web. Here, ontology provides an explicit specification of conceptualization which helps to connect the information on the existing web pages with the background knowledge. Ontology based search overcomes the semantic gap between the keyword found in documents and those in query. The use of ontology in the search process provides an interaction between machine and human.

Traditional search process is based on term-based searches, retrieving resources conceptually related to the user informational need. Queries can be expressed in several ways, and mapped on the semantic level defining topics that must be retrieved from the web or local sources.

3.3. Other approaches for searching in digital library

The retrieval of resources within the context of a single digital library is a multifaceted process that includes features such as cataloguing, metadata, and indexing. The convolution of a digital library is proportional to the number of aspects (indexing and cataloguing) that are active at the same time when a user searches for contents across many collections that employ various metadata systems. The retrieval of such semantically related documents in digital libraries is not a new problem and there are many approaches to overcome it. For example, in many digital libraries the metadata of documents is enriched by set of tags like keywords or subject categories. However, these meta-data may lack a standardized taxonomy and also suffer from the coverage as it is usually a human annotated process. Besides, there have been also a long-time experience in using keyword based and knowledge-based searching approaches. Following this, there have been also other approaches like topic modeling-based keyword extraction and using these keywords for single document retrieval. On the other hand, these days advanced semantic space text representation techniques such as word embedding approaches have been widely used. Similarly, traditional matrix factorization methods of text-co-occurrence matrix such as Latent semantic analysis (LSA) have been used comparatively.

3.4. Integrated ontology based and embedding searching

Along with the ontology-based searching approaches recently the word embedding based approaches has showed a breakthrough in the research community. To do so, in this study the researcher has proposed a hybrid based digital library searching approach using the power of both ontology-based concept matching and word embedding.

3.5. Proposed Architecture

In this section the study we introduce the proposed architecture for digital library semantic searching framework based on existing attempts of semantic web technology solutions in digital library systems. The proposed model is comprised of various pipelines as explained in detail in the following subsections.

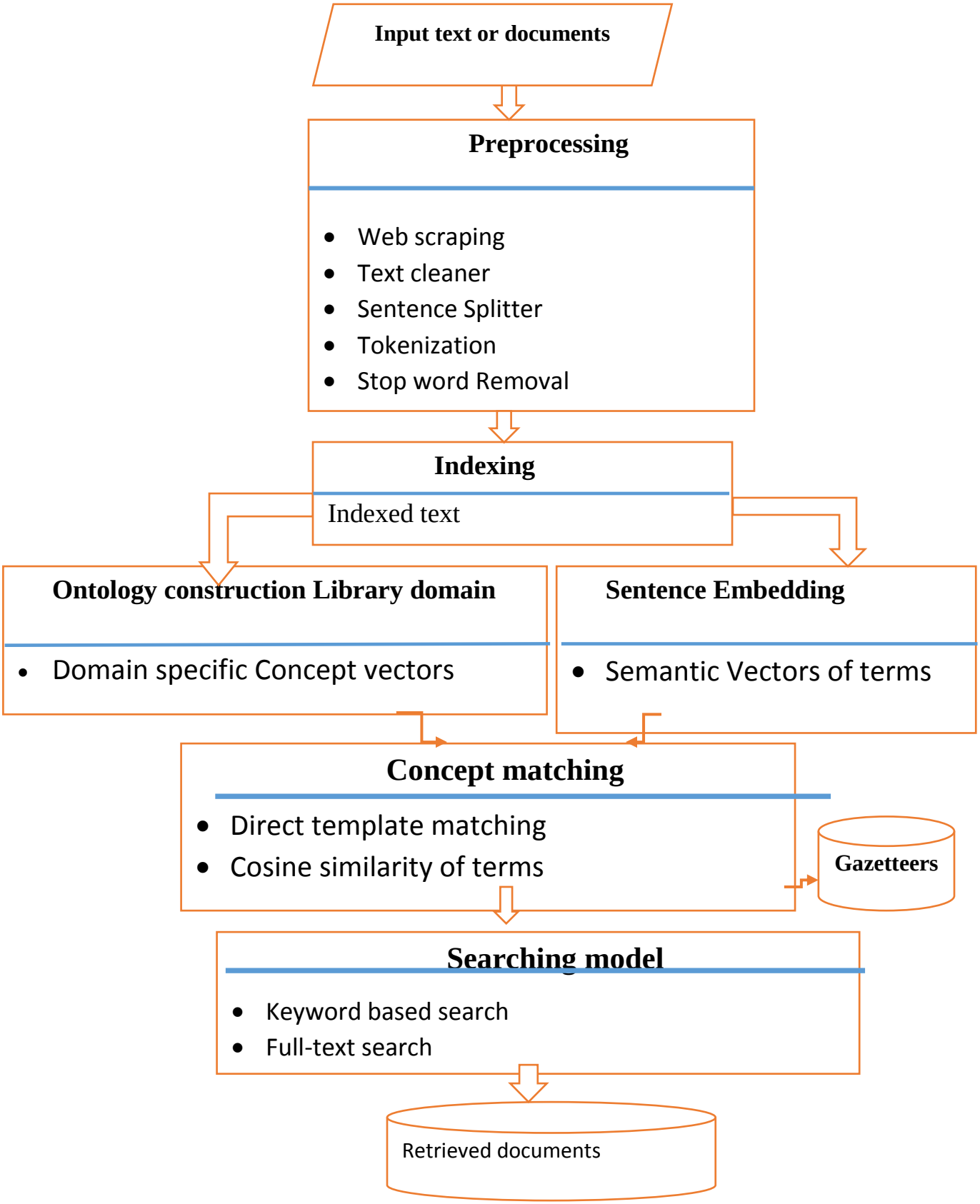


Figure 2 Proposed Architecture

3.5.1. Preprocessing

Preprocessing is the most significant and preliminary task of feature engineering approaches in many webs search engine and other text data manipulation applications which helps to clean and made ready the corpus for the further processing. The main role of data preprocessing is formatting or normalizing the input documents, so that later tasks can be computed easily. The document preprocessing component handles task specific issues that are imposed by the nature of the tasks to make the data ready for remaining pipelines. In this research the researcher has employed various preliminary text preprocessing activities such as stop word removal, tokenization and unwanted token normalization. It's envisaged that removing those relevant contents could possibly improves the indexing and searching process. Ultimately, the searching process would be fast whenever the most representative contents have been clearly filtered.

3.5.2. Indexing

Indexing is the process of storing the information for a faster retrieval upon a search query. A Search engine must maintain all the contents it finds during the crawling process, and store it in an index for the easy retrieval. The goal of indexing is to speed up the matching, without an index the retrieval process would require streaming through the collected web pages. Similarly, in this study an inverted index technique that help to store the list of documents for each word in the source articles were employed.

3.5.3. Ontology construction in case of communicable diseases

Ontology is an important element underlying the semantic web technology. It contains the domain specific knowledge to increase the powerfulness of a particular search engine. The ontology design is very important since it affects the search effectiveness and efficiency. Designing ontology and evaluating it are challenging tasks. Ontology contains several concepts in the object-oriented style. The relationships are the properties which map from a domain to a range and can be viewed as functions or relations. The property characteristics are needed to increase the possibility of the knowledge inference. To develop a domain specific ontology, it requires a domain knowledge. In addition, it's mandatory to understand the object concept, the hierarchy of subclasses, the property mapping, and property characteristics. In this study, the

researcher has have selected a specific domain which is communicable diseases. The study developed a domain ontology for specific information of few communicable diseases. In the digital libraries there are many source documents available with their metadata's and it's one relevant discipline which has gained more attention in ontology development. In this case the researcher has selected this domain because of availability of existing sources of sample ontologies on the web. Readers might explore any sources using short query, key words, and long queries. These short or long queries have been parsed and matched to source contents. In order to match source contents and query strings concepts play great role beside other semantic notation of contents. To do so, concepts of those source contents and query elements could be driven with the help of concept rich domain ontology. Sample domain ontologies for few communicable diseases have been incorporated in experiment section of this document.

3.5.4. Sentence embedding

The initial embedding techniques dealt with only words. Given a set of words, the study would generate an embedding for each word in the set. The simplest method was to one-hot encode the sequence of words provided so that each word was represented by 1 and other words by 0. To represent sentences, in the study the researcher can't one-hot encode them as the study did to words because there are so many possible sentences that it would be unpractical. A possible and straightforward way to create sentence representations is to take advantage of the embedding of each word and calculate the embedding of the whole sentence based on those. Sentence-BERT is currently the state-of-the-art algorithm to create sentence embedding. In this study, the search query and source contents of digital resources such as books and articles have been embedded using sentence embedding techniques. So that the similarity between vectors of query strings and contents have been calculated using different ranking algorithms. Different pre trained semantically dense vector representation such as; Infer Sent, Sentence Transformer and Doc2Vec have been used to encode query strings and digital library source contents manly abstracts, titles and related information. An embedding of query string instances has been constructed using rich vocabularies of pre trained models.

Doc2vec (also known as: paragraph2vec or sentence embedding) is the modified version of word2vec the main objective of doc2vec is to convert sentence or paragraph to numerical vector representation form. Infer Sent is a sentence embedding method that provides semantic

representations for English sentences. It is trained on natural language inference data and generalizes well too many different tasks. The sentence Transformer framework provides also an easy method to compute dense vector representations for sentences and paragraphs (Gruber, 1995). The models are based on transformer networks like BERT / RoBERTa / XLM-RoBERTa etc. and achieve state-of-the-art performance in various tasks. Text is embedding in vector space such that similar text is close and can efficiently be found using cosine similarity. In this case, a simple DistilBERT sentence transformer pre-trained model has been used. Distil- BERT maps sentences paragraphs to a 768-dimensional dense vector space

3.5.5. Concept matching

As the core components of the semantic web ontologies has been played great role as explicit specialization of conceptualization. However, due the design conventions used by different domain ontologies sharing a knowledge and information across various ontologies has become a challenge. To do so, an ontology matching approach and concept matching becomes a remedy to solve this constraint. Different ontology matching techniques such as element-level and WordNet supported matching have been widely used. The element level approach fundamental uses the lexical information or string surface matching strategies though it's limited to explore semantic similarities of words. Similarly, the WordNet approach has been criticized because of its insufficiency of vocabularies. As a result, in this research we have employed the word embedding based concept matching techniques to acquire the semantic similarity of concepts without scale constraint of WordNet and limited lexical information-based approaches.

3.5.6. Searching in digital library

Semantic search is search with meaning other than static dictionary meaning of user's query in which that seeks to improve search accuracy by understanding a searcher's intent through contextual meaning. Through concept matching, synonyms, and natural language algorithms, semantic search provides more interactive search results through transforming structured and unstructured data into an intuitive and responsive database. As it discussed in subsection 3.2 various semantic search approaches have been utilized for semantic search in digital library systems. In this research, the study employed a robust hybrid semantic search approach which utilizes both the power of semantic web domain specific ontologies and distributional sentence embedding techniques. To evaluate the semantic similarity degrees of user queries and digital library resource articles of books the researcher has employed a widely used efficient pairwise similarity metrics such as; cosine similarity, Euclidean-distance and linear-kernels.

CHAPTER FOUR

4. DEVELOPING SEMANTIC WEB FOR DIGITAL LIBRARIES

In-depth review of related literature shows that the future digital libraries will offer large variety of information objects. These objects would be in form of text, tables, images, scientific data, annotations, 3D images and videos. Heterogeneous information sources would be required for handling new form of information. Future digital libraries will offer greater services to the users. According to (Castelli, 2006).The potential users of future digital library will demand more dynamic digital library models. The Semantic Web offers potential for digital library to offer web-based library services (Burke, 2009). The next decade's digital libraries have grown into architecting data models and semantic technologies to combine the existing information retrieval systems with Semantic Web frameworks and Semantic layers to make the digital collections more searchable, interoperable and discoverable (Macgregor, 2008). Semantic Web facilitates the digital library services. Librarians should use RDF and web Ontologies to develop Semantic Digital Libraries and provide access to information available all around the world by offering semantic searching (Moran, 2010) .

The next generation digital libraries are dependent on networked knowledge organization systems, Ontology models, flexible metadata standards and Semantic query languages. Different attempts are being made to reuse their data by making it accessible through Semantic layers and Semantic resource discovery system. There are different web applications that can be used to integrate the digital libraries with Semantic Web technology. Some scholars like Balaji, Sarangi and Madalli (2012) studied to find out that how the knowledge management systems of digital libraries can be integrated with semantic technology to utilize semantic potential and approaches for organizing the knowledge on the web.

4.1. Evaluating and Experimenting.

In this study various existing digital library searching systems such as traditional keyword based and dictionary-based systems have been assessed. Following this, the study come up with a robust integrated semantic web supported digital library searching systems. The integrated system incorporates domain specific ontologies and semantic sentence embedding approaches. The domain specific ontologies for communicable diseases were particularly adopted from international diseases library ontologies. The researcher has extracted rich source of dictionaries and relevant terms from entities and properties of major communicable diseases. On the other hand, the robust sentence embedding pre trained models were augmented on the top of our small size communicable diseases articles metadata. The relevant attributes of the articles were identified as metadata of titles, abstracts, authors and publisher information. Basically, the researcher aim was to demonstrate the semantic web technology for digital library specifically the domain ontologies prominently contribute to improve the digital library services. In this case the study has taken the searching capabilities categories. To do so, the newly proposed framework has been evaluated in the following section using small-scale dataset collected.

4.1.1 The role of semantic web and web ontology to enhance digital library

In the digital era, currents Digital library users are more demanding than web search engine users. In many cases digital libraries are considered as an outreach of classical libraries Use of digital information tools and processes are changing the library environment and library services. Libraries and librarians are required to become more relative with technological environment. The aims of the semantic web are Organization of knowledge in conceptual spaces according to its meaning, usage of automated tools to support maintenance by checking for inconsistencies and extracting new knowledge, replacement of keyword-based search by query answering: requested knowledge being retrieved, extracted, and presented in a human friendly way and query answering over several documents will be supported. The ultimate goal of the Semantic Web is to allow machines to understand the meaning of digital objects, rather than just the key words used to describe them. This will revolutionize the search and retrieval of digital objects, a key function for digital libraries. One of the main benefits of the Semantic Web is that it operates

within the current Web environment to add logic and meaning to digital objects. (Jiang, X., & Tan, A, 2009)

4.1.2. Semantic descriptions of resource (RDF)

Semantic Web facilitates the digital library services. Librarians should use RDF and web Ontologies to develop Semantic Digital Libraries and provide access to information available all around the world by offering semantic searching (Moran, 2010). Due to the widespread importance of integration and interoperability for intra- and inter-business processes, the research community has tackled this problem and developed semantic standards such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL). RDF and OWL standards enable the Web to be a global infrastructure for sharing both documents and data, which make searching and reusing information easier and more reliable as well. In this study, the existing web ontology (OWL) has been used as source of dictionaries in order to evaluate the semantic based digital libraries system.

4.2 Experiment and Evaluation

In this section the researcher discussed the details of experimental setting, implementation details, parameter configuration, tuning and discussion of results found. In addition, evaluation of search engine strategies in digital library systems has been clearly stated in this section.

4.2.1 Dataset preparation

In this study, the researcher has collected a dataset by crawling domain specific potential articles, books and conference papers on communicable diseases. The dataset has been collected using powerful python Beautiful Soup libraries. The domain specific dataset source which is called communicable disease has been selected because of the availability of existing extensible ontology resources in order to extract and map dictionary of relevant searching terms. The credible sources that have been used in order to scrape the dataset found at [1] [2] that contains

[1] https://www.semanticscholar.org/me/research

[2] https://scholar.google.com/

numerous relevant articles. To utilize the scraping tools, the researcher used list of keywords which are founds in[3] that helps us to hunt communicable disease related articles.

Keywords used to search	**Remarks /Description**
2019-nCoV	✓
CRE	✓
Ebola	✓
Enterovirus D68	✓
Flu	✓
Hantavirus	✓
Hepatitis A	✓
Hepatitis B	✓
HIV/AIDS	✓
Measles	✓
MRSA	✓
Pertussis	✓

[3] https://acphd.org/communicable-disease/list-of-communicable-diseases/

Rabies	✓
Sexually Transmitted Disease	✓
Shigellosis	✓
Tuberculosis	✓

Table 1 List of Key words that helps to search articles

The metadata the researcher extracted during collection of the dataset includes title, author, year, publisher, ISBN, edition, description, URL/DOI and abstracts. These attributes have been selected based on the conventional database management metadata storage trends for book information in digital libraries. Users intended to search resources using either of the relevant metadata that has been incorporated with books. A pipelines of data preprocessing steps has been employed in order to clean our dataset. Unwanted characters, stop words and irregular characters have been removed using special regular expressions. It's envisaged that the preprocessing steps improves the performance of the overall system. To do so, intensive preprocessing steps have been implemented clearly. The following table 2 contains the details of sample dataset collected for the purpose of this study.

Title	**Author**	**Year**	**Publisher**	**ISBN**	**Edition**	**URL/Doi**	**Abstract**
Pediatric Board Study Guide	Osama. L Naga	2020	SPRIGER	978-3-030-21266-7	second edition	https://doi.org/10.1007/978-3-030-21267-4	Indexed as d1
Handbook of	Bonny Libbey	1998	**Jones and BartlettPub**	0-7637-0624-8	3rd editi	DOI:10.1056/NEJMOA020047	Indexed as

oncology nursing	Johnson, Jody Gross.		**lishers**		on		d2
Ebola	H. Feldmann, ÿA. Sprecher,ÿ T. Geisbert	ÿ2020	Medicine	--	--	DOI:10.1056/nejmra1901594	Indexed as d22
Progress and challenges in TB vaccine development	G. Voss, D. Casimiro	2018	Medicine	9996111636	--	https://doi.org/10.12688/f1000research.13588.1	Indexed as d41
Hepatitis A	Khrystyna Hrynkevych,ÿH. Schmitt	2021	VacciTUTOR	---	--	DOI:10.33442/vt202136	Indexed as d123

Table 2 Sample dataset collected to evaluate digital library searching strategies

As indicated above**, Table 2** describes the metadata of domain specific communicable disease articles collected from the web in order to evaluate the semantic based digital library systems. Over 2000 digitalized manuscripts have been collected manually in order to construct the dataset. Each article has 8 relevant attributes that the researcher needs to embed latter along with the dictionaries extracted from the domain ontologies. Thus, the table shows the sample dataset constructed to evaluate the proposed model.

4.1.1. Ontologies of communicable diseases

In this study, an existing ontology of infectious diseases have been employed that are publicly available in[4]. The ontologies contain major entities, properties and relationships of concepts. The ultimate goals of reusing existing ontologies mainly focused on improving the qualities of term dictionaries the researcher used in combination with the powerful word embedding approach. The ontologies incorporate entities that are relevant to both biomedical and clinical aspects of most infectious diseases. The researcher has created a rich dictionary of terms which were comprised of different pathogens, vectors and related infectious disease terms. For instance, tuberculosis is one of the major communicable diseases in which there are multiple related terms that have incorporated to develop tuberculosis ontology. Where TB by itself is conceptually defined as a primary bacterial infectious disease that is located in lungs, located in lymph nodes, located in pericardium, located in brain, located in pleura or located in gastrointestinal tract, has materia_basis_in Mycobacterium tuberculosis, which is transmitted by droplets released into the air when an infected person coughs or sneezes.

The following figure 3 and figure 4 shows a sample existing ontology for tuberculosis bacterial infectious disease. Where the Error! Reference source not found. **3** shows the tree-based views of concepts and relationships and **figure 4** depicts the graphical visualization of the similar ontologies which infers the is_a, has_a and transmited_by relationship between categories of tuberculosis infectious disease.

[4] https://bioportal.bioontology.org/ontologies/IDO

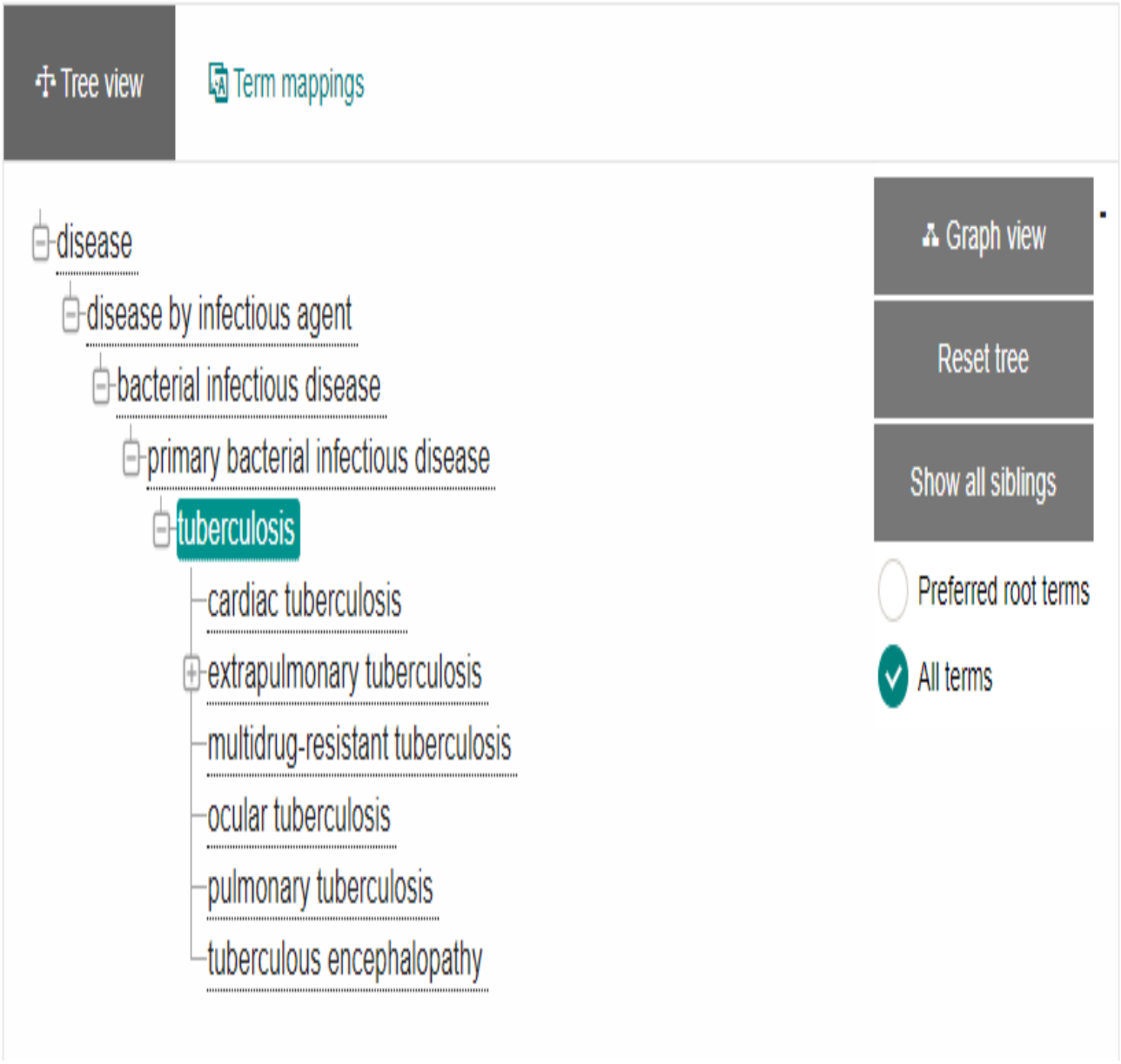


Figure 3 Tree based views of one of the infectious disease ontologies

In the above **figure 3** the tree-based representation of sample communicable disease such as Tuberculosis ontologies has been clearly stated. As it have stated in subsection 0 our data source is limited to domain specific communicable diseases. Thus, tuberculosis is one major categories of communicable diseases which has been widely represented in rich form of web ontologies by domain experts in order to extract concepts. Thus, in the above figure we have tried to illustrate the major categories of tree-based ontologies that infer categories, relevant concepts and

attributes of tuberculosis. Those rich conceptual terms have been used with the word vectors of articles in order to search semantically.

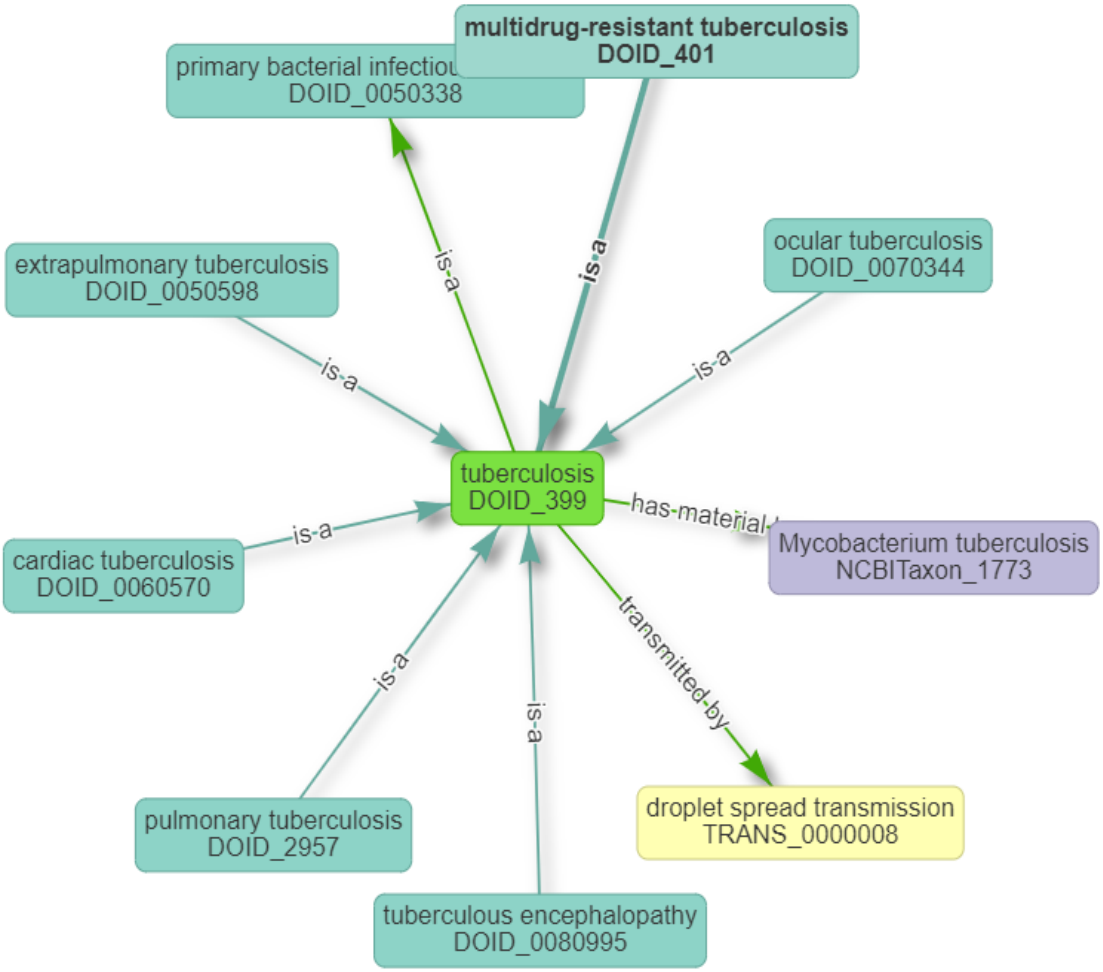


Figure 4 sample ontologies for tuberculosis infectious disease

The above figure 4 the graphical representation of sample communicable disease such as Tuberculosis ontologies has been clearly stated. The ontology describes major entities such as the type's relationship or variants of Tuberculosis and the major proprieties it possesses as a has_a relationship between the main entities. Basically, we are interested on the potential rich

dictionaries of each communicable disease that has been represented interims of the has_a and is_a relationships.

4.1.2. Experimental setup

To conduct the experiment of this study the researcher has setup an experimental environment on googles colab to utilize high performance computing machines (GPU) and freely available limited storage capabilities of Google drives. Besides, in order to construct vectors of documents related to infectious disease vocabularies we could use large vocabularies of existing pre trained models. The researcher used the default settings of google colab in order to run our experiment.

In this study, a semantic search technique has been employed in order to evaluate the effectiveness of proposed method semantic web supported digital library semantic search systems. To do so, the researcher has used a semantic robust sentence embedding techniques along with rich set of dictionaries of infectious disease terms that has been already extracted from domain specific ontologies which stated in **subsection** Error! Reference source not found. of this document. Initially, once the study preprocess the domain-specific articles dataset along with its metadata a semantic vector of titles and other related information has constructed for the corresponding articles or books. Robust sentence embedding pre trained models such as InferSent has been tuned to incorporate new vocabularies of infectious disease. These set articles or books are basically indexed resources of digital library systems. Following this, the semantic similarity of the indexed digital library resources instance vectors that has been trained on the top of pre trained models and infectious disease related dictionary of terms computed using widely used efficient pairwise similarity metrics such as; cosine-similarity and Euclidean-distance. Cosine similarity calculates the angle between two vectors based on their semantic relatedness.

4.1.3. Experimental Results

In this study, a standard evaluation metrics such as sklearn pairwise similarity metrics such as Cosine similarity or Cose kernel approach has been used to evaluate the best and worst semantic similarity based digital library search system.

Table 4.31 below shows the experimental results of semantic search supported digital library system using dictionary of rich terms of infectious diseases.

S.NO.	Techniques	Standard metrics	Top search terms used
		Cosine similarities	
1	DOC2VEC	0.755	Tuberculosis
		0.799	Influenza
		0.813	Hepatitis
		0.811	Mumps
		0.796	Rabies
		0.901	Viral Hemorrhagic
		0.825	Yellow Fever
		0.789	Fever Animal
		0. 692	Zika
2	InferSent	0.768	Tuberculosis
		0.898	Influenza
		0.767	Hepatitis

		0.826	Mumps
		0.866	Rabies
		0.963	Viral Hemorrhagic
		0.873	Fever Animal
		0.812	West Nile Virus
		0. 695	Yellow Fever
		0.914	Zika
3	TF-IDF	0.235	Tuberculosis
		0.456	Influenza
		0.567	Hepatitis
		0.876	Mumps
		0.645	Rabies
		0.723	Viral Hemorrhagic
		0.885	Fever Animal
		0.745	West Nile Virus
		0.566	Yellow Fever
		0.888	Zika

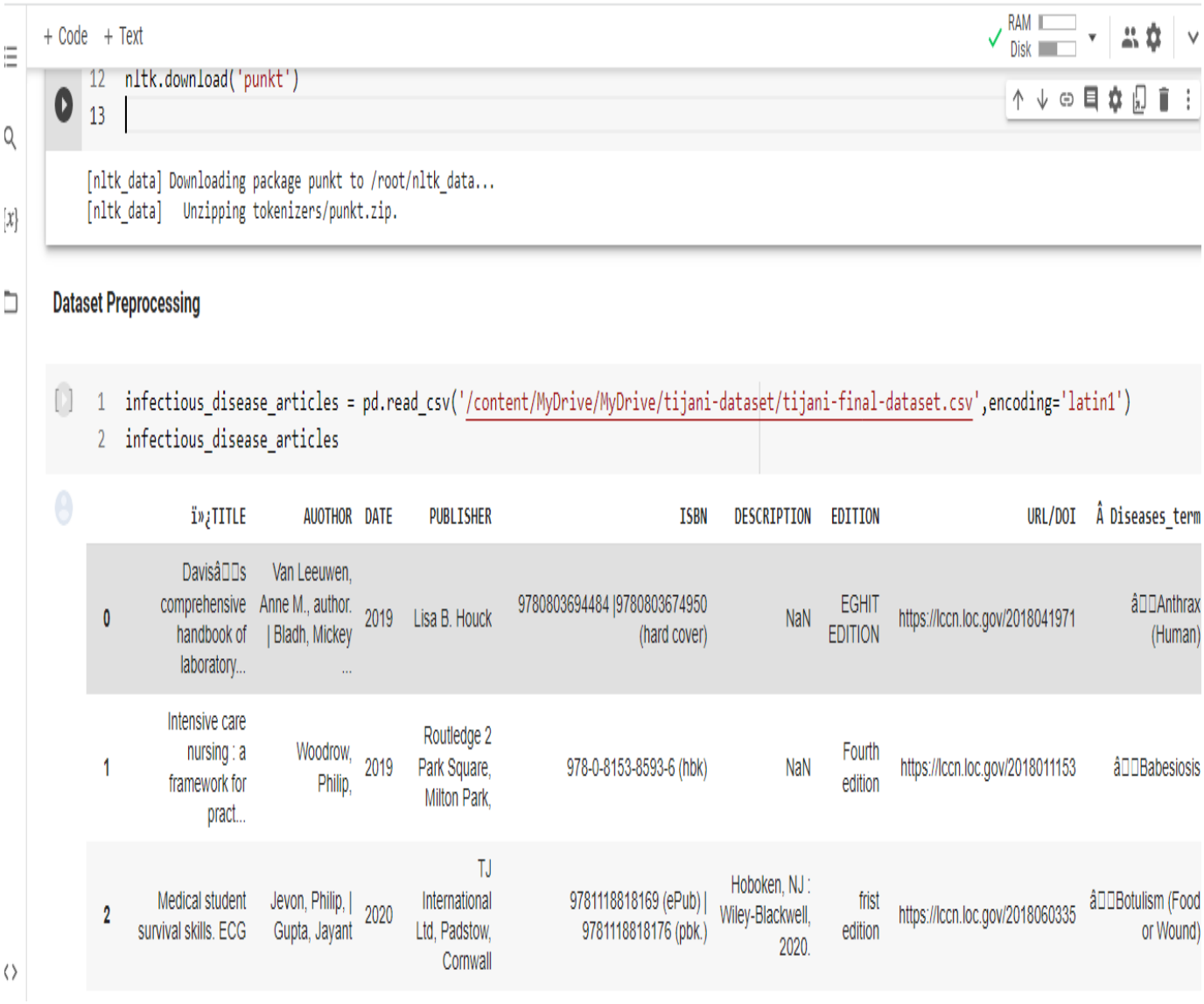
Table 3 Experimental results for semantic searching in digital library

Tables 3 indicate that the comparison of infer Sent and DOC2Vec in which an InferSent pre trained model and DOC2Vec have shown a better result than the classical TF-IDF based semantic search strategies.as it indicated from the above table the TF-IDF techniques has showed the least scores in comparison to other metrics.

4. 2. Prototype development

This chapter presents a simple demo of the generic framework proposed in Section 3.5 of this document and the dataset prepared in subsection 4.1.1**.** In initial development of the prototype, the semantic based digital library searching models has been trained on three robust sentence embedding models along with the dictionaries extracted from domain ontologies.

To do so, the prototype is developed using cloud based Google Collaborators that supports GPU enabled robust sentence embedding network architectures with Keras in the back-end. Python programming language and dependent text processing libraries such as pandas and Numpy has been utilized. The google drive has been also used as a temporal storage for training and testing dataset and also for trained models pickle file location. The prototype has been implemented to demonstrate the proposed architecture partially at initial stage. Thus, the demo the researcher has implemented shows the capability of digital library searching system using semantic text similarity based robust sentence embedding and domain ontologies.



	ï»¿TITLE	AUOTHOR	DATE	PUBLISHER	ISBN	DESCRIPTION	EDITION	URL/DOI	Â Diseases_term
0	Davisâ□□s comprehensive handbook of laboratory...	Van Leeuwen, Anne M., author. | Bladh, Mickey ...	2019	Lisa B. Houck	9780803694484 |9780803674950 (hard cover)	NaN	EGHIT EDITION	https://lccn.loc.gov/2018041971	â□□Anthrax (Human)
1	Intensive care nursing : a framework for pract...	Woodrow, Philip,	2019	Routledge 2 Park Square, Milton Park,	978-0-8153-8593-6 (hbk)	NaN	Fourth edition	https://lccn.loc.gov/2018011153	â□□Babesiosis
2	Medical student survival skills. ECG	Jevon, Philip, | Gupta, Jayant	2020	TJ International Ltd, Padstow, Cornwall	9781118818169 (ePub) | 9781118818176 (pbk.)	Hoboken, NJ : Wiley-Blackwell, 2020.	frist edition	https://lccn.loc.gov/2018060335	â□□Botulism (Food or Wound)

Figure 5 Dataset loading and visualization demo

As the above figure indicates the dataset collected from different source is loading and visualization process on python the google drive has been also used as a temporal storage for training and testing dataset and also for trained models pickle file location. So figure 5 used to shows the demonstration of the proposed architecture partially at initial stage.

Figure 6 Dataset preprocessing demo

As it indicated on above figure 6, Preprocessing is the most significant and preliminary task of engineering approaches in many webs search engine and other text data manipulation applications which helps to clean and made ready the corpus for the further processing. The main role of data preprocessing is formatting or normalizing the input documents, so that later tasks can be computed easily. Therefore, figure 6 depicts that the preprocessing of dataset collected.

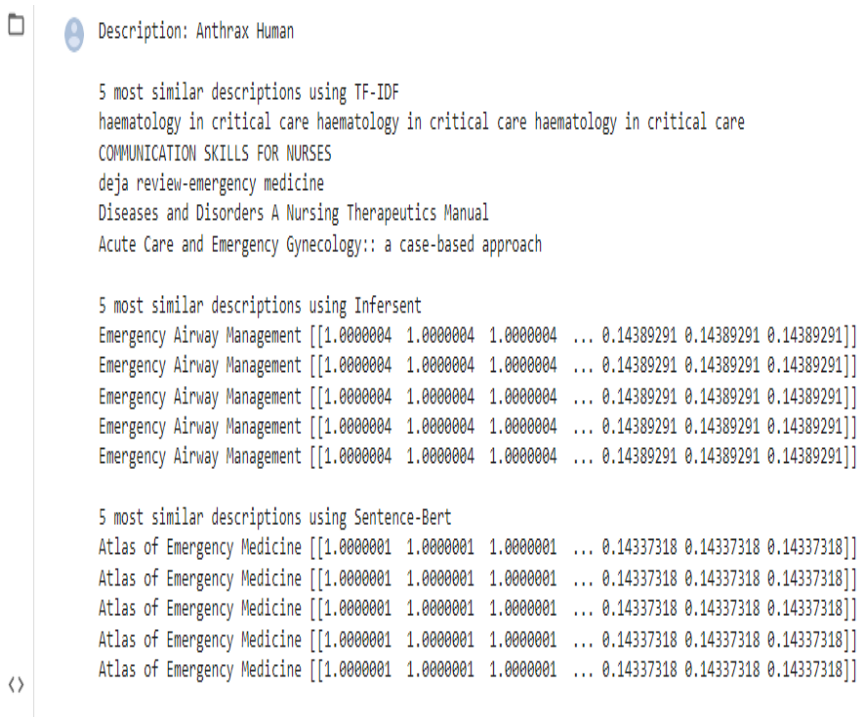


Figure 7 Interactive python Console based evaluation demo

Figure 7 indicate that the interacve python console for evaluating the developed prototype, as it depicted on the above figure The prototype implemented shows the capability of digital library searching system using semantic text similarity based robust sentence embedding and domain ontologies on python for console.

4.2 Discussion of the Findings

In this study as shown in table 3 an InferSent pre trained model and DOC2Vec have shown a better result than the classical TF-IDF based semantic search strategies. To inspect the experimental results of both techniques the researcher pre-set the maximum number of top similar documents which are very close to the search vocabularies. The DOC2Vec and InferSent have large set of pre trained vocabularies that helps to augment limited size dataset. In addition, semantic relatedness of search terms and digital resources such as articles has been calculated based on the maximal neighboring words relatedness in specific window size. However, in case of the classical TF-IDF based similarity measures the similarities of search terms and documents has been calculated based which terms are topically relevant (or irrelevant) by analyzing how often a term appears on a document (term frequency — TF) and how often it's expected to

appear on an average page, based on a larger set of documents (inverse document frequency — IDF).

From our experimental evaluation the TF-IDF techniques has showed the least scores in comparison to other metrics. On the other hand, the rich set of vocabularies for infectious disease was very helpful to scan a match in the digital library contents. Moreover, the set of vocabularies were used to search articles and books from the beginning. Indeed, in this study the roust sentence embedding techniques along with set of terms that has been extracted from domain specific ontologies has showed an improved accuracy to search contents or articles in digital library system. Therefore, the semantic web element which is a domain ontology has a powerful contribution to boost the searching capabilities of digital libraries. Besides, the ontology by itself recently couldn't assured to retrieve documents with optimal accuracy this is actually because of the limitation of domain ontologies to incorporate and encode all concepts in particular domains. To do so, in this study the researcher observed that the hybrid effects of both the domain ontologies and the robust sentence embedding approaches were efficient and has contributed to score an optimal searching result in a dynamic keyword and phrase based semantic searching systems.

In Connections with similar activities carried out by other research groups and initiatives at an international level different studies are established in order to achieve a global and stable level of consensus on the model, but is difficult establish a consensus because they concern several different areas, which makes it difficult to compare or combine the results achieved in these areas since it's not always clear how they are related, and how they can impact on or constrain one another.

Nguyen (2013) in her research on Digital Library Research Trends she said that DL have been discussed in various international digital library conferences, i.e. Joint Conferences on Digital Libraries (JCDL), The European Conference on Research and Advanced Technology for Digital Libraries (ECDL), International Conference on Asia-Pacific Digital Libraries (ICADL) etc, also there is so much research in different areas (digital library architecture, systems, tools, and technologies; digital content and collections; metadata; standards; interoperability; knowledge organization systems; users and usability; legal, organizational, economic, and social issues) presented an overview of trends in digital library research, but to date, to the best

of the researcher's knowledge, there has not been any study that conducted to enhance digital library search using existing ontology in the digital library field.

For the accomplishment of the study, the researcher decided to select the domain specific dataset source which is called communicable disease because of the availability of existing extensible ontology resources in order to extract and map dictionary of relevant searching terms. The credible sources that have been used in order to scrape the dataset found at [5] [6] that contain numerous relevant articles. To utilize the scraping tools, the researcher used list of keywords which are founds in[7] that helps to hunt communicable disease related articles.

[5] https://www.semanticscholar.org/me/research

[6] https://scholar.google.com/

[7] https://acphd.org/communicable-disease/list-of-communicable-diseases/

CHAPTER FIVE

5.1. CONCLUSION AND RECOMMENDATION

In this chapter the study presents reflection on the overall progress of the study starting from research question, dataset preparation, methods used, experiments and results found.

5.1.1. Conclusion

The broad concept of digital libraries ideally represents the needs of heterogeneous information resources combining the development of complex systems issues such as interoperability among existing data providers, distributed retrieval, and long-term preservation. These days most of the existing web-based digital libraries offer search as well as navigation services. Search services are mainly based on a set of metadata such as domain related keywords, author, title, or journal name that carries no semantic information for the search engine. Besides, digital libraries may support graphical user interfaces to aid the formation of queries in order to search documents. Queries can contain one or more domain keywords to be searched in the full text document, or in one of the metadata fields. The Semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. In addition, the semantic web has played a significant role to improve the efficiency of different search engines including digital library systems. Study concludes that future digital libraries will offer more effective and dynamic information services. The use of Semantic applications increased to structure the metadata of digital contents in well-organized and meaningful way to present semantic retrieval of information. Different studies indicate the future digital libraries greatly rely on web intelligent agents, context-based technology and artificial intelligent applications. Due to these applications interoperability among various heterogeneous information systems will be increased and digital contents that are available on the WWW will become more ubiquitous. Librarians' role in future digital library shall be more technology based and web-oriented. Librarians must develop their digital skills to utilize modern applications and Semantic tools for developing digital libraries. They should increase their knowledge of modern metadata schemas and advanced web languages to understand and harness the Semantic applications for digital libraries.

In this study the researcher have evaluated the effect of semantic web technology and web ontology language for digital library service. The study analyzed various digital library services with the support of semantic web technologies. Mainly, digital libraries have been benefited from ontology language, semantic web and syntactical features in order to enhance the searching capabilities of their systems. To do so, in this study the researcher have discussed all the digital libraries searching strategies and come up with a new integrated robust framework with the help of domain specific ontologies and semantic sentence embedding approach. The study has used an existing ontology for infectious disease portal in order to extract dictionaries of terms, concept relationships and properties. On the other hand, in the study the capabilities of sentence embedding techniques on the top of rich set of pre trained vocabularies pre trained models have been used to evaluate the experiment. In order to evaluate the proposed framework the researcher has prepared a small size dataset on domain specific communicable diseases. The dataset consists of relevant metadata of articles such as author, title, URL, abstract, ISBN, year of publication, and disease term or vocabularies.

The proposed framework has been evaluated using the dataset that has been prepared for the purpose of this study. Necessary, preprocessing techniques has been employed in order to clean, normalize and transform our dataset. the researcher configured and tuned hyper parameters of robust sentence embedding pre trained models in order to augment our small size dataset on the top of vocabulary rich trained vectors. The major sentence embedding techniques that have been evaluated along with the domain ontologies dictionaries are DOC2Vec, InferSent and the classical TF-IDF methods. To evaluate the proposed method a pairwise distance-based similarity metrics such as cosine similarity has been employed. The experimental result has shown that the robust InferSent sentence embedding technique returns clear significant result to find the semantic relatedness between dictionaries of domain specific infectious disease ontologies and articles. Indeed, the semantic web technologies such as domain ontologies along with robust sentence embedding techniques could enhance the searching capabilities of digital libraries system. Integration of Semantic Web applications with digital libraries shall present effective retrieval of digital contents. Use of these applications in digital libraries shall be helpful in making the digital information easily accessible and retrievable. It enhance the interoperability among different digital information systems which would play its dynamic role in finding the required information on the web.

5.1.2. Recommendation

Based on reviewed studies and conclusions drawn from this research, the followings recommendations are given to cope with the challenges of semantic digital library In the future interested researchers could extend our work by considering various techniques such as advanced semantically language models other than those embedding based resources. The current dynamic information technologies demand proactive role by library professionals.

- Librarians should embrace the emerging technologies in libraries and information environment and increase their digital skills for developing Semantic Digital Libraries.
- They should develop their skills and expertise to use different web languages i.e. XML, Web Service Modeling Language, RDF schemas and different ontology based applications for digital libraries.
- In addition; it is recommended also increasing the size of the dataset and evaluating the proposed framework on other domains by incorporating hybrid dictionaries and digital library resources in order to validate the robustness of the proposed framework. Moreover, it is very important to extend this study by considering the combination of three approaches such as knowledge graphs, domain ontologies and syntactical linguistic features of digital library resources.
- The research recommends that LIS Schools must incorporate the use of Semantic Web languages in their curriculum and train their graduates in developing web-based information services. Future digital library development software shall provide more opportunities to use Semantic applications in developing digital libraries.

Reference

Abdulelah et al. (2017). WEB EVOLUTION - THE SHIFT FROM INFORMATION. *International Journal of Artificial Intelligence and*, 8.1-3.

Berners-Lee et al. (2001). The Semantic Web. *Scientific American* .

Burke, M. (2009). The semantic web and the digital library. *Aslib Proceeding New Information Perspective*, *61*, pp. 316-322.

C. M. Owusu-Ansah. (2019). Conceptions of digital libraries: an African perspective.

Candela et al. (2011). the Digital Library Reference Model.

Castelli, D. (2006). Digital libraries of the future – and the role of libraries. *24*, 596-503.

Castro, L.J., Giraldo, O.L., & Castro, A.G. (2010). Using the Annotation Ontology in. *SEMWEB*.

Chawner, B. (2008). Spectators, not players: information managers‟ use of Web 2.0. *The Electronic Library, 26*, 630-649.

Chowdhury, G.G., & Chowdhury. (2002). Introduction to Digital Libraries.

G. T. ,. T. A. ,. A. Norbert Fuhr. (2007). *Evaluation of digital libraries.* Springer-Verlag.

Gruber, T. (1995). Toward Principles for the Design of Ontologies Used for knowledge sharing. *Journal of Human-Computer Studies, 43*, 907-928.

Hendrix. (2010). Checking out the future: perspectives from the librarycommunity on information technology and 21st-century libraries.

Ján Hreňo et al. (2010). Integration of Government Services using Semantic Technologies. *Journal of Theoretical and Applied Electronic Commerce Research*, 23-31.

Jiang, X., & Tan, A. (2009). Learning and inferencing in user ontology for the personalized in Semantic Web search. *information science*, 16.

Jorge C, et al. (2014, may 16). THE SEMANTIC WEB AND ITS APPLICATIONS. *Journal of Intelligent Information Systems (JIIS)*.

Jorge Cardoso et al. (2006). *THE SEMANTIC WEB AND ITS APPLICATIONS.* lab.

Kamlesh et al. (2016). Semantic Digital Library. *International Conference on Advanced computing*, (pp. 2-6).

Kruk., S. R. (2014). Advanced search and browsing in digital libraries. *Digital Enterprise research institute*.

L. Guo, Y. Zhou. (2010). Ontology Based Digital library Search Model Research College of Information Science and Technology,.

Macgregor, G. (2008). Introduction to a special issue on digital libraries and semantic web. *context, applications and research, 57*, 157-177.

Moran, C. (2010). The use of semantic web technologies in digital libraries.

Owusu-Ansah, C. M. (2019). Conceptions of digital libraries ;an African perspective in Digital library perspectives.

Patel-Schneider et al. (2002). The Yin/Yang web: XML syntax and RDF semantics. 11th. *11th international conference on World Wide Web.* USA.

Peffers, K et al. (2008). A Design Science Research Methodology for Information Systems. *Journal of Management Information Systems*.

S. A. Khan. (2015). *Use of Semantic Web & Ontology-based Applications for Future Semantic.* University of the Punjab.

Sebastian R et al. (2014). The Role of Ontologies in Semantic Digital Libraries.

Senthil et al. (2006). Semantic Web: Emerging Technology for Digital Library. *Defence Scientific Information & Documentation Centre (DESIDOC)*.

Sreeja. (2012). agent based semantic web. *journal of computer science*, 1-23.

Appendix

Appendix 1: Sample Data collection instruments

Questionnaires for Library and information science professional (LISP)

Dear Respondent:

This is a questionnaire developed to extract data for analysis digital library service for its enhancement through semantic web technology and ontology by researcher ***The Case of Jimma university library and information services, electronics and digital library services***. The purpose of the study to explore the semantic web feature and design semantic based digital library framework, to improve digital library services and knowledge management.

Instruction. **Tick** "√" the **Boxes** □ Attached to the **Item(s)** that Correspond Your **Answers** (e.g. like this☑) and Write **Brief Response**(s) on the **Spaces** Provided where Necessary!

Part I: **general information**

Questionnaires for Library and information science professional (LISP)

Dear Respondent:

This is a questionnaire developed to extract data for analysis digital library service for its enhancement through semantic web technology and ontology by researchers ***The Case of Jimma university library and information services, electronics and digital library services***. The purpose of the study to explore the semantic web feature and design semantic based digital library framework, to improve digital library services and knowledge management.

Instruction. **Tick** "√" the **Boxes** □ Attached to the **Item(s)** that Correspond Your **Answers** (e.g. like this☑) and Write **Brief Response**(s) on the **Spaces** Provided where Necessary!

Part I: **general information**

1. **Sex:**

   Male □ Female □

2. **Age:**

   [18-35] □ [36-50] □ [≥51] □

3 Marital statuses

□Single □ married □divorced

3. **Educational** Qualification:

□ Diploma/certificate □BSC □ MSC □ PHD

4. Professional qualification /field of study

□ Information science □ library and information science
□ Information technology □ computer sceince □software engineering

□Others

Part II: **Questions Relating to Study Objectives**

1. Is there functional digital library in your organization?

☐ Yes ☐ No

2. If the above question is yes, what type of digital services and what type of collections are there in the Digital libraries?

☐ Theses/Dissertations.

☐ Institute Publications/Articles

☐ Multimedia Lectures

☐ E Books

☐ Others**

3. What are the features of your digital libraries and types of the Software used in digital library development?

☐ Greenstone ☐ Fedora

☐ Dspace ☐ JeromeDL ☐ Eprints ☐ Others**

4. Are you aware of the semantic features of digital library software?

☐ Yes ☐ No

5. If your response is **yes** for the above question, what are the semantic web solutions is available in your digital library system based on semantic web technology?

☐ Simple Search. ☐ Advanced Search. ☐ Natural Language Search.

☐ Browsing. ☐ Collaborative Browsing. ☐ Semantic Tagging

☐ Bookmarking Resources. ☐ Bookmark Sharing

☐ Blogging Resources. ☐ Ranking Resources. ☐ Bookmark Recommendation.

☐ Resource Recommendation. ☐ Taxonomy View. Query Building Mechanism

6. What are the Search (Technique) Features used in your digital library system? (two or more option possible)

☐ Proximity Searching

☐ Truncation

☐ Boolean

☐ Defined Dictionary

☐ Clustering

7. What are Browsing (Technique) Features are available to assist your digital library user to navigate among correlated searchable terms?

☐ By Author/Title/Year. ☐ Alphabetically. ☐ By Subject.

☐ Browse with Exhibit. ☐ Tag Filtering. ☐ Taxonomy View.

☐ Other

8. Due to the information explosion digital libraries are facing challenges in managing, organizing and retrieving information from the digital resources.

☐ Strongly Agree ☐ agree ☐ disagree ☐ strongly disagree

9. Digital library needs active information access facilities and to acquire trusted and reliable information.

☐ Strongly Agree ☐ agree ☐ disagree ☐ strongly disagree

Appendix 2: Sample Dataset collected

Title	Author	Year	Publisher	ISBN	Edition	URL/Doi	Abstract
Pediatric Board Study Guide	Osama. L Naga	2020	SPRIGER	978-3-030-21266-7	second edition	https://doi.org/10.1007/978-3-030-21267-4	Indexed as d1

Handbook of oncology nursing	Bonny Libbey Johnson, Jody Gross.	1998	**Jones and BartlettPublishers**	0-7637-0624-8	3rd edition	DOI:10.1056/NEJMOA020047	Indexed as d2
Ebola	H. Feldmann,ÿA. Sprecher,ÿT. Geisbert	ÿ2020	Medicine	--	--	DOI:10.1056/nejmra1901594	Indexed as d22
Progress and challenges in TB vaccine development	G. Voss, D. Casimiro	2018	Medicine	9996111636	--	https://doi.org/10.12688/f1000research.13588.1	Indexed as d41
Hepatitis A	Khrystyna Hrynkevych, ÿH. Schmitt	2021	VacciTUTOR	---	--	DOI:10.33442/vt202136	Indexed as d123

Appendix 3: sample code

```
# -*- coding: utf-8 -*-
"""TijaniThesisSearchingInDigitalLibraries.ipynb

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1XTKN7zmUlFDDCe5eJdOypvEm-QhHJxli
"""
```

```
from google.colab import drive

drive.mount('/content/MyDrive')

!pip install -U sentence-transformers

"""**Importing libraries**"""

import nltk
import torch
import numpy as np
import pandas as pd
from nltk import word_tokenize
from nltk.stem import WordNetLemmatizer
from models import InferSent
from gensim.models.doc2vec import Doc2Vec
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sentence_transformers import SentenceTransformer
nltk.download('punkt')

"""**Dataset Preprocessing**"""

infectious_disease_articles = pd.read_csv('/content/MyDrive/MyDrive/tijani-dataset/tijani-final-dataset.csv',encoding='latin1')
infectious_disease_articles

infectious_disease_articles.columns
```

```
infectious_disease_articles.columns =
infectious_disease_articles.columns.str.replace('[#,@,&,ï»¿,Â ]', '')

infectious_disease_articles.columns

infectious_disease_articles.drop(infectious_disease_articles.columns[infectious_disease_articles.
columns.str.contains('unnamed',case = False)],axis = 1, inplace = True)

infectious_disease_articles


#infectious_disease_articles.to_json (r'/content/MyDrive/MyDrive/tijani-dataset/tijani-final-
dataset.csv')


infectious_disease_articles.TITLE= infectious_disease_articles.TITLE.fillna('',inplace=False)

#tweets.tweets=tweets.tweets.fillna('',inplace=False)


from nltk import word_tokenize

articles_titles = infectious_disease_articles["TITLE"]

print(articles_titles)


import ast


articles_titles= articles_titles.str.encode('ascii', 'ignore').str.decode("utf8")

articles_titles


"""The first step is to create numerical representations of each description the news.


## TF-IDF

The first approach uses TF-IDF to create description representation. Each description is lower
cased and lemmatized with WordNet, before create the TF-IDF matrix.

"""


lemmatizer = WordNetLemmatizer()

tfidf_vectorizer = TfidfVectorizer()
```

```
nltk.download('wordnet')


descriptions_lemmatized = [" ".join([lemmatizer.lemmatize(token.lower()) for token in
word_tokenize(description)]) for description in articles_titles.values ]


from wordcloud import WordCloud

from nltk import word_tokenize, sent_tokenize, FreqDist

fdist = FreqDist(descriptions_lemmatized)


import matplotlib.pyplot as plt


wordcloud = WordCloud(background_color='white').generate(' '.join(fdist))

plt.figure(figsize=(12, 12))

plt.imshow(wordcloud, interpolation="bilinear")

plt.axis("off")

plt.show()


descriptions_representation_tfidf = tfidf_vectorizer.fit_transform(descriptions_lemmatized)


"""## Doc2Vec

The second approach to create description representations uses the Doc2Vec library,
implemented in GenSim. The pre-trained model was downloaded from
https://github.com/jhlau/doc2vec. Make sure you download one of the pre-trained Doc2Vec
models and save it in the models folder.

"""


file = "/content/MyDrive/MyDrive/doc2vec.bin"

file


doc2vec_model = Doc2Vec.load(file)
```

```
!pip install 'gensim==3.2.0.'


import os
os.chdir("/content/MyDrive/")
!ls


start_alpha = 0.01
infer_epoch = 1000


documents = [[token for token in nltk.word_tokenize(description.lower())] for description in articles_titles]
embeddings_doc2vec = []
for document in documents:

embeddings_doc2vec.append(doc2vec_model.infer_vector(doc_words=document,alpha=start_alpha, steps=infer_epoch))


emb


"""## InferSent
The third approach creates description embeddings using InferSent. The pre-trained model was downloaded from https://github.com/facebookresearch/InferSent. You can download the word embeddings from https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M.vec.zip and the infersent model from https://dl.fbaipublicfiles.com/infersent/infersent2.pkl. Both should be saved in the models folder.
"""


V = 2
MODEL_PATH = '/content/MyDrive/MyDrive/infersent%s.pkl' % V
params_model = {'bsize': 64, 'word_emb_dim': 300, 'enc_lstm_dim': 2048,
                'pool_type': 'max', 'dpout_model': 0.0, 'version': V}
```

```
infersent = InferSent(params_model)

infersent.load_state_dict(torch.load(MODEL_PATH))

W2V_PATH = '/content/MyDrive/MyDrive/crawl-300d-2M.vec'

infersent.set_w2v_path(W2V_PATH)


infersent.build_vocab(articles_titles.values, tokenize=True)


embeddings_infersent = infersent.encode(articles_titles.values, tokenize=True)


embeddings_infersent


import pickle


# Save in line of code

pickle.dump(embeddings_infersent ,
open("/content/MyDrive/MyDrive/article_titles_embed.pickle", "wb"),
protocol=pickle.HIGHEST_PROTOCOL)


loaded_infersent_trained_vector =
pickle.load(open("/content/MyDrive/MyDrive/article_titles_embed.pickle", 'rb'))

loaded_infersent_trained_vector


loaded_doc2vec_trained_vector =
pickle.load(open("/content/MyDrive/MyDrive/conflict_tweet_embed_doc2vec.pickle", 'rb'))

loaded_doc2vec_trained_vector


"""## Sentence-Bert

The last approach creates description embeddings using Sentence Transformers.

"""


model = SentenceTransformer('distilbert-base-nli-stsb-mean-tokens')
```

```
embeddings_distilbert = model.encode(articles_titles.values)

embeddings_distilbert


pickle.dump(embeddings_distilbert  ,
open("/content/MyDrive/MyDrive/article_titles_embed_sentTransformer.pickle", "wb"),
protocol=pickle.HIGHEST_PROTOCOL)


loaded_distilbert_trained_vector=pickle.load(open("/content/MyDrive/MyDrive/article_titles_embed_sentTransformer.pickle", 'rb'))

#loaded_distilbert_trained_vector


from sentence_transformers import SentenceTransformer, util


from sentence_transformers import SentenceTransformer, util


model = SentenceTransformer('all-MiniLM-L6-v2')


# Single list of sentences

sentences = ['The cat sits outside',

        'A man is playing guitar',

        'I love pasta',

        'The new movie is awesome',

        'The cat plays in the garden',

        'A woman watches TV',

        'The new movie is so great',

        'Do you like pizza?']


#Compute embeddings

embeddings = model.encode(sentences, convert_to_tensor=True)
```

```
#Compute cosine-similarities for each sentence with each other sentence
cosine_scores = util.pytorch_cos_sim(embeddings, embeddings)

#Find the pairs with the highest cosine similarity scores
pairs = []
for i in range(len(cosine_scores)-1):
    for j in range(i+1, len(cosine_scores)):
        pairs.append({'index': [i, j], 'score': cosine_scores[i][j]})

#Sort scores in decreasing order
pairs = sorted(pairs, key=lambda x: x['score'], reverse=True)

for pair in pairs[0:10]:
    i, j = pair['index']
    print("{} \t\t {} \t\t Score: {:.4f}".format(sentences[i], sentences[j], pair['score']))
```

"""Now that the description representations are calculated, it is needed to calculate the representation of the input query and its similarity with all other descriptions. For that, the function below uses the cosine sinimlarity between vectors and returns the K indexes with the most similar representations."""

```
def find_similar(vector_representation, all_representations, k=1):
    similarity_matrix =cosine_similarity(vector_representation, all_representations)

    np.fill_diagonal(similarity_matrix, 0)
    similarities = similarity_matrix[0]
    if k == 1:
        return [np.argmax(similarities)]
    elif k is not None:
```

```
    return np.flip(similarities.argsort()[-k:][::1])


from sklearn.metrics.pairwise import linear_kernel


infectious_disease_articles.columns=infectious_disease_articles.columns.str.lstrip()

infectious_disease_articles.columns

infectious_disease_articles.Diseases_term =
infectious_disease_articles.Diseases_term.fillna('',inplace=False)

infectious_disease_articles.Diseases_term=infectious_disease_articles.Diseases_term.str.encode(
'ascii', 'ignore').str.decode("utf8")

infectious_disease_articles.Diseases_term.to_list()


descriptionss = ['Anthrax Human',

 'Babesiosis',

 'Botulism Food or Wound',

 'Brucellosis',

 'Chikungunya',

 'Cholera and Other Vibrio Illness',

 'Coccidioidomycosis',

 'Creutzfeldt-Jakob Disease CJD and Transmissible Spongiform Encephalopathy TSE',

 'Cyclosporiasis',

 'Cysticercosis',

 'Dengue and Severe Dengue',

 'Diphtheria',

 'Ehrlichiosis/Anaplasmosis',

 "Guillian-Barre' Syndrome",

 'Gonorrhea',

'Haemophilus influenzae, Invasive Disease',

 'Hantavirus',

 'Hepatitis A',
```

'Hepatitis B, Perinatal',

'Hepatitis B, Acute',

'Or',

'Hepatitis C, Acute',

'Hepatitis C, Perinatal',

'Hepatitis E',

'Influenza, Severe',

'Influenza, Novel',

'Legionellosis',

''Leprosy Hansen's Disease'',

'Leptospirosis',

'Listeriosis',

'Lyme Disease',

'Malaria',

'Measles Rubeola',

'Meningococcal Infections',

'Mumps',

'Pertussis Whooping Cough',

'Plague Human',

'Poliovirus',

'Psittacosis',

'Q Fever',

'Rabies, Human',

'Rabies, Animal',

'Relapsing Fever',

'Reye Syndrome',

'Rocky Mountain Spotted Fever',

'Respiratory Syncytial Virus RSV',

'Rubella, German Measles',

'Congenital Syndrome',

'Salmonellosis',

'Shiga Toxin-Producing Escherichia coli STEC and/or Hemolytic Uremic Syndrome HUS',

'Shigellosis',

'Spotted Fever Rickettsioses Including Rocky Mountain Spotted Fever',

'Staphylococcus Aureus, Severe Community-Associated',

'Syphilis',

'Taeniasis',

'Tetanus',

'Toxic Shock Syndrome Non-Streptococcal',

'Trichinosis',

'Tuberculosis',

'Tularemia',

'Typhoid Carrier',

'Typhoid and Paratyphoid Fever',

'Typhus Case Report',

'Varicella Chickenpox',

'Vibrio Illness Non-Cholera',

'Viral Hemorrhagic Fever',

'Viral Hemorrhagic Fever Animal',

'West Nile Virus',

'Yellow Fever',

'Zika']

```
#just to test cosine
def cosine(u, v):
    return np.dot(u, v) / (np.linalg.norm(u) * np.linalg.norm(v))
```

"""Let's test the most similar news with different queries and test the performance of the different algorithms."""

```
from sklearn.metrics.pairwise import euclidean_distances

#from sklearn.metrics.pairwise import jacard


#descriptionss = ["Bombarding villages or towns", "willful killing", "Pillaging property",
"Sexual slavery","Taking hostages", "Destructing Properties", "Enforced pregnancy "

#"Looting",     "Raping ",       "Bombarding villages or towns", "Humiliating",        "Arrest" ,
          "seizing property ",

#"Genocide","conscripting or enlisting children for war", "Spy", "displacing civilians ","Rebel"]

description = ["displacing civilians ", "Massacare", "Pillaging property",     "Raping ",
"Humiliating","AmharaGenocide",  "Bombarding villages or towns", "Genocide", "conscripting
or enlisting children for war", "Willful killing", "Arrest", "Rebel"]


for description in descriptionss:

    print("Description: {}".format(description))

    print()



    tf_idf_similar_indexes = find_similar(tfidf_vectorizer.transform(["
".join([lemmatizer.lemmatize(token.lower()) for token in word_tokenize(description)])]),
descriptions_representation_tfidf, K)

    print("5 most similar descriptions using TF-IDF")


    for index in tf_idf_similar_indexes:

        print(articles_titles[index])


    #print(" tweets percentage: {} %".format(100*len(tf_idf_similar_indexes)/len(articles_titles)))

    print()



    infersent_similar_indexes = find_similar(infersent.encode([description], tokenize=True),
loaded_infersent_trained_vector, K)
```

```
    print("5 most similar descriptions using Infersent")

#print(cosine_similarity(infersent.encode([description],tokenize=True),embeddings_infersent))

#similarity=cosine_similarity(infersent.encode([description],tokenize=True),loaded_infersent_trained_vector)
    for index in infersent_similar_indexes:

print(articles_titles[index],np.fliplr(np.sort(cosine_similarity(infersent.encode([articles_titles[index]],tokenize=True),loaded_infersent_trained_vector,K))))
    print()


    distilbert_similar_indexes = find_similar(model.encode([description]), loaded_distilbert_trained_vector, K)
    print("5 most similar descriptions using Sentence-Bert")
    for index in distilbert_similar_indexes:

print(articles_titles[index],np.fliplr(np.sort(cosine_similarity(model.encode([articles_titles[index]]), loaded_distilbert_trained_vector, K))))
    print()


doc2vec_similar_indexes = find_similar([doc2vec_model.infer_vector([token for token in nltk.word_tokenize(description.lower())], alpha=start_alpha, steps=infer_epoch)], loaded_doc2vec_trained_vector, K)
    print("5 most similar descriptions using Doc2Vec")
    for index in doc2vec_similar_indexes:
        #print(news_description[index])

print(news_description[index],np.fliplr(np.sort(cosine_similarity([doc2vec_model.infer_vector([token for token in nltk.word_tokenize(news_description[index].lower())], alpha=start_alpha, steps=infer_epoch)], loaded_doc2vec_trained_vector, K))))
    print()
```

Appendix 4: tools used

cloud based Google Collaborators that supports GPU enabled robust sentence embedding network architectures with Keras in the back-end.

Python programming language and dependent text processing libraries such as pandas and Numpy has been utilized.

The google drive has been also used as a temporal storage for training and testing dataset and also for trained models pickle file location.

a distribution of the Python and R programming languages for scientific computing such as Anaconda and jupyter notebook.