Addis Ababa University

College of Natural Science

Design of a Spatially Aware Amharic Web Content Retrieval

Masreshaye Worku

A Thesis Submitted to the Department of Computer Science in Partial Fulfillment for the Degree of

Master of Science in Computer Science

Addis Ababa, Ethiopia

Date: Sene 2010 [June 2018]

Addis Ababa University
College of Natural Sciences

Masreshaye Worku Edo

Advisor: Dr. Solomon Atnafu

This is to certify that the thesis prepared by Masreshaye Worku Edo, titled: *Design of a Model for Spatially Aware Amharic Web Content Retrieval* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Name	**Signature**	**Date**
Advisor:	_____	
Examiner:	_____	
Examiner:	_____	

Abstract

It is obvious that we are living in the information era. WWW has been central to the development of the information age and is the primary tool billions of people use to interact on the Internet. It fundamentally changed how we connect with each other. Location information is a kind of information people are interested in. The analysis and use of geographical content from web resources is currently an area of increased interest and research. Almost everything that we do can be regarded as having some form of geographical context. Over the past few decades, accessing geographical information has focused on the combination of digital maps and databases that characterise the majority of geographic information systems. In geographic information systems, geographic objects are generalized into geometric points, lines and polygons.

The objective of this work is to design a model which uses Amharic geo-ontology for the development of a spatially-aware Amharic web content retrieval. The model receives and parses the user request in order to identify the existence of spatial features, spatial relationships to discover the exact location phrases. The Amharic geo-ontology maintains a place name, the geographical footprints which indicate its spatial extent, and its topological relationships with other places. It plays a key role in the process of geo-parsing of web documents and generation of spatial indexes. The geo-parsing process is identifying the presence of place names and spatial relationships in the document. While the geo-coding process involves disambiguating place names with multiple spatial references. Once significant place names have been detected in a document, the geographical ontology will be used to provide footprints to the document. A spatial index of web documents will be created based on the footprints. The spatial relevance ranking is based on measures of distance between the query footprint and the document footprint.

A prototype application is developed to evaluate the proposed solution. A query is formulated in three different formats, which are: <Place Name >, <A Place Name, Spatial Relationship, Geographic Feature Type>, and <Geographic Feature Type>. The evaluation is performed using 275 known place names, 38 cities in Ethiopia, and 35 geographical feature types.

Key Words:

geo-ontology, place name, geographic feature, geographic feature-type, spatial relationship, geo-coding, geo-parsing, spatial indexing.

This thesis work is dedicated to

my mother

Shegie Chiquala

Acknowledgements

I would first and foremost like to thank God for the strength and wisdom He gave me to bring this thesis to completion.

I would also like to thank my thesis advisor Dr. Solomon Atnafu for his support throughout the course of writing this thesis.

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction

1.1 Background

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Information retrieval deals with the representation, storage, organization of, and access to information items. Nowadays, information retrieval systems are used to provide access to a variety of resources audio and video files, books, journals and other documents. The Web has made major collection of information accessible to everyone, anywhere. However, due to the huge amount of information and some inherent properties of the Web, accessing information on the Web is not a trivial task. Web search engines are one and the most used information retrieval applications [1].

Fetching information from the Web can be done through search engines. These days, a number of search engines are available online. Most of them are general purpose search engines and work primarily for the English language or other popular languages. There are other language specific search engines which are designed and developed to work with a particular language [2] . Amharic language is one of those languages a search engine is designed for [2,10,11,12].

Geographical information is recorded in a wide variety of media and document types. There are innumerable paper-based books, reports, images and maps; there are computer databases and digital maps along with vast number of web pages containing text, images and digital versions of articles, books and reports. Over the past few decades, information technology for accessing geographical information has focused on the combination of digital maps and databases that characterise the majority of geographic information systems (GIS) [3]. Nowadays, generating data in geographic domain is growing explosively, which leads to increasing the applications of GIS. Due to this, GIS faced the problem of effective management of the spatial information from different resources and in different forms [4].

Geographic search can be seen as a key element of a semantic web. Conventional search engines lack the ability to consider semantics, and don’t consider geographic qualifiers

which are used to describe spatial relationships such as ‘inside’ or ‘near to’ [5]. Documents relating to somewhere inside a large place may not mention about that place name but may refer instead to a district inside that large place. Geographic names may be different in different languages or may be used in a non-geographic context. These and other limitations can only be addressed if geographic information is treated accordingly, and in particular if spatial relationships can be in some way stored and used in analysis of the results of a query [5].

Egenhofer [6] describes some typical issues through the example of the query “lakes in Maine”. If this query was submitted to a conventional search engine, the spatial relationship “in” would generally be treated as a stop word and discarded from the search. The author suggests that it is possible that data describing such lakes might be available but described only in documents which name counties in Maine, but not Maine itself. Thus, some form of spatial join must be made between data describing the lakes and counties (and their geometry) and counties and the state of Maine (and its geometry). Egenhofer suggests that a central challenge for a geospatial web is therefore that “it captures, analyses and tailor geospatial information, much beyond the purely lexical and syntactic level”.

Adopting a geo-ontology will enable the search engine to detect that the query refers to geographic locations. Information can be integrated based on its meaning. This can be done by integrating ontologies that are linked to sources of information. The goal of building ontology for the study domain is to capture the knowledge of the domain, to provide a corporate cognition, to define terminologies accepted collectively and to present the relation of these vocabularies in varied formalization modes [7]. Geographical ontologies have been introduced to the geographical domain not only as a concept model which can represent objects on semantic and knowledge level, but also to integrate geographical data from different sources and in different forms for reasoning and to facilitate knowledge sharing, in addition to assist in recognizing spatial terms [4].

The subject of ontology also is an important field of research in geographic information science. Geospatial ontologies have been used as a means of knowledge-sharing among different user communities, thus improving interoperability among different geographic databases [8]. Geo-ontology is used for realizing the organization, sharing and interoperability of knowledge in the geoscience field. By using semantic for building geo-ontology cooperatively, it is possible to manage varied semantic layers of geo information in the flat [7].

In order to provide a description of geographic entities, a geo-ontology can be conceptualized in two different views of the world. The field view considers spatial data to be a set of continuous distributions while the object view conceives the world as occupied by discrete, identifiable entities. A geo-ontology is different from other ontologies because topology and part-whole relations play a major role in the geographic domain. Geographic objects can be connected or contiguous, scattered or separated, closed or open. They are typically complex and have constituent parts. Representing geographic entities either constructed features or natural variation on the surface of the earth is a complex task. These entities are not merely located in space; they are tied intrinsically to space. They take from space some of its structural characteristics [9].

A geo-ontology takes into consideration not only semantic relations but also spatial relations like adjacency, spatial containment and connectedness. It describes entities, semantic relations, and spatial relations. A geo-ontology also has two basic types of concepts: concepts that correspond to physical phenomena in the real world and concepts that correspond to features of the world that we create to represent social and institutional constructs.

1.2 Motivation

We are living in the information era, in which information becomes a very important issue in humans' day to day activities. Information can be gathered in a variety of ways. The world-wide web is the most sophisticated and easy to retrieve information form. The web provides large number of references for information that we are in need. The growth of the Web due to increases of network speed, number of internet users, storage capabilities and other internet technologies, has suggested that there would exist a much higher potential for exploiting the Web if tools were available that better match human requests [6]. Search engines are tools that make information retrieval, from the web, too easy. A large proportion of information available on the world-wide web refers to objects that may include one or more spatial components.

Location information is a kind of information people are interested in. The analysis and use of geographical content from web resources is currently an area of increased interest and research. Given that a large proportion of what people do is based around location, many web resources contain some kind of geographical context [3]. The geospatial domain is a challenging area due to the variety of data models, formats, semantics and relations.

Moreover, reasoning with geo-spatial information would benefit users by providing a richer set of information that they do not have to explicitly ask for in their queries [10].

Current geo-spatial search engines do not provide users with the flexibility they need in order to express their queries with their familiar vocabulary. For example, users have to adhere to specific terms that are explicitly defined in the target databases. This is not practical especially for Web search engines, which are most typically available to public and general users. On the other hand, semantic search enables users to ask for queries where a composite query is divided into parts and each part is answered by a different data source. While combining the answers into one final answer might require expensive geometric operations because of the nature of the geographical information, semantic search provides a better approach. Semantic search provides dedicated tools and mechanisms to efficiently process rules that infer knowledge and provide geometric comparison [11] that could contribute to the user query.

The nature of the graph pattern represented by RDF, OWL and SPARQL enables semantic spatial search to accept users' sub-queries defined in their vocabularies in a more flexible manner. This will increase user's acceptance to the semantic search. The required task will then be translating these vocabularies into a semantic query and process them where each sub-query answers a separate question and probably from the same or different data sources. Combining the answers to meet the final solution of the user can be done with the help of ontologies as they facilitate the validation process [10].

The variety of spatial relations such as contains, near, etc., are motivators for defining a geo-ontology in which the properties of these relationships better interpret at and accurate for geo-queries. Also, it was necessary to approximate these terms with their possible contexts in order to deal with them correctly when they were encountered as properties of some resources in the knowledge base. Consider the examples provided below:

For instance, if the user asked for "በአዲስ አበባ የሚገኙ ትምህርት ቤቶች", then we can instead retrieve "ትምህርት ቤቶች" to which the "አዲስ አበባ" entity has originated a relation "የሚገኙ" which is a containment spatial relationship. However, such an assumption can only be made about the spatial relation if "የሚገኙ" was defined somewhere in the geo-ontology.

Consider the query, "በአዲስ አበባ ዩኒቨርሲቲ አቅራቢያ ያሉ ሆስፒታሎች". While the "አቅራቢያ" relation has no exact boundaries, proposing a boundary by the system can yield a behavior

that seems strange to some users. Therefore, a fuzzy [12] boundary can be used and applied, but very close to "አዲስ አበባ ዩኒቨርሲቲ", can be considered and given a rank value.

1.3 Statement of the Problem

As a matter of fact, Amharic language is one of the morphologically complex languages in the world [2]. Search engines that are designed for specific languages are more effective at handling queries of that language than general purpose search engines.

Search engines for Amharic web content retrieval have been done so far by different researchers. Those are research works for retrieving information from Amharic web content based on the given Amharic search keywords. In addition to that general-purpose search engines such as Google, provide service of Amharic language-based searching, but it doesn't support or provide spatial-aware searching service for Amharic language. For instance, if we submit a query "በአዲስ አበባ የሚገኙ ሆስፒታሎች" to Google, it considers the spatial relations "በ … የሚገኙ" as a search key and returns a search result for it.

Geographic search can be seen as a key element of a semantic web. Conventional search engines lack the ability to consider semantics, geographic qualifiers describing spatial relationships such as inside, north of or near are not treated geographically.

As far as our knowledge is concerned, which is based on the review of related works, none of the research works done before didn't design a model for spatially-aware Amharic web content retrieval. That means there is no Amharic search engine that considers spatial relationships while retrieving information from any Amharic web content. So that, while we are searching for an event which are related to location, the key word which shows the spatial relationship may be considered as a stop word and/or may be thrown away by the search engines or the search engines may consider them as a single search keyword and lookup for their matching word on the web contents. The results returned may become far away from the user's request.

These limitations can be manifested in a failure to distinguish between different instances of the same place name, a lack of ability to deal with spatial qualifiers such as "near" or "north", a lack of methods to rank and explore results with respect to their geographic relevance and the non-retrieval of resources which are geographically relevant but use a place name different from that specified in the query.

Therefore, there is a need to improve the quality of geographically-specific Amharic web content retrieval that understands the spatial relationship of events.

1.4 Objectives

General Objective

The main objective of this work is to design a model which uses Amharic geo-ontology for the development of a spatially-aware Amharic web content retrieval.

Specific Objective

The specific objectives of this work are to:

- explore related works on spatially-aware web document retrieval;
- design a model for spatially-aware Amharic web content retrieval;
- detect geographical references in the form of place names and associated spatial natural language qualifiers within text documents and in users queries;
- design an algorithm that disambiguates place names to determine which particular instance of a name is intended;
- index documents with respect to their geographic context as well as their non-spatial thematic content;
- develop a prototype for spatially-aware Amharic web content retrieval;
- evaluate the prototype for effectiveness.

1.5 Methods

In order to achieve the objectives of the research the following methods will be deployed.

i. Literature Review

Different literatures that are relevant to the subject of this thesis work will be reviewed. Documents which can add value from a variety of sources, such as journals, proceedings of conferences, books, research works, and materials from the internet will be considered.

ii. Identification and Usage of Tools

It is obvious that a variety of tools are required to design, implement, build prototype and to evaluate the findings of the research. So that, tools which are necessary at every stage

of the intended research work will be identified and selected while the research is going through. A tool which best fits for each task will be used.

iii. Build a Prototype

In order to evaluate, experiment and to show how our spatially aware Amharic web content retrieval perform, a prototype system will be developed. The performance of the prototype will be evaluated.

iv. Evaluation

One method of achieving the objective of the research work is performing experiment. So that experimentation will take place based on the finding of the research. The result of the experiment will be evaluated.

1.6 Scope and Limitations

The proposed research work is limited to the designing and implementation of a geo-ontology and adopting it for Amharic web content retrieval search engines.

Defining an Amharic ontology other than that of spatial relationships and the implementation of a semantic search engine for Amharic web content retrieval is out of the scope of this research work.

1.7 Application of Results

The output of this work will add a value for the Amharic language web content retrieval by providing additional feature which is awareness to spatial relationships to the existing Amharic search engines.

It will be applicable in industries which are concerned on Amharic language-oriented location-based services. In addition to that, it will contribute for the semantic web content retrieval applications which are focused on Amharic language.

It will also be an input for researchers who are interested to study on related areas.

1.8 Organization of the Rest of the Thesis

The remaining part of this work is organized as follows. Chapter Two deals with the review of literatures in the domain of our interest, which covers discussion on Information Retrieval, Geographical Information Retrieval, Geo-ontology, and Spatially-Aware web content retrieval. Chapter Three presents works that are related to the theme of our thesis work. Those related works are summarized and we identified the gaps at the end of the chapter. The Fourth Chapter presents the "Proposed Solution" with its high-level architecture and details the components of the proposed solution. We present the evaluation of our prototype and summarized the findings in Chapter Five. Finally, conclusions are drawn from the work and recommendations are given on possible future works in Chapter Six.

Chapter Two

Literature Review

This chapter gives an overview of the state-of-the-art concepts related to our work. The core elements of our work such as Information Retrieval, Geographical Information Retrieval, geo-ontology, and the concepts of spatially aware search engine are reviewed.

2.1 Information Retrieval

The World Wide Web has touched the lives of billions of people around the world and fundamentally changed how we connect with others, the nature of our work, how we discover and share news and new ideas, how we entertain ourselves and how communities form and function [13]. The World Wide Web or the Web is an information space where documents and other web resources are identified by Uniform Resource Locators (URLs), interlinked by hypertext links, and accessible via the Internet. The World Wide Web has been central to the development of the Information Age and is the primary tool billions of people use to interact on the Internet [14].Web pages are primarily text documents formatted and annotated with Hypertext Markup Language (HTML). In addition to formatted text, web pages may contain images, video, audio, and software components that are rendered in the user's web browser as coherent pages of multimedia content [15].

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs) [16]. The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler. Internet content that is not capable of being searched by a web search engine is generally described as the deep web [17].

Information retrieval is finding documents, usually text documents, that satisfy an information need from a large collections. Information retrieval systems can also be distinguished by the scale at which they operate. Information retrieval supports users in browsing or filtering document collections or further processing a set of retrieved

documents. For thousands of years, people have realized the importance of archiving and finding information [18]. With the advent of computers, it became possible to store large amounts of information; and finding useful information from such collections became a necessity. As a result of this, the field has matured considerably. Nowadays, information retrieval is becoming the dominant form of information access. Hundreds of millions of people need to deal with information retrieval every day. Several IR systems particularly web content retrieval systems are used on a daily basis by a wide variety of users [19].

Figure 2.1 shows the overview of information retrieval [12] which is broken up into five stages, those are Digital Library, Information Extraction (IE) Process Flow, Indexes, Information Retrieval (IR) Process Flow, and Information Visualization (IV) Process Flow. IE consists in extracting information from unstructured documents, represents it with appropriate descriptors, and organizes it into structured indexes for supporting IR. IR consists in matching user needs with indexed document descriptors. Finally, Information Visualization (IV) displays results of an IR process and supports browsing scenarios, as well as feedback for query reformulation [11,13].

The effectiveness of information retrieval can be measured with the collection of three things: A document collection, a test suite of information needs, expressible as queries, and a set of relevance judgments [19].

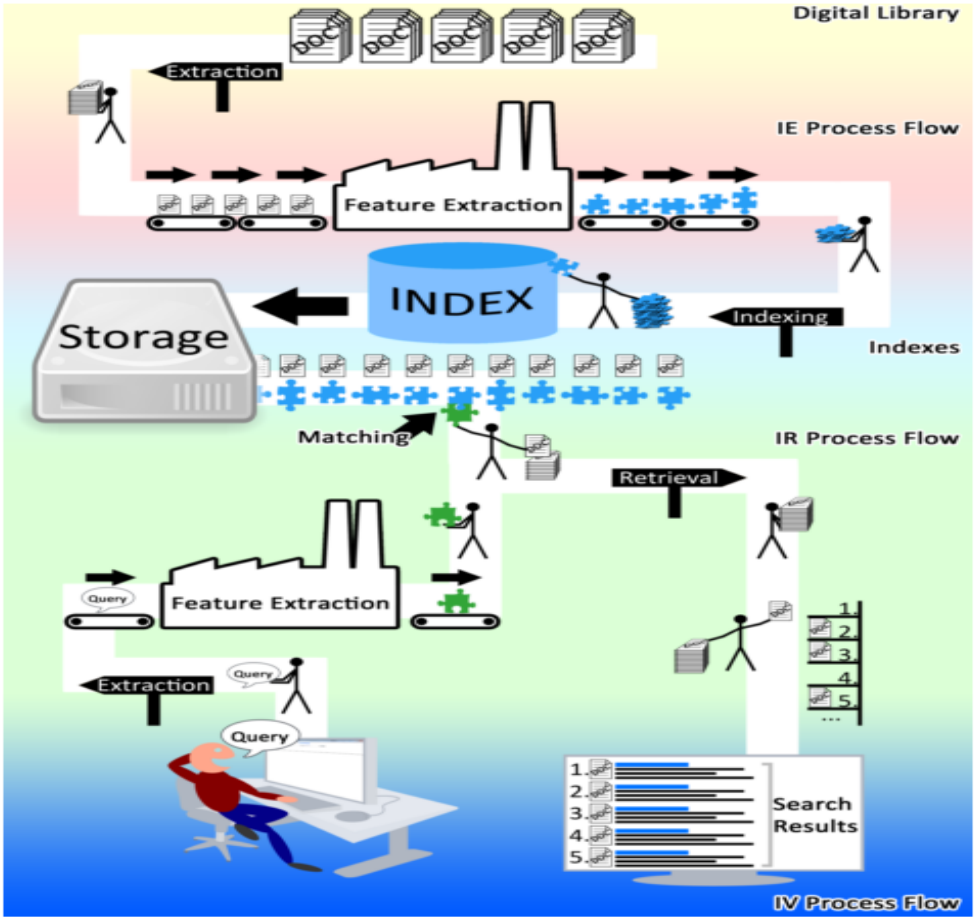

Figure 2.1: Stages in Information Retrieval

2.2 Geographic Information Retrieval

Almost everything that we do can be regarded as having some form of geographical context, and therefore many information resources refer in some way to location. Geographic information retrieval is necessary as most human activities are rooted in geographical space in some aspect, a large portion of web contents may be associated to some geographical locations. At least 20 percent of web pages included one or more easily recognisable and unambiguous geographic identifiers [5]. Web pages often contain place names in order to provide geographic context. Many web documents report on activities that take place in some locations on the Earth typically by means of place names. Some place names refer to places that have well defined boundaries, such as a country, while others refer to vague regions which do not have precisely defined boundaries, such as the South of Addis Ababa [4, 14]. Therefore, it is needed to enrich gazetteers with knowledge of such vague places and hence improve the quality of place name-based information

retrieval. It is important therefore that these vague places can be recognized within natural language queries and correctly interpreted with regard to their spatial extent [14, 15].

These types of data, mentioned as geographic identifiers before, are called Spatial or Geospatial data. Spatial data is a data that contains positional values. The more precise and a further refined phrase is geospatial data, which then means spatial data that is georeferenced. Spatial data that is not georeferenced can have positional data unrelated to the Earth's surface. Example, in molecular chemistry, the position of atoms in molecules are defined relative to each other, and in industrial design engineering, in which the parts of a car engine are defined relative to each other. So, they can be called spatial data but not geospatial/georeferenced data. In this paper, we will use 'spatial data' as a synonym for 'georeferenced data' [20].

Christopher J. Bones et.al mentioned in their work, to improve the quality of geographically-specific information retrieval, the following issues in GIR should be considered [3] :

- Detecting geographic references: it is the process of analyzing text to identify the presence of place names and other spatial language, which is geo-parsing.
- Disambiguating place names: since, there are many names that are shared between different places, determine uniquely the place to which the name refers.
- Vague geographic terminology: place names that users employ when searching on the web are of an informal nature, without precise boundaries. The spatial language, such as near, close, between and north of, that accompanies place name terminology can be as vague as some of the place names.
- Spatial and textual indexing: When web documents have been categorized according to their geographical context they must be indexed in a way that enables them to be found quickly in response to user queries.
- Geographical relevance ranking**:** relevance with respect to the query and the retrieved documents can be represented by a score that the frequency of occurrence of query terms within retrieved documents. The spatial score can be some measure of the geometric match between the query footprint and the document footprints. These two scores can then be combined to find an overall relevance.
- User interfaces: Query formulation generally requires specification of a triplet of <theme><spatial relationship><location>. This can easily be facilitated by a simple

structured interface. Other approaches to query formulation allow users to sketch a region of interest on a map and enter a related concept within a textbox.

2.3 Spatially Aware Search Engine

The semantic web is an extension of the current web, in which information is given explicit meaning, to make it easier for machines to automatically process and integrate information available on the web. The geospatial semantic web adds space and time dimensions to the semantic web [21] . Since location is a common element of information on the web, queries submitted to search engines also contain a reference to location. Since the conventional information retrieval methods are based on keyword matching, such geographic terms are treated by conventional search engines in the same way as other terms. Thus, they lack the ability to consider semantics, geographic qualifiers describing spatial relationships such as "inside", "north of" or "near" are not treated geographically, hence important semantic information will be discarded [4,18]. This has led to develop:

i. ontologies that specify the concepts and their relationships in the domain of geographical information, and

ii. spatially-aware search engines to find places or other resources corresponding to the places that are referred to in a query [22].

A geo-ontology has a key role in the development of a spatially aware search engine, to assist in recognizing place names and spatial relationships, which encodes geographical terminology and the semantic relationships between geographical terms [23]. For the geo-ontology, it is important to figure geo-spatial property of geographical concepts besides showing property of attribute, especially geo-spatial relation [24]. It provides support for query disambiguation, query term expansion of the required query, spatial indexing, geographical relevance ranking of the retrieved search results, the extraction of geographical references, creation of the spatial indexes, methods for the formulation of spatial queries and interaction with the results of geographic search [4,20,21].

"For geographical search, there is a need to match the geographical component of the query with the geographical context of the documents as represented by the document footprints. GIR queries can be characterized as a triplet of <theme><spatial relationship><location> composed of a topic of interest in combination with a place name qualified by a spatial preposition such as near, in, or north of. The combination of place name (after

disambiguation) and spatial preposition can be used to generate a query footprint. This query footprint can then be used to access the relevant part of the spatial index and hence find document footprints that intersect the query footprint" [3].

Spatial relations are the core contents of the geographic information, and they play a major role in spatial data model and spatial query. In geographic information systems geographic objects are generalized into geometric points, lines and polygons. However, in the real-world geographic objects are not simply geometric objects but spatially distributed objects with geographic semantics. Therefore, the geo-ontology knowledge base needs to be integrated with the spatial relation query system to realize the specific query [25].

After a query has been interpreted, it must be submitted to a search engine which incorporates techniques to deal with the thematic and geographic aspects of the query. Final the search engine will retrieve and present a sorted ranked list of results to a user, where more relevant documents are listed higher than less relevant ones. Geographical relevance is defined as a relation between a human's geographical information needs and geo-referenced information objects (e.g. documents, maps, etc.) [4,11].

2.3.1 Query Formulation

The first step in the search process is the specification of a query. For any retrieval process to be effective, it has to be possible to formulate queries, extract the right features from the query, and retrieve the most matching documents. Matching documents then need to be presented in an adequate way to the user. In GIR systems, a query can be characterised by a triplet containing a thematic component, a geographic component (place name) and some form of spatial relationship that links them, <theme><spatial relationship> <location> [3]. In other words, the triplet is composed of a topic of interest in combination with a place name qualified by a spatial preposition such as near, in, or north of. The combination of place name and spatial preposition can be used to generate a query footprint. An example is <Hospitals><in><Addis Ababa>. This query footprint can then be used to access the relevant part of the spatial index and hence find document footprints that intersect the query footprint. [4,11].

2.3.2 Information Extraction

Before spatial information can be extracted from texts, they need to be appropriately processed. GIR systems thus scan texts for occurrences of geographic content and convert them to geographic features. These features can then be assessed in the context of

geographic relevance. There should be a method to detect the presence of place names and other geographical references from a given set of web pages, this is called geo-parsing. Once a place name has been identified it is “grounded”, i.e. a map coordinate is allocated to it using the place name resources described in geo-parsing. The process of finding and extracting locations from texts and assigning coordinates to these places is called geocoding. It can only be accomplished through geoparsing, where **N**atural **L**anguage **P**rocessing (NLP) is used to identify geographic references in texts. A location that is found, disambiguated and assigned a coordinate (e.g. a GPS coordinate with latitude and longitude in degrees) is then geo-referenced. It knows its geographic location , which can then be exploited in further processing steps [11, 14].

2.3.3 Spatial Indexing

An index enables efficient filing and fast retrieval. If no index is used, in the worst case, all documents need to be searched thoroughly to find the desired one. The purpose of indexes is to decrease search time to a bearable minimum [26]. A spatial index of web documents can be created in a similar way to a spatial index of a geographic data set. Each document is allocated one or more geometric footprints, typically in the form of polygons. When web documents have been categorized according to their geographical context, they must be indexed in a way that enables them to be found quickly in response to user queries. The text indexing can be combined with a spatial index that records which documents relate to particular regions of space. Searching this index requires text query terms to be concatenated at run time with the identifiers of spatial cells intersecting the query footprint, prior to matching the transformed query terms with the spatio-textual index terms. Building a spatial index of documents can be done if each document has one or more document footprints that represent the regions of geographic space to which the document refers [3,23].

During the computation of spatial indexing MS-SQL server treated geographic objects as a geodetic ellipsoid. To decompose this space, the geography grid tessellation scheme divides the surface of the ellipsoid into its upper and lower hemispheres and then performs the following steps as shown in Figure 2.2:

i. Projects each hemisphere onto the facets of a quadrilateral pyramid.
ii. Flattens the two pyramids.
iii. Joins the flattened pyramids to form a non-Euclidean plane.

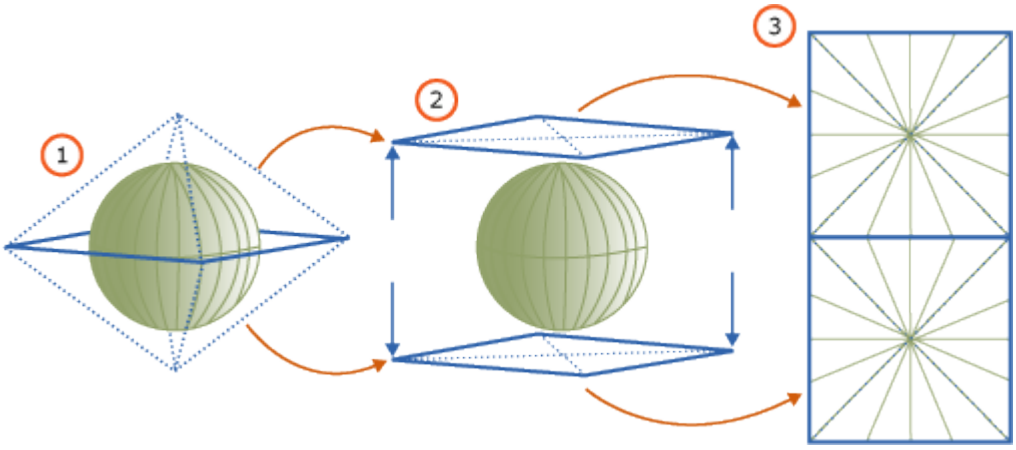

Figure 2.2 : three steps of Decomposition Process [36]

Once the space has been projected onto the plane, the plane is decomposed into the four-level grid hierarchy.

In SQL Server, spatial indexes are built using B-trees, which means that the indexes must represent the 2-dimensional spatial data in the linear order of B-trees. Therefore, before reading data into a spatial index, SQL Server implements a hierarchical uniform decomposition of space. The spatial indexing flow is illustrated in Figure 2.3.

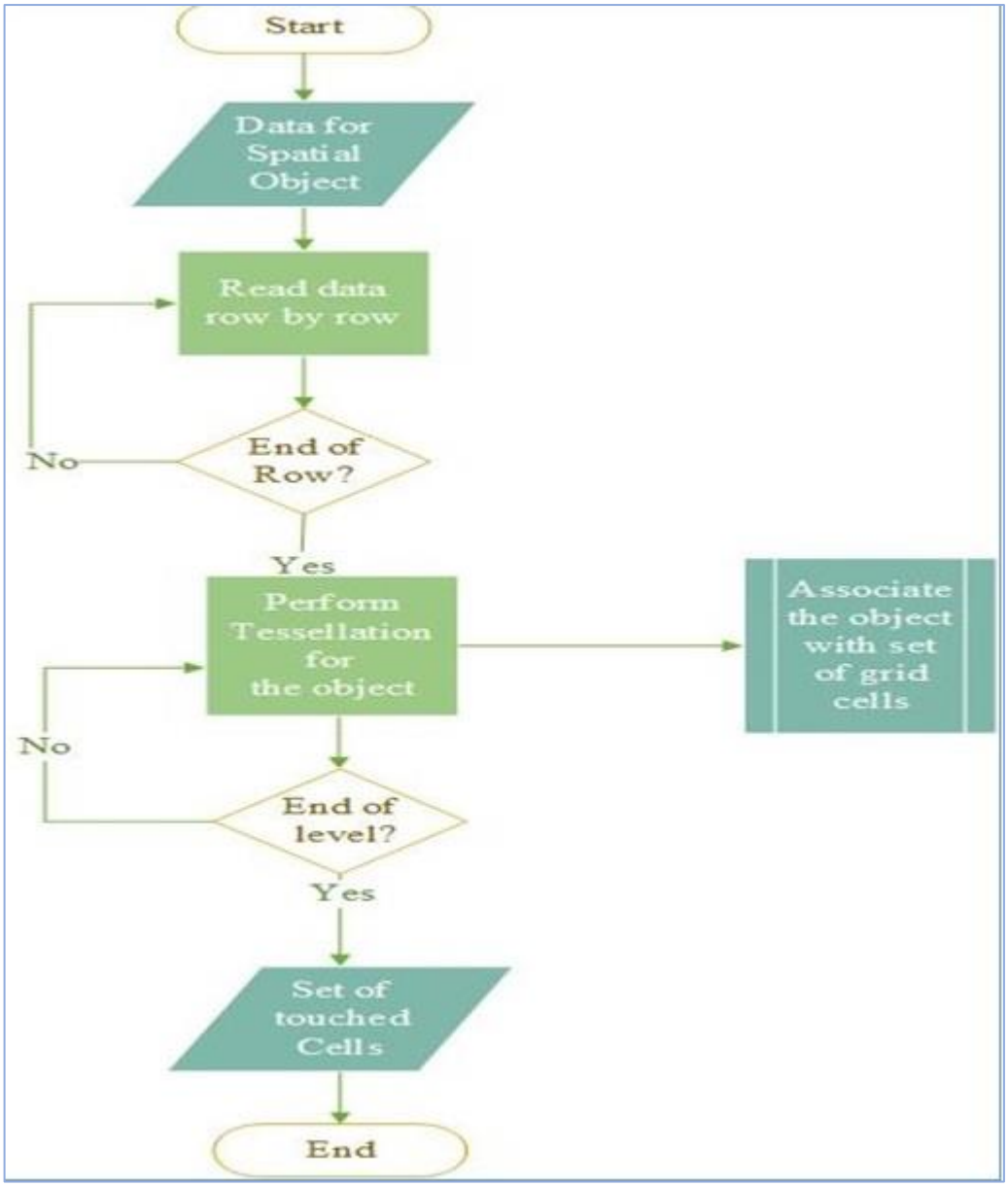

Figure 2.3: the flow of Spatial Indexing in SQL server

The index-creation process decomposes the space into a four-level grid hierarchy. These levels are referred to as level 1 (the top level), level 2, level 3, and level 4 [27]. Each successive level further decomposes the level above it, so each upper-level cell contains a complete grid at the next level. Figure 2.4 shows the decomposition for the upper-right cell at each level of the grid hierarchy into a 4x4 grid.

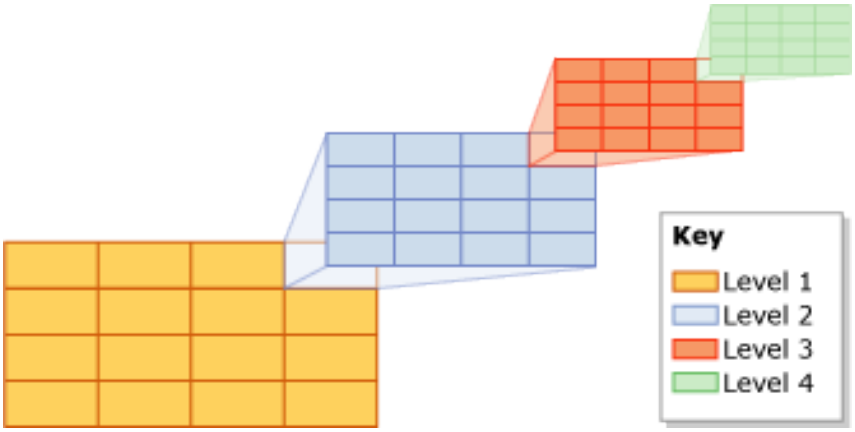

Figure 2.4: Decomposition of cells [36]

2.3.4 Relevance Ranking

Web documents identified by the search engines are ranked according to both textual and spatial relevance. Spatial or geographical relevance is defined as a relation between a human's geographical information needs and geo-referenced information objects [26]. From a geo-spatial perspective, each document in the web document collection is represented by footprints following the grounding of web locations to places, and a query is also represented as a footprint. When all footprints in a document are assigned a similarity score with respect to the query footprint, a document spatial similarity score for the document can be calculated [5]. The relevance ranking component combines the spatial and textual document scores to generate a single ranking. However, practically, geographical relevance estimation is also implemented as similarity measures. The purpose of relevance ranking in GIR is to return a ranked list of documents satisfying topical, spatial, and temporal criteria altogether [28].

The authors of SPIRIT in [5] list the steps to be taken to produce a relevance ranking of documents with respect to a query, which are:

- For every document footprint, a footprint similarity score is produced with respect to the query footprint.
- For every document, a document spatial similarity score is produced based on the footprint similarity scores of all the footprints contained in the document.
- Document spatial similarity scores are usually combined with textual scores into a document similarity score.
- Documents are ranked in descending order of their document similarity scores.

2.4 Geo-Ontology

Geo-ontology, a kind of domain ontology, is used to make the knowledge, information and data of concerned geographical science in the abstract to form a series of single object or entity with common cognition. This single object or entity can compose a specific system in some certain way and can be disposed on conception and given specific definition at the same time [24]. In the work of Hong Wang et.al [28] geo-ontology is defined as:

Geo-ontology = {C, R, A, X, I}, Where C (concepts) represents the concept set of geographic objects; R (relation) is a relation set and it mainly describes the relation set among concepts; A (attribute) shows attribute's set of geographic objects; X (axioms) is a lot of axioms and it is a constraint rules among the concepts, relation and attributes; I (instances) in a material object and I is a set of definition about instance.

Geo-ontology allows for definition and reasoning about real world spatial objects by combining information from multiple sources. A geospatial semantic approach to building applications allows for machines to understand and exchange qualitative spatial concepts [21]. A geo-ontology defines concepts that correspond to:

i. things from the physical and social world having a location on the surface of the earth and
ii. semantic and spatial relations (i.e. topology) between these things. Furthermore, these concepts are associated either with discrete geographic objects that have well defined boundaries or with continuous fields over space [22].

2.4.1 The Hierarchy of Geo-Ontology

Geo-Ontology systems may be divided into three layers' structure as shown in Figure 2.5, which are the Top-Level Geo-ontology with the highest-level abstraction, the Domain Geo-ontology used by specific application field and the Application Geo-ontology [29].

Application Geo-ontology is atomic geo-ontology with the smallest granularity and richest spatial information contents, as well as the least sharing and reusing facets. Domain Geo-ontology is a concept system, which contains the common sharing knowledgebase used in a spatial information domain. Top level Geo-ontology is a concept system which describes the most general spatial features and phenomena and has the biggest sharing extent [3]. Meanwhile, Figure 2.5 shows that the extent and scope of information sharing, interoperation and generalization will be going up with the geo-ontology level going up.

However, the semantic content of geo-ontology will become not so detail and the user will get less information content in the top-level geo-ontology.

The application ontology layer is consisted of formal application sub-layer and specific application sub-layer. Any layer of the above may be empty layer [29].

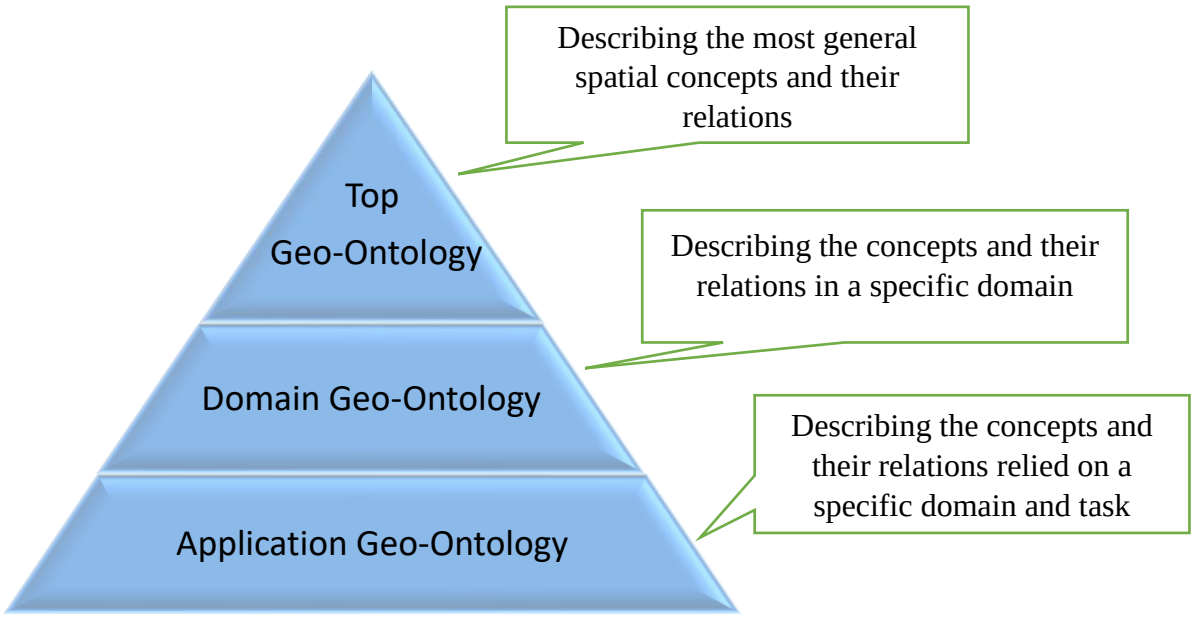

Figure 2.5: Hierarchy of Geo-Ontology

2.4.2 Building Geo-Ontology

The main aim of constructing geo-ontology is to get the knowledge of the domain of geography, and to provide the common vocabularies in the domain [24]. The SPIRIT [30] project proposes the main components of geo-ontology as shown in Figure 2.6.

The main processes of building geo-ontology are listed below, as described in [31]:

i. Confirm the scope of geo-ontology
ii. List the ontology property of geographical concept
iii. Ensure the relationship among geographical concepts
iv. Collect concepts' meaning, attributes, and instance
v. Prototype system of geo-ontology

The work, "Research on Geo-Ontology Construction based on Spatial Affairs" describes the principles of constructing geo-ontology as follows [24]:

i. Legibility: Geo-ontology must explain the meaning of defined glossary by using natural language in manner of being background-independent and integrated.
ii. Consistency: all the defined axioms by using Geo-ontology and the documents explained in natural language should have the property of consistency.
iii. Extensibility: geo-ontology can provide the foundation to define new attribute under the condition of existent concept without modifying the definition of concept being used.
iv. Coding with the least leaning degree: the description of concepts should not rely on a certain kind of expressing methods.
v. Least restrictions: users can realize the specialization and instantiation the geo-ontology they needed in fact, reducing possible restrictions of modeling objects

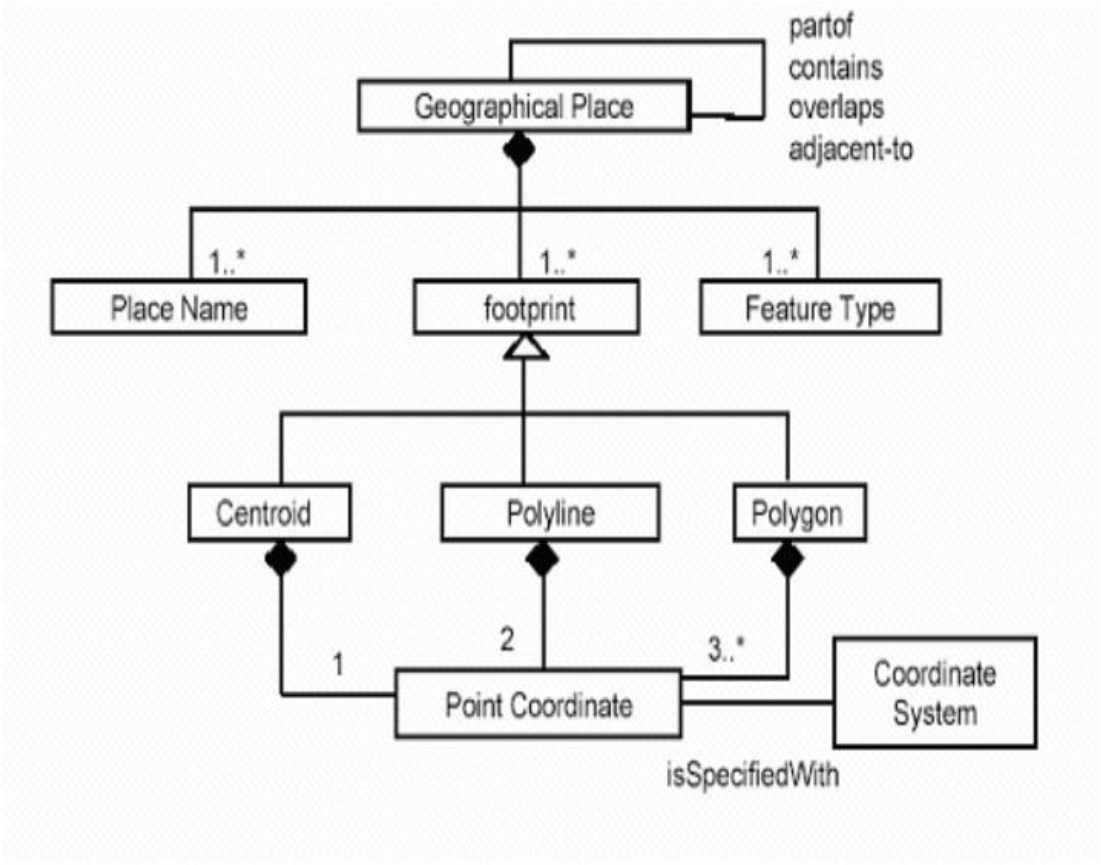

Figure 2.6: Conceptual Design of Geo-ontology: SPIRIT

Maintaining the consistency and the integrity of the geo-ontology is essential for supporting the correct functionality of the search engine. Possible maintenance tasks needed when building the ontology base are:

- Ensuring that all mandatory relationships are satisfied, e.g. every geographic feature belongs to every at least one geographic type and has at least one associated footprint.
- A polygon footprint with more than two points, must have at least four points, with the first point being equal to the last point.
- For two features in a containment relationship, the bounding box of the child must be enclosed in the bounding box of the parent.
- For two features in an overlap relationship, the building boxes of both must intersect [23].

2.5 Amharic Language and Geographic Information Retrieval

In the era of information technology, the Web becomes a universal channel for delivering information. It is clear that varieties of documents are available on the web presented with different languages, including Amharic language. Due to an increase in Internet population within Ethiopia, the number of web documents that are written in Amharic language is also increasing. Amharic language is highly diversified Semitic language throughout Ethiopia. With more than 25 million speakers, it is the second most spoken Semitic language in the World, next to Arabic. It is also the working language of the Federal Democratic Republic of Ethiopia. Amharic language has its own alphabet, ፊደል/ fidäl, which is inherited from the Geez (Ethiopic) language. The language's writing system contains thirty-four consonants. Each of which occurs in a basic form and six other forms known as orders. Each consonant has seven forms. The seven orders represent syllable combinations consisting of a consonant following vowel [2,31-33]. The word order in Amharic clauses is generally SOV, subject , object, verb. Verbs agree with their subjects in number, gender and person and objects precede verbs within the verb phrase [32].

Natural Language Processing (NLP) is a field of study in Artificial Intelligence and Linguistics that deals with teaching computers to understand Natural or Human Language. NLP technology can be used in virtually all the stages of information retrieval such as document processing, query processing, and matching. NLP can improve the performance of IR systems in many different aspects. Many NLP tools and techniques, including stop word removal, stemming, compound and statistical phrases, head-modifiers, word sense disambiguation, noun parsing, morphological analyzer and others have been used in information retrieval [33]. NLP platforms provide possibilities for spatial named entity

recognition and extraction from textual documents. Table 2.1 shows a summary of what NLP can do for GIR [26 - 27].

Table 2.1: NLP steps required to extract spatial features

Term	Description
Part of Speech (POS) tagging and Named Entity Recognition (NER)	Two ways of linguistic analysis: • POS: Process of sequentially labelling tokens with syntactic labels (e.g. nouns, verbs, etc.). • NER: process of finding mentions of predefined categories (e.g. names of locations).
Named Entity Validation (NEV)	• Knowledge-based resources validate candidate named entities (e.g. spatial features). • Real candidates are distinguished from false ones and geo-referenced (e.g. by assigning a coordinate). • Requires Gazetteer lookup (geographical dictionary containing location names, alternative location names, population, coordinates, etc.).
Named Entity Interpretation (NEI)	• Last step: find relations between tokens and collect meaningful token groups. • Uses knowledge-based resources for disambiguation and association of representations to spatial features and to analyze spatial relationships.

2.6 Tools and Language

In the process of modeling and maintaining the geo-ontology, there is a need of an ontology representation language. Which should be capable of the basic functionalities listed below:

i. Representation of real-world geographic concepts.

ii. Representation of data properties for each geographical concept.

iii. Representing relationships between concepts.

iv. Representing specialization/generalization concept hierarchies.

v. Representing simple composition hierarchies, i.e., concepts made up of sets of other concepts.

Desirable functionalities of the language:

vi. Representing constraints on data properties of concepts.

vii. Representing constraints on relationships.

viii. Expressing integrity rules over individuals

ix. Expressing integrity rules between individuals belonging to different concepts.

x. Representation of advanced composition hierarchies, i.e. a class of all houses which are within 10 miles of a motorway [36-23].

Li Bin et.al [24] declares that geo-ontology can be effectively constructed by using OWL (Ontology Web Language) combining with some tools, such as protégé. OWL is the standard for ontology languages. It is developed within a European research project and is supported by the W3 Consortium. OWL has strong ability of expressing semantic knowledge as well as features of property on geo-ontology. Among the languages of describing geo-ontology, OWL can be used as a comparatively good and appropriate language tool to deal with needed information. OWL can be used to express distinctly the geographical meanings of vocabulary entry in the glossary and relations between them. Such expression content is no other than geo-ontology [24].

In [34] it is mentioned that most of the authors, who works on the area, prefer Protégé over all other tools. Protégé has a wide range of import and export options and a good effort/result ratio, meaning that while it is not the most powerful tool it provides many users with as many capabilities as they need for an acceptable amount of learning.

Chapter Three

Related Work

Research works related to this thesis work have been reviewed. Those related works can be grouped into three general categories Amharic web content retrieval systems, geographic information retrieval, and geo-ontology (geospatial/geographic ontology)

3.1 Amharic Web Content Retrieval

The work of Tessema Mindaye [2] is one of Amharic search engines designed and implemented for the retrieval of Amharic language web documents. The search engine has three components, these are a crawler component, an indexer component, and a query engine component. These components are optimized for the language they are designed, which is Amharic language. The crawler component crawls the Web and collects web pages that have Amharic content with Unicode encoding. It stores the downloaded Amharic web documents in a file structure (repository) based on document type. The indexer component processes the documents in the repository to create the index. It tokenizes character streams into words based on Amharic punctuation marks and white space. Finally, it indexes the words in an inverted index structure with some statistics using Lucene. The query engine component used this index to search related document set and rank the results returned for the submitted queries. The Query Engine component accepts the user query in Amharic and searches the Lucene based index for the parsed query. After parsing and pre-processing the user query, the query engine component first calculates the text-based similarity of the query with that of the document in the index. After getting this similarity, it will combine it with link-based ranking and present the result in descending order to the user.

Hassen Redwan redesigned and developed an enhanced Amharic Search Engine [33]. This search engine allows indexing and searching of Amharic documents written in multiple character set representations of the Amharic language. The author redesigned the engine in such a way that non-Unicode based Amharic documents can also be recognized by the crawler component of the engine. The non-Unicode Amharic documents are converted to Unicode based encoding and passed to the Indexer component where by the document index is created for both types of documents together. The indexer component performs analysis of Amharic documents by using Amharic analyzer. Ordered processing is done

by the analyzer are Amharic tokenization, stop word removal, alias analysis, stemming and finally indexing [33]. It also implements multi-threading to improves the performance of the search engine's crawling process.

Tessema et al. [35] designed to index only Amharic language documents available on the Web and make them available for search. It has three components: An Amharic Crawler, the Amharic Indexer, and the Amharic Query Engine Component. These components have unique sub-components that give special consideration to the typical characteristics of the language they are designed for. The Crawler is responsible for collecting Amharic language documents from the Web. The Indexer component builds indexes from documents that it gets from the Crawler. It also tokenizes, removes stop words, apply stemming to the words before it indexes. This index is used by the Query Engine to satisfy user needs.

This search engine takes the typical feature of Amharic language into consideration. Such as Morphological variant of words, Repetitive alphabets (fidels), Short form of compound words, and spelling variation of a word. These features affect the retrieval of an Amharic language document from the Web. The Normalization component will take care of these Amharic language unique features. The downloaded Amharic pages are passed through different processes before they are stored in the Index. The character streams will be tokenized into words by taking white space and Amharic punctuation marks as word demarcations. Shorter forms of a word will be replaced by its expanded form. If the word is a stopword or if it is a variant of a stopword, it will be removed. The remaining non-stopwords will be stemmed to reduce them to their common form. The final result of all these processes is stored as an index in a structure that is appropriate for fast searching.

The work of Solomon Atnafu and Mequanint Munye [36] is a generic model bilingual Web search engine for Amharic and English languages. The search engine has different components Amharic query preprocessing, English query preprocessing, Amharic query translation, English query translation, Amharic and English search engines are the major components of the work.

The query preprocessing is done in two different modules. The Amharic query preprocessing module is responsible for tokenizing the Amharic queries into words, normalization of Amharic redundant symbols which have the same sound and expanding short form of Amharic words, eliminating stop-word, and stemming inflectional and some

derivational Amharic morphemes. The output of Amharic query preprocessing module is a set of Amharic preprocessed bag of words which is used as an input for the Amharic query translation component. The English query preprocessing module, is responsible for tokenizing the English queries into words, eliminating English stop-words", and stemming inflectional and some derivational English morphemes. The query translation process is responsible for lexical transfer or dictionary lookup in the bilingual Amharic-English dictionary and transliterating the out of dictionary words assuming that they are proper names or borrowed words. The search engine component is responsible for crawling, indexing, and ranking of the Web documents.

3.2 Geographic Information Retrieval

In the work of Christopher J. Bones and Ross Purves [3], it is recommended that taking user needs into consideration is very important while developing Geographical Information Retrieval (GIR) techniques and to develop techniques to evaluate the quality of approaches to GIR. It also identifies that information retrieval methods usually identify web pages that contain query terms and rank documents using statistical methods that are intended to highlight the most relevant. For some forms of geographical search, particularly when looking for common resources within relatively large geographic extents, this approach can work but it is fraught with limitations. From the user's perspective these limitations can be manifested in a failure to distinguish between different instances of the same place name, a lack of ability to deal with spatial qualifiers such as "near" or "north", a lack of methods to rank and explore results with respect to their geographic relevance and the non-retrieval of resources which are geographically relevant. GIR is, therefore, concerned with improving the quality of geographically-specific information retrieval with a focus on access to unstructured documents such as those found on the web.

The authors also identified the issue to be considered in the process of developing geographical information retrieval system. These issues are detecting geographical references in the form of place names and associated spatial natural language qualifiers within text documents and in users' queries, disambiguating place names to determine which particular instance of a name is intended; geometric interpretation of the meaning of vague place names, and of vague spatial language such as "near", indexing documents with respect to their geographic context as well as their non-spatial thematic content;

ranking the relevance of documents with respect to geography as well as theme; developing effective user interfaces that help users to find what they want.

The SPIRIT (Spatially-Aware Information Retrieval on the Internet) project, which has demonstrated that retrieval from web resources can be improved by making search spatially-aware [5]. It allows both textual and graphical query formulation together with results presentation and a mapping backdrop. It defines geographical ontology which acts as a repository of knowledge about place names, and relationships between them, for the regions covered by the search engine. The work introduces spatial-awareness into search engine technology and the architecture used to enable this. The project provides a solution that allows documents to be indexed spatially as well as thematically which in turn enables a full set of geographical query operators, graphical query formulation, the ranking of results according to conceptual as well as spatial match to the original query, and the graphical display of search results.

SPIRIT comprises a number of components responsible for key spatial-awareness functionalities. It has a multimodal interface, allowing both textual and graphical query formulation, together with results presentation and a mapping backdrop. A *geographical ontology* acts as a repository of knowledge about place names, and relationships between them, for the regions covered by the search engine. Data in this repository are used to recognise the presence of place names in web resources. The geographic ontology is also used to disambiguate place names in a user query and to generate a query footprint that reflects the region of space to which the query refers. The query footprint is a geometric interpretation of the place name and the spatial relationship employed in the query. **A** *metadata component* attempts to associate the documents from the original web crawl with one or more footprints representing the regions of space to which individual documents relate. This is used in generation of a spatial index and in spatially-aware relevance ranking. Spatial indexing supplements text indexing by associating documents with one or more cells of a subdivision of geographic space.

3.3 Geo-ontology

The role of a geo-ontology in the development of a spatially-aware web document retrieval systems is identified and described by Christopher J. Bones et.al [37]. The authors stated that the geo-ontology may be used to disambiguate the place name expression in user queries and subsequently generate alternative place names and associated place names for

query expansion. They added that the geo-ontology could also be used to identify the presence of place names, spatial qualifiers and domain-specific terminology in the web repository. Geographical relevance ranking in the search engine needs to use the geo-ontology for the derivation of footprints and the footprints associated with the web resources. They also propose a base ontology model for the geo-ontology which aims to provide a foundation for subsequent implementation and experimentation.

In this paper the possible types of queries that a spatial-aware search engine is expected to handle are identified. The basic query expressions are listed below:

- A Place Name, or,
- A spatial Entity with a Relationship to a Place Name, or,
- A spatial Entity with a Spatial Relationship to a Place Name, or,
- A Place Name with a Spatial Relationship to a Place Name, or,
- A Place Type with a Spatial Relationship to a Place Name, or,
- A Place Type with a Spatial Relationship to a Place Type, or,

The geo-ontology proposed in this work has three components, a geographic feature ontology, a geographic type ontology, and a spatial relation ontology.

R. Henriksson et. al [22] proposes that the geo-ontology building process begins with the definition of an underlying conceptual model, which serves as a base for the geo-ontology. However, the conceptual model lacks consideration of the meaning of the geographical concepts, and hence can miss some information. The authors emphasize the fact that the theories and thoughts behind the geographical concepts should be made clear. In this paper a set of core geographical concepts, their mutual relations and properties are examined.

This work summarizes two points the content of geo-ontologies, and the approach of geo-ontology development process. The authors suggested that geo-ontologies should contain classes that describe the spatial aspects of places (e.g. location), regional geography (e.g. administrative regions), patterns based on human interaction with nature (e.g. land use), and aspects related solely to the physical environment (e.g. landforms).

Z. Xue and X. Jun [25] build a knowledge base mainly on the base of analysis of geographic semantic information implicated in the description of spatial relations, which

is constructed to assist spatial relation query. In order to solve the problem which related to geographic semantic and spatial relations, a geo-ontology knowledge base about spatial relations is established. An ontological knowledge base is established to store the knowledge related to spatial relations between geographic objects based on the semantic analysis of geographic relations, and human's cognition. Their main target is to establish the relationship between entities and attribute which are involved in spatial relations.

This work focuses on the problem of solving spatial relation query in different context, which mainly is performance. The knowledge base is integrated to the spatial relation query system to realize the specific query. The query system mainly includes three parts: the parsing model, the geo-ontology knowledge base, and the spatial relation query model. The parsing model is used to analyze the input natural language query sentences. The geo-ontology knowledge base used to provide the relations between the geographic objects, and the properties of the objects. The spatial relation query model calculates the indices which are used to quantifiably represent the spatial description words and call spatial operators. Spatial operators are functions which fulfil the spatial relation queries.

The authors of this paper use OWL as modelling language, and open-source software Protégé, an open source ontology editor, as modelling tool. Their reason is that owl can not only provide user with amounts of readable documentations, but also process the information of documentations and clearly express the words meaning and relations.

H. Wang et. al [31] defined geo-ontology as a set of concept, relation, attribute, axiom, and instances. They also defined the main process of building geo-ontology.

Li Bin et. al [24] proposed some important contents on constructing geo-ontology, such as principles and Logical language of constructing geo-ontology. They proposed to OWL (Ontology Web Language) as a logical language of constructing geo-ontology. Being a special and logical language describing network ontology, OWL has strong ability of expressing semantic knowledge as well as features of property on geo-ontology. The authors also mentioned that OWL can be used to express distinctly the geographical meanings of vocabulary entry in the glossary and relations between them. Such expression content is no other than geo-ontology. For the geo-ontology, it is much more important to figure geo-spatial property of geographical conceptions besides showing correlative property of attribute, especially geo-spatial relation. Geo-ontology can be effectively

constructed by using OWL combining with some tools. They used Protégé, which is a better tool of open source, to construct a geo-ontology based on spatial affairs.

3.4 Summary

Overall, the set of related works reviewed provide a snapshot of existing research works within the field of spatially aware web content retrieval as well as Amharic language web document retrieval. Having reviewed those documents, we learnt that none of the works provide an Amharic search engine with the ability of geographic information retrieval. The gaps we identified are summarized and listed in Table 3.1.

During the course of related works review we acquired a cumulative knowledge about the works in the area of our study. We observed that our work, to be fully functional, should have to address three areas of knowledge. Each of them is composed of various components and functionalities to provide services that are expected from them. There should also be a mechanism to integrate these three basic components of our research. In addition to that, we also understood what components a language specific search engine should contain and how its architecture looks like, fundamentals of geographical information retrieval, and the basic idea behind geo-ontology including geographical concepts, geographic features, geographic objects with their properties and the spatial relationship between each other. And we learnt what a geo-ontology is, linking the Amharic geo-ontology to the Amharic web content retrieval, and also the integration of the spatial and textual indexes.

We should also identify geographic features with the spatial relationships between each geographic object and populate geographic instances within the geo-ontology. The geo-ontology will be an input for the Amharic web content to footprint mapping, it will be a basic component for the spatial indexing. Our design will consider the points mentioned above, in order to deliver a functional model.

Table 3.1 : Summary of Related works

	Related Work	Amharic content Retrieval	Geographic Information Retrieval	Design of Geo-Ontology
1.	Design and Implementation of Amharic Search Engine (Tessema Mindaye, 2007, MSc thesis AAU) [2]	Unicode encoded Amharic web contents only	No	No
2.	Enhanced Design of Amharic Search Engine (Hassen Redwan, 2008 MSc Thesis AAU) [33]	Unicode + non-Unicode encoded web contents. Multithreading	No	No
3.	Searching the Web for Amharic Content (Tessema Mindaye et.al, 2010, Journal of Multimedia Processing and Technologies) [35]	Normalizing the query	No	No
4.	Amharic-English Bilingual Web Search Engine (Mequannint Munye, 2012, Conference Proceeding – on Management of Emergent Digital EcoSystem) [36]	Amharic-English query translation	No	No
5.	Geographical Information Retrieval (Christopher J. Bones et.al; 2008, International Journal of Geographical Information Science) [3]	No	Issues to be considered before GIR systems development are identified	No
6.	SPIRIT (Christopher J. Bones et.al, 2008, International Journal of Geographical Information Science) [38]	No	Spatially aware search engine	Geo-ontology implemented
7.	Maintaining Ontologies for Geographic Information Retrieval on the Web (Christopher J. Bones et.al, 2003, OTM Confederated International Conference) [37]	No	Basic query expressions are listed	Geo-ontology components are proposed

	Related Work	Amharic content Retrieval	Geographic Information Retrieval	Design of Geo-Ontology
8.	Core Geographical Concepts: Case Finish Geo-Ontology (Henriksson, Riikka; Kauppinen, Tomi; Hyvönen, Eero; 2008, The first International Workshop on Location and the Web) [22]	No	No	The content of geo-ontologies, and the approach of geo-ontology development process are examined
9.	Construction of Geo-Ontology Knowledge Base about Spatial Relations (Z. Xue and X. Jun, 2011, IEEE International Conference) [25]	No	No	Focuses on solving spatial relation query in different context, mainly performance
10.	Design of Geo-Ontologies based on Concept Lattice (H. Wang et.al, 2003, International Archives of Photogrammetry, Remote Sensing and Spatial Information Science) [31]	No	No	Main processes of geo-ontology building are defined
11.	Research on Geo-Ontology Construction based on Spatial Affairs (B. Li et.al, 2008, ICEODPA) [24]	No	No	Contents on constructing geo-ontology are proposed

Chapter Four

Spatially Aware Amharic Web Content Retrieval

In the previous chapter, we elaborate the different related works done before and we identified the gaps which resides in the existing works. In this work, we proposed a general model for the *Spatially Aware Amharic Web Content Retrieval* that incorporates features that are not considered in the previous Amharic web content retrieval systems. Figure 4.1 shows the architecture of the proposed model. This architecture encompasses components that identify Amharic language web documents and special analyzer that analyzes Amharic web documents.

In the following sections we will provide a summary of the high-level mechanism and communication processes between the different components of the *Spatially Aware Amharic Web Content Retrieval*. The functionality associated with each of these components and the interactions between the components required to support the functionality of each component, inputs, outputs and the issues associated with the task will be addressed. The components are explained in the same order they are executed in our program during run time.

4.1 System Architecture

The system architecture incorporates various components which are Amharic Search Engine, Geo Parsing, Geo Coding, Amharic Geo-Ontology, Spatio-textual Index , Request Management, Matching, and Ranking components. From this point onward, we will describe the functionality associated with each of these components and the interactions between the components. The modified components are shaded in the high-level abstraction as illustrated in Figure 4.1.

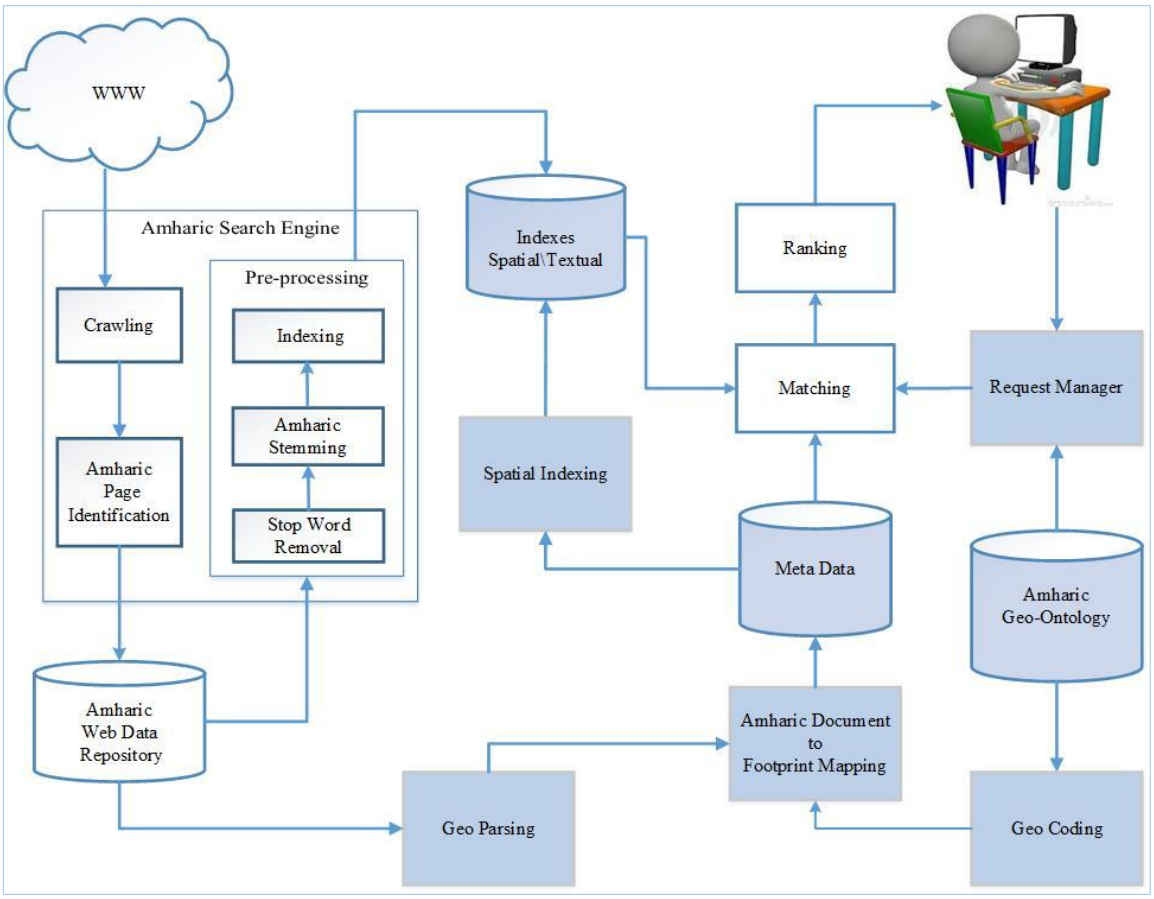

Figure 4.1: Architecture for Spatially-Aware Amharic web Content Retrieval

4.2 The Amharic Search Engine

The search engine component is responsible for the crawling, Amharic language identification, stop word removal, stemming, and indexing tasks. In this work Amharic search engine developed before will be adopted [2, 33].

The Web Crawler component starts visiting list of URLs, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the fetch list. The crawling component recursively visits these lists of URLs. The crawler copies and saves the information to the archives.

The Amharic Page Identification component filters the Amharic web documents and make them available for further processing. It discards web documents and links with other than Amharic languages.

The archive is known as the repository and is designed to store and manage the collection of web pages. The repository only stores HTML pages and these pages are stored as

distinct files. The repository stores the most recent version of the web page retrieved by the crawler.

The Indexing component collects, parses, and stores the parsed data to facilitate fast and accurate information retrieval. The output of the Crawler component, which is the Amharic Web Data Repository, is an input for the Indexing component. The Indexer component is responsible for creating the logical representation of the documents in the repository.

4.3 Amharic Web Document to Footprint Mapping

The metadata component is a storage that stores the Amharic web document to footprint mapping. It associates the crawled web documents with one or more footprints, such as longitude and latitude, that represent the geographical location to which individual documents relate. Once a place name is detected in a document the geo-ontology can be used to provide footprints to those place names. Algorithm 4.3 will be used to implement the mapping. The document to foot print mapping process involves two process, namely geo-coding and geo-parsing.

- **Geo-coding**

The geo-coding process involves disambiguating place names with multiple spatial references. Which is assigning spatial co-ordinates to candidate locations after extracting them from the web document. Once significant place names have been detected in a document the geographical ontology will be used to provide footprints that contribute to the geographical metadata associated with the document. The algorithm for the process of assigning footprint to web document is given in Algorithm 4.3.

- **Geo-parsing**

The geo-parsing process is responsible for identifying the presence of place names and other spatial qualifiers. Once a place name has been identified it is “grounded”, i.e. a map coordinate is allocated to it using the place name resources described in geo-parsing.

```
Input:
   Crawled Web document
   Geo-ontology
Variables
   Word : string
   PlaceName : string
   ListOfWords : List
   ListOfPlaceNames : List
   Latitude : Geography
   Longitude : Geography
   DocumentURL : String
   DocumentTitle : String
   DocumentFootprint : List
Begin
   //Read web documents
   If Crawled web document is not empty
       For each word in Crawled web document
           Add word to ListOfWords
           Next
     End For
   End If
   //Read Geo-ontology
   If Geo-ontology is not empty
     For each PlaceName in Geo-ontology
          Add PlaceName to ListOfPlaceNames
          Next
     End For
   End If
   //Identify the existence of place names in web document
   For each Word in ListOfWords
     For each PlaceName in ListOfPlaceNames
       If Word matches with PlaceName
         Extract(Latitude,Longitude)from Geo-ontology
         Extract(DocumentURL)from Crawled web document
         DocumentFootprint.Add(
                    PlaceName.Longitude,Word.DocumentURL,
                    Word.DocumentTitle)
       End If
     Next
     End For
     Next
   End For
     Return DocumentFootprint
End
```

Algorithm 4.1: Algorithm for assigning footprint to web document process

4.4 Request Management

The Request Management component interfaces with different components that allows the user to specify the subject of the query, a place name and/or a spatial relationship to the place name. It receives the user request, formats it, interprets it, requests the search execution, and requests ranking. After the user enters a query expression, the query will be parsed and disambiguated in order to identify the existence of spatial features, spatial relationships, and to discover the exact location phrases. The combination of the place name and spatial relationship are then used to determine a geographic query footprint. It uses the Amharic geo-ontology in order to identify the existence of any spatial terms or place names.

- **Amharic Geo-Ontology**

The Amharic geo-ontology component provides knowledge of places within the geographic coverage of the search engine. For each place, the geo-ontology maintains all of the names that a place is known by, the place types with which it can be categorized, the geographical footprints which indicate its spatial extent, and its topological relationships (such as part of and containing) with other places.

The geo-ontology is used both at the pre-processing and at run time processes. At pre-processing, the geo-ontology plays a key role in the process of geo-parsing of web documents and generation of spatial indexes. At run time process, the Amharic geo-ontology will help in the interpretation of user queries. When the search engine is interacting with the user, the geo-ontology will be used to recognize the presence of place names and perform place name disambiguation in the query. It will also be used in the expansion of spatial query and provide data to enable ranking of retrieved documents and extract metadata.

4.5 Spatial Indexing

The index component of our model stores spatio-textual index. In the textual indexing each term in the index is associated with a list of the documents that include the term. Spatio-textual indexing combines text indexing with spatial indexing of documents with respect to their document footprint.

Once document footprints, which is the geographic coverage, have been established. It will be very helpful for spatial indexing of the documents to facilitate fast access to documents pertaining to a given query footprint. Depending on the development of reliable techniques of geo-parsing, the process of spatial indexing will take place.

Here we adopted Hierarchical Tessellation algorithm [39]. It is based on an extensible indexing framework B+-Tree infrastructure and an adaptive quadtree-like multi-level grid and reuses the existing optimization framework of its relational query processor. During the process of spatial indexing, it uses hierarchical data structures techniques to sort data based on its spatial possession. Spatial indexing decomposes the space from which the data is drawn into regions which can be further decomposed into sub-regions. These regions are grouped hierarchies and represented as a tree structure which facilitates fast spatial operations such as search or intersect.

```
Tessellation Function(geometry g, cardinality card, out
CellQ)
 Begin
  Init OutputCellQueue using CellQ
  Init ProcessQueue
  Insert TopCell into ProcessQueue
  While ProcessQueue is not empty
  Begin
   CellToTraverse = ProcessQueue.Remove
   ComputeCoverage(g, CellToTraverse, out CoverageMasks)
   If (card < CoverageMask.Count + OutputCellQueue.Count)
    Output CellToTraverse
   Else
   Begin
    Output Interior (fully covered) Cells
    Output Leaf-Level Cells
    Output Touched Cells
    Insert non-leaf PartiallyCoveredCells to ProcessQueue
   End
 End
End
```

Algorithm 4.2: Hierarchical tessellation

4.6 Matching and Ranking

In the process of matching we will decide whether two terms are matched or not. Matching is important if we want to compare two terms which are the same but they were annotated differently. Here, the query terms are matched against the terms found in the spatio-textual index which is grabbed from the web documents.

The ranking process is responsible for the ranking of web documents that are identified by the Amharic search engine component according to their relevance to the query. It ranks them with respect to the non-spatial and spatial elements of the query. Once a geo-parsing process took place each document in the web document collection is represented by a footprint. The ranking component accesses the metadata, which is the web document to footprint mapping, to retrieve geometric footprints of places as well as associated data providing the geographical context of a place.

Chapter Five

Prototype and Evaluation

In this Chapter, we will describe the implementation of the proposed solution. The development environment, the tools, and languages we used for the development are briefly discussed .

5.1 The Development Environment

The development environment that we used are:

A laptop computer with a specification of:

- Intel core i3 processor with processing speed of 2.5GHz,
- An installed Memory of 4.0GB RAM,
- 500GB Hard disk, and
- Windows 10 Operating System

Tools and Languages

During the process of developing a prototype for our model we used the tools and programing languages listed below.

- **Protégé 5.2.0**

We used Protégé 5.2.0 for the development of geo-ontology. It is a free, open-source ontology editor and framework developed by Stanford University. It is written in Java; it provides a graphic user interface to define ontologies.

- **ArcGIS 10.1**

ArcGIS is a product of an international supplier of geographic information system (GIS) software company known as ESRI. It provides a contextual tool for mapping and spatial reasoning. We used ArcMap, a component of ArcGIS, to process the spatial data, and to extract place names with their location from a shape file.

- **Apache Nutch 1.4**

Apache Nutch is a well matured, production ready Web crawler. It is a highly extensible and scalable open source web crawler software product licensed by the Apache Software Foundation, that can be used to aggregate data from the web [40].

- **Apache Solr 4.3**

Solr, pronounced as "solar", is an open source enterprise search platform, written in Java, from the Apache Lucene project. Apache Solr is produced by Apache Software Foundation Its major features include full-text search, hit highlighting, faceted search, real-time indexing, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling. It is highly reliable, scalable and fault tolerant, providing distributed indexing, replication and load-balanced querying, automated failover and recovery, centralized configuration and more. Solr powers the search and navigation features of many of the world's largest internet sites [41].

- **.NET Framework 4.6**

The .NET Framework is a software framework developed by Microsoft that provides a comprehensive programming model for building all kinds of applications on Windows, web applications, desktop applications, or mobile applications. The .Net Framework provides the necessary compile time and run-time foundation to build and run any language that conforms to the Common Language Specification (CLS).

- **Entity Framework 6**

Entity Framework is Microsoft’s data access technology for applications. It is a set of technologies in ADO.NET that support the development of data-oriented software applications. Entity Framework enables working with data in the form of domain-specific objects and properties. With the Entity Framework, developers can work at a higher level of abstraction when they deal with data.

- **ASP.NET MVC 5**

The Model-View-Controller (MVC) is an architectural pattern which separates an application into three main components: the model, the view, and the controller. The framework provides an alternative to the ASP.NET Web Forms pattern for creating Web

applications. The ASP.NET MVC framework is a lightweight, presentation framework that is integrated with existing ASP.NET features.

- **Visual Studio 2017**

Visual studio is an IDE (Integrated Development Environment) from Microsoft. It is used to develop computer programs for Microsoft Windows, as well as web sites, web apps, web services and mobile apps. Visual Studio uses Microsoft software development platforms. We used Visual Studio 2017 for the development of the prototype system on top of .Net Framework.

- **Microsoft SQL Server 2008 R2**

We used this tool to store and retrieve data within our system. Microsoft SQL Server is a relational database management system developed by Microsoft. It is a software product for storing and retrieving data as requested by other software applications. It provides a mechanism for the storage of spatial data either as a geography or as a geometry data types.

- **Web Ontology Language (OWL)**

The OWL is a Semantic Web language designed and built up on W3C XML standard. It is a family of knowledge representation languages for authoring ontologies. It represents rich and complex knowledge about things, groups of things, and relations between things. We used this language for the implementation of the geo-ontology.

- **C# Programing Language**

C# is an object-oriented programing language that enables developers to build a variety of applications that run on the .NET Framework. C# can be used to create Windows client applications, XML Web services, distributed components, client-server applications, database applications. It is developed by Microsoft within its .NET initiative.

- **Google Map API V3**

Google Maps API is a set of methods and tools that allows to display maps on a web site or any web application. It is a JavaScript API that lets developers to customize maps with their own content and imagery for display on web pages and mobile devices. The Maps JavaScript API features four basic map types (roadmap, satellite, hybrid, and terrain) which can be modified using layers and styles, controls and events, and various services and libraries [42].

- **SolrNet**

SolrNet is an open source Apache Solr client for .NET. Since we built our model over .NET framework and SQL Server we prefer to use SolrNet which provides an abstraction layer using Solr in a .NET application.

5.2 Data Set

In order to populate individuals to our geo-ontology we need to have a data which consists of geographic features with their respective place names and location. So that to satisfy our need we extract a freely available OSM (OpenStreetMap) data from a bbike website [43] . OpenStreetMap is a map of the world created to be used freely under an open license [44]. We downloaded a shape file for Addis Ababa city from the web site mentioned above with a provided download link. After the shape file is downloaded, it is processed with ArcGIS software. During the processing of the shape file we loaded the file on ArcMap, which is a component of ArcGIS, and we added the fields necessary for our work, like coordinates of geographic objects. Then we calculated geometry for geographic object to get the x-coordinate and y-coordinate values of each and every point on the map. Figure 5.1 shows geographic objects plotted under the boundaries of Addis Ababa city and one of the attribute tables with calculated geometry. In this work, we populated a total of 275 individuals consisting of 275 known places located in Addis Ababa city and 38 cities all over Ethiopia.

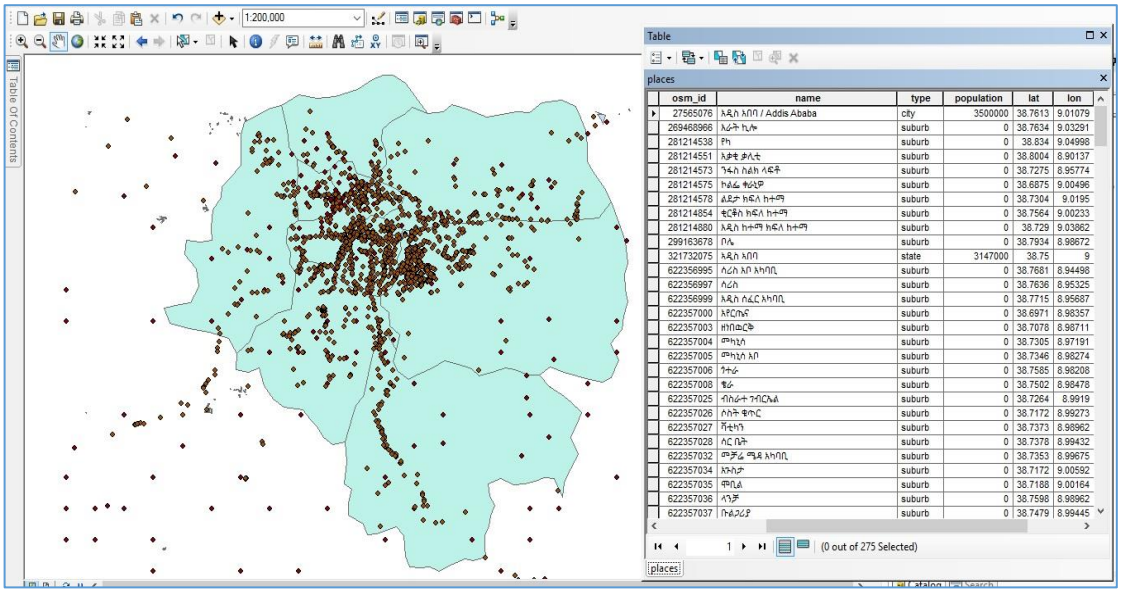

Figure 5.1 : Processing the shape file in ArcMap software

In order to collect the Amharic web documents while testing our prototype we used the websites listed in Table 5.1.

Table 5.1 : List of Websites visited by the crawler

Name of the Website	URL
Addis Admass Newspaper	http://www.addisadmassnews.com/
Ethiopian Reporter Newspaper	https://www.ethiopianreporter.com
Ethiopian Government Portal	http://www.ethiopia.gov.et/
Fana Broadcasting Corporate	http://www.fanabc.com
Ethiopian News Agency	http://www.ena.gov.et
Mereja.com	http://www.mereja.com/
Ethiopian Press Agency	http://www.ethpress.gov.et
Wikipedia	https://am.wikipedia.org/

5.3 Prototype

5.3.1 The Crawling Process

At the time of experiment, we crawled the seed URLs provided in Table 5.1 using a web Apache Nutch crawler. We executed the following command on Cygwin command-line interface: *$ bin/nutch crawl urls -dir crawl/ -depth 5 -topN 1000*

Figure 5.2 illustrates the execution of crawling with Nutch's crawl command.

```
Admin@think /cygdrive/c/MyThesis/nutch-1.4/runtime/local
$ bin/nutch crawl urls -dir crawl/ -depth 5 -topN 1000
crawl started in: crawl
rootUrlDir = urls
threads = 10
depth = 5
solrUrl=null
topN = 1000
Injector: starting at 2018-05-28 09:36:00
Injector: crawlDb: crawl/crawldb
Injector: urlDir: urls
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2018-05-28 09:36:07, elapsed: 00:00:07
Generator: starting at 2018-05-28 09:36:07
Generator: Selecting best-scoring urls due for fetch.
Generator: filtering: true
Generator: normalizing: true
Generator: topN: 1000
Generator: jobtracker is 'local', generating exactly one partition.
Generator: Partitioning selected urls for politeness.
Generator: segment: crawl/segments/20180528093611
Generator: finished at 2018-05-28 09:36:13, elapsed: 00:00:05
Fetcher: Your 'http.agent.name' value should be listed first in 'http.robots.agents' property.
Fetcher: starting at 2018-05-28 09:36:13
Fetcher: segment: crawl/segments/20180528093611
Using queue mode : byHost
Fetcher: threads: 10
Fetcher: time-out divisor: 2
Using queue mode : byHost
Using queue mode : byHost
```

Figure 5.2: Crawling with Nutch Screenshot

5.3.2 Textual Indexing

Using data from all possible sources (CrawlDB, LinkDB and Segments), the Indexer creates an index and saves it within the Solr directory. For indexing, the popular Lucene library [45] is used as shown in Figure 5.3.

```
Admin@think /cygdrive/c/MyThesis/nutch-1.4/runtime/local
$ bin/nutch solrindex http://127.0.0.1:8983/solr/ crawl/crawldb -linkdb crawl/linkdb crawl/segments/*
cygpath: can't convert empty path
SolrIndexer: starting at 2018-05-28 12:01:34
Adding 1000 documents
Adding 1000 documents
Adding 1000 documents
Adding 618 documents
SolrIndexer: finished at 2018-05-28 12:02:15, elapsed: 00:00:40
```

Figure 5.3: Textual Indexing with Apache Solr

5.3.3 Spatial Indexing

For the implementation of spatial indexing we used MS-SQL Server since it supports spatial data and spatial indexes. A spatial index is a type of extended index that allows indexing a spatial column. A spatial column is a table column that contains data of a spatial data type, such as geometry or geography [27]. SQL Server allows creating an index even before there is data in the table [46]. Table 5.2 presents a sample SQL server syntax that shows how we created spatial index on our database "GeoOntology". The full SQL Server syntax is attached on Annex B.

SQL Server 2008 Spatial Indexing component is built using the extensibility hooks available in the relational query processing engine. SQL Server's spatial indexing method adopts the well-known strategy of *hierarchical decomposition of space* to implement the index. The spatial indexing system is designed to be able to handle objects of varying shapes and sizes efficiently. Non-uniform data distribution patterns are the weak points of fixed decomposition schemes. In order to address this, the system allows for multiple localized indexes on a spatial column instance [39].

The system decomposes the indexed space into an ordered collection of axes-aligned cells using a four-level hierarchy of grids. The cells are disjoint except when they are part of ancestor descendant relationships. The cells of all levels form an ordered domain, in which all of a cell's descendants immediately follow it. The linear ordering is important and necessary because the underlying storage and access methods make use of existing one-dimensional B+-Tree storage keyed on hierarchical cell identifiers. The linear ordering used provides for spatial locality in the index [39].

Table 5.2:Partial Syntax for Spatial Indexing

```
CREATE SPATIAL INDEX SpatialIndex_PointObject_Centroid
   ON PointObject (GeoOntology)
   USING GEOGRAPHY_GRID
   WITH (
    GRIDS = (LEVEL_1= MEDIUM, LEVEL_2= LOW,
             LEVEL_3= MEDIUM, LEVEL_4 = HIGH),
             CELLS_PER_OBJECT = 64
   );
```

The query creates a spatial index, “SpatialIndex_PointObject_Centroid”, on the “Centroid” column of the PointObject table. The query specifies GEOGRAPHY_GRID as the tessellation scheme. GEOGRAPHY_GRID specifies the geography grid tessellation scheme. It can be specified only on a column of the geography data type [46]. It also specifies different grid densities on different levels and 64 cells per object.

The cells-per-object limit defines the maximum number of cells that tessellation can count per object. The extent of tessellation of each object depends primarily on the cells-per-object limit of the spatial index.

5.3.4 Geo-Processing

As a result of the crawling process, a total of 3,618 web documents are crawled and each document is parsed, indexed and stored in a database accordingly. During the geo-processing phase, we perform the geo-parsing task and look for the existence of the geographic qualifiers in those documents. Following the geo-parsing process, web pages identified as they contain place names are grounded, in other words they become mapped to their corresponding footprint.

5.3.5 The Amharic Geo-Ontology Development Process

During our literature review we observed that there is not a standard method for building geo-ontology. Different works gave different methods about the building of geo-ontology. We follow the methods presented in [28,41]. During the development of the ontology the main processes we follow are listed below.

Identification and Scoping the Geo-Ontology

Table 5.3 : Sample List of Concepts identified

Point Objects	Linear Objects
ቤተ ክርስቲያን	ወንዝ
መስጊድ	መንገድ
የጤና ተቋማት	ሐዲድ
ሆስፒታል	
ጤና ጣቢያ	
አውሮፕላን ማረፊያ	**Areal Objects**
ኤምባሲ	አገር
ባንክ	ክልል
መድን	ከተማ
ቲያትርና ሲኒማ ቤት	ክፍለ ከተማ
ሕንጻ	ወረዳ
ትምህርት ቤት	ቀበሌ
ኮሌጅ	ሐይቅ
ዩኒቨርሲቲ	ደን
መናኸሪያ	
ሆቴሎች	

During this activity we define the scope of Geo-ontology as concepts which belong to the land use information limited to the urban area. After defining the scope of the geo-ontology, we identified and classified the geographical concepts with their attributes and relationships with the help of a GIS expert. The main concepts are listed in Table 5.3.

Building the Geo-ontology

This process involves explicitly representing the acquired knowledge in the previous step in a formal language. This activity is broken down into four steps.

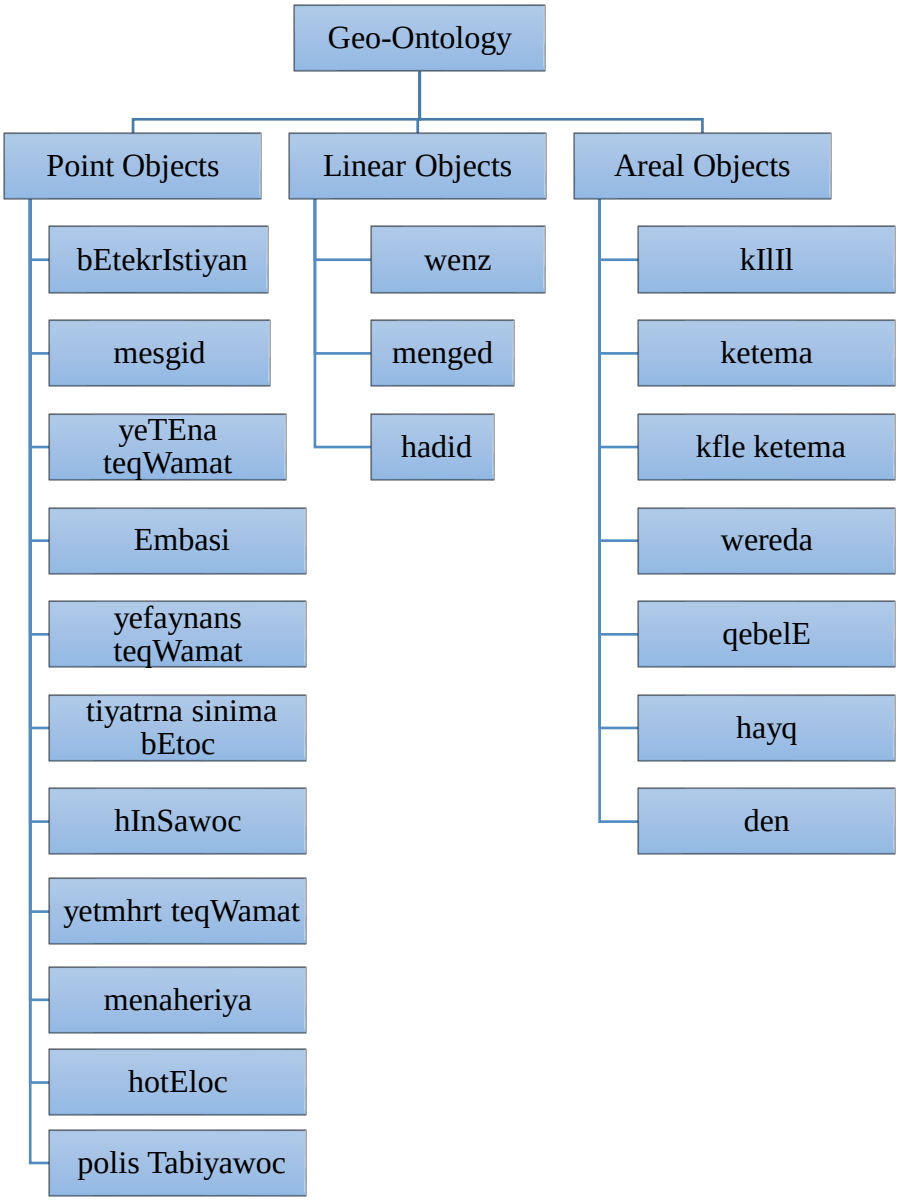

Figure 5.4 : Class hierarchy of Sample Concepts

The Construction of Classes

To provide a service as a basis for the semantic integration of different knowledge bases containing geographical data, we propose a geo-ontology composed of three high level concepts: Point Objects, Linear Objects, and Areal Objects, as shown in the class hierarchy Figure 5.4. The first represents objects or entities which can be represented as a single point in the coordinate system. The second one represents geographical objects like road, river, and railways. These objects are represented with a minimum of two points in the

coordinate system. The third contains geographical features which are represented by their boundaries.

The Construction of Property

After finishing the basic classes, we use object properties model and data properties model to build the properties of classes. In Protégé, defining the "Domain" and "Range" of the properties can express constrain of concrete concept classes. The property and the class that can use the property are connected by domain. And the scope of the property is determined by range. Table 5.4 lists part of the domain and range of the properties in the knowledge base.

Table 5.4: List of Object properties and Data properties with their respective domain and range

Property		Domain	Range
Object Properties	አቅራቢያ (Near)	መንገድ፣ ወንዝ፣ ሃዲድ (Linear Objects)	ሕንጻ፣ ሆቴል፣ ትምህርት ቤት (All Point objects)
	ክፍል (Part of)	ሕንጻ፣ ሆቴል፣ ትምህርት ቤት …. (Point objects)	ከተማ፣ ክ/ ከተማ …. (Areal Object)
	የተደረበ (Overlaps)	ከተማ፣ ደን፣ ሐይቅ …. (all Areal Objects)	ከተማ፣ ደን፣ ሐይቅ …. (all Areal Objects)
	አጠገብ (Adjacent to)	ከተማ፣ ደን፣ ሐይቅ …. (all Areal Objects)	ከተማ፣ ደን፣ ሐይቅ …. (all Areal Objects)
	ይዟል (Contains)	ከተማ፣ ክ/ከተማ፣…. (all Areal Objects)	ሕንጻ፣ ሆቴል፣ትምህርት ቤት (all Point objects)
Data Properties	መለያ (Id)	All Classes	Integer
	ስም (Name)	All Classes	string
	ማዕከላዊ ነጥብ (Centroid)	ሕንጻ፣ ሆቴል፣ትምህርት ቤት …. (Point objects)	float
	ክልል (Area)	ከተማ፣ ደን፣ ሐይቅ …. (All Areal Objects)	float
	ርዝመት (Length)	መንገድ፣ ወንዝ፣ ሃዲድ (Linear Objects)	float

The Construction of Individuals

According to the previous construct classes, properties and the permissible value of the classes and properties, in the Protégé using the Individuals model to add the specific individuals. The first thing we did in this process is collecting a spatial data which is the main source to populate individuals in the geo-ontology.

5.3.6 Implementation of the Geo-ontology

During the process of geo-ontology development, OWL (Web Ontology Language) is selected as an ontology development language. We used OWL due to its ability to reason about classes and individuals to the degree permitted by the formal semantics [47]. The geo-ontology is implemented using an ontology editor tool called Protégé. We used the class model in Protégé to create classes and subclasses it is shown with the ontology graph Figure 5.5.

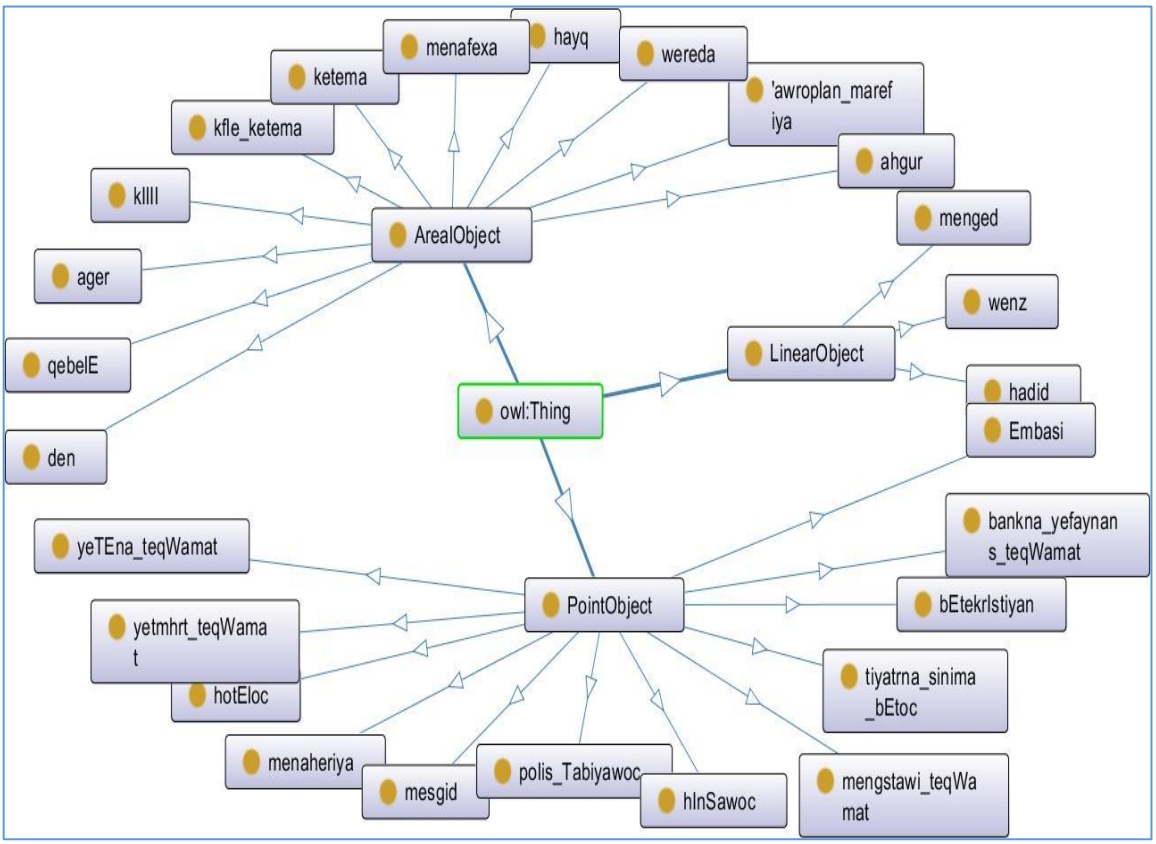

Figure 5.5: Amharic geo-ontology graph implemented with protégé

Since the tool, i.e. protégé, doesn't support writing with Amharic characters while creating ontologies, the concept names are written in Roman form. We used a tool developed by Michael Gasser named HornMorpho [48]. It is a Python program that analyzes Amharic, Oromifa, and Tigrinya words into their constituent morphemes (meaningful parts) and generates words, given a root or stem and a representation of the word's grammatical structure. For instance, the concept "የጤና ተቋማት" is written as "? yeTEna teqWamat". How the conversion is done using HornMorpho is shown in Figure 5.5. This tool takes a word written in geez alphabets as an input and converts it to Roman form and vice versa.

After Romanizing process is completed, we have organized concepts of the geo-ontology hierarchically as shown in Figure 5.6. Using these concepts, we implemented the Amharic geo-ontology, which is one of the core components of this work, using protégé version 5.2.0. The ontology graph developed with this tool is presented in Figure 5.5. Sample source code for the generated geo-ontology is presented in Annex A.

Evaluation of the Geo-ontology

Ontology evaluation is the process of validating the developed ontology by assessing its capability to represent the domain. The constructed geo-ontology is evaluated to make a technical judgement of the ontologies. In this research the capability of the ontology in representing facts correctly is assessed by domain experts, GIS experts, as per the defined geographic concepts and their relationships during the scoping of the geo-ontology activity.

5.3.7 Mapping the Geo-ontology to a Relational Database

Once the implementation of the knowledge base is finalized, we mapped the geo-ontology into a relational database as shown in Table 5.5.

The need for this mapping is to store the ontology data and execute queries on that data. It is important to fully realize the power of ontologies and to enable efficient and flexible information gathering, persistent storage of geo-ontology and its subsequent retrieval. Due to the popularity of SQL and the efficiency of executing these queries, a complimentary approach is to use RDB (Relational Database) [49].

During the mapping between an OWL (Web Ontology Language) ontology and an RDB, an OWL class corresponds to an RDB table, an OWL data property corresponds to a table field, and an OWL object property corresponds to a foreign key [49] [50].

Table 5.5 shows how OWL classes and sub classes are mapped to a relational database table. Data from ontology classes and database tables is listed on the table.

Table 5.5: OWL Class to Database Tables Mapping

Owl Class	Table Name	Filtering Expression
PointObject	*PointObject*	
ቤተ ክርስቲያን	PointObject	FeatureType = "ቤተ ክርስቲያን"
መስጊድ	PointObject	FeatureType = "መስጊድ"
ሆስፒታል	PointObject	FeatureType = "ሆስፒታል"
ጤና ጣቢያ	PointObject	FeatureType = "ጤና ጣቢያ"
አውሮፕላን ማረፊያ	PointObject	FeatureType = "አውሮፕላን ማረፊያ"
ኤምባሲ	PointObject	FeatureType = "ኤምባሲ"
ባንክ	PointObject	FeatureType = "ባንክ"
መድን	PointObject	FeatureType = "መድን "
ቲያትርና ሲኒማ ቤቶች	PointObject	FeatureType = "ቲያትርና ሲኒማ ቤቶች"
ሕንጻዎች	PointObject	FeatureType = "ሕንጻዎች"
ትምህርት ቤት	PointObject	FeatureType = "ትምህርት ቤት"
ኮሌጅ	PointObject	FeatureType = "ኮሌጅ "
ዩኒቨርሲቲ	PointObject	FeatureType = "ዩኒቨርሲቲ"
መናኸሪያ	PointObject	FeatureType = "መናኸሪያ"
ሆቴሎች	PointObject	FeatureType = "ሆቴሎች"
ፖሊስ ጣቢያዎች	PointObject	FeatureType = "ፖሊስ ጣቢያዎች"
LinearObject	*LinearObject*	
መንገድ	LinearObject	FeatureType = "መንገድ"
ወንዝ	LinearObject	FeatureType = "ወንዝ"
ሃዲድ	LinearObject	FeatureType = "ሃዲድ"
ArealObject	*ArealObject*	
ክልል	ArealObject	FeatureType = "ክልል"
ከተማ	ArealObject	FeatureType = "ከተማ"
ክፍለ ከተማ	ArealObject	FeatureType = "ክፍለ ከተማ"
ወረዳ	ArealObject	FeatureType = "ወረዳ"
ቀበሌ	ArealObject	FeatureType = "ቀበሌ"
ሐይቅ	ArealObject	FeatureType = "ሐይቅ"
ደን	ArealObject	FeatureType = "ደን"

As presented in Table 5.6 the data type properties of OWL will be mapped to a database table column of an RDB.

The object property of our geo-ontology, which represents the spatial relationships, mapping to the database table is shown in Table 5.7. It presents how the domain and range objects reference to each other.

Table 5.6 : OWL datatype property class mappings to database table column

Data Type Property	Table Name	Column Expression
መለያ	All Tables	Id
የቦታ ስም	Place Name	Name
ማዕከላዊ ነጥብ	PointObject	Centroid
ርዝመት	LinearObject	Length
ስፋት	ArealObject	Area

Table 5.7 : OWL object property mappings to database tables pairs

Object Property	Table Name (Domain)	Table Name (Range)	Primary Key (Source Column)	Foreign Key (Target Column)
Near	LinearObject	PointObject	LId	LId
Part of	PointObject	ArealObject	PId	PId
Overlaps	ArealObject	ArealObject	AId	AId
Adjacent to	ArealObject	ArealObject	AId	AId
Contains	ArealObject	PointObject	AId	PId

Based on the geo-ontology to relational database mapping results we designed the database as shown in Figure 5.6.

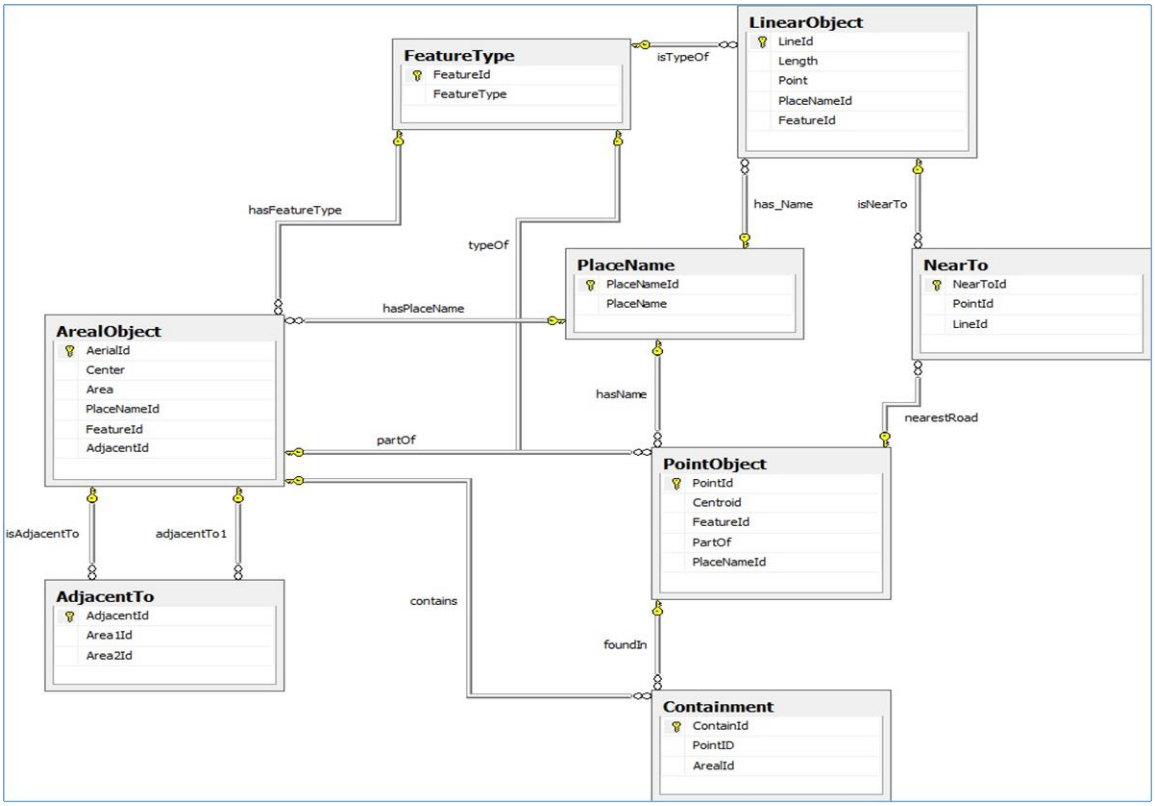

Figure 5.6 : Database diagram

Screenshot of the Prototype Application

The illustration for the user interface of our prototype system is shown in Figure 5.7. While the user submits a query, the result will be presented in two columns the left column is dedicated for the spatial results and the list of web documents will be displayed on the right side of the map.

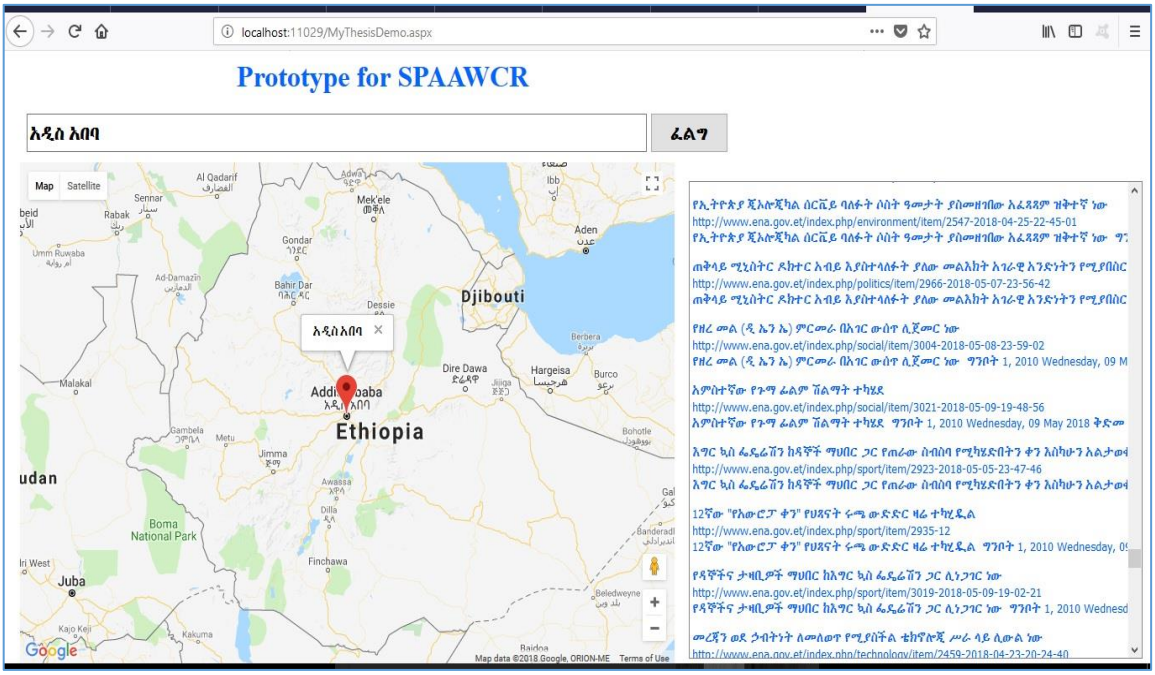

Figure 5.7: Screenshot of our Prototype with sample query "አዲስ አበባ"

5.3.8 Ranking

The task of spatial relevance ranking is based on measures of distance between the query footprint and the document footprint. The relevance will be evaluated by the spatial relationships between documents and queries in geographical space [51]. In order to perform relevance ranking we used a distance function. SQL server provides a functionality that help us when we are performing ranking based on distance between the query and the document footprints. Spatial indexes support methods which are provided by SQL Server and we used for relevance ranking, we used two methods, which are STContains and STDistance methods [27] . These methods we used to rank are dependent on the spatial relationship.

For Containment relationships we use STContains method: this method returns 1 if a geometry instance completely contains another geometry instance. Returns 0 if it does not. Coordinates are checked for containment.

For relationships such as near-to/near-by we used STDistance method: this method returns the shortest distance between a point in a geography instance and a point in another geography instance.

5.4 Evaluation

The evaluation for our work is conducted in comparison with the three well known web search engines Google, Yahoo, and Bing.

Since the ultimate goal of this work is to design a model that adds spatial awareness for Amharic web content retrieval, we evaluated the developed prototype application:

- For query disambiguation: checking the existence of a place name or spatial relationship within the query string
- Returning a result set containing web documents with their respective footprints
- Plotting each footprint on a map based on the coordinate (longitude and latitude) values extracted from the document to footprint map.

For the sake of evaluation, we formulated three types of queries that our model should handle. These query expressions consist of a reference to:

- Place Name only,
- A Place Name with a Spatial Relationship to a Geographic Feature Type, and
- A Geographic Feature Type only

A Place Name is an actual name for a geographic object, such as አዲስ አበባ ዩኒቨርሲቲ. A Feature Type represents the categorization of geographical objects, e.g. ከተማ፣ ዩኒቨርሲቲ and so on. A spatial relationship is an instance of a relationship between any types of a geographic objects, such as a containment relationship የሚገኙ.

- **Evaluation with a Query containing only a Place Name**

To evaluate that we submit “አዲስ አበባ ዩኒቨርሲቲ”, which is a place name without any spatial relationship, as a query string for Google, Yahoo, Bing, and as well to our Model. For a given query string “አዲስ አበባ ዩኒቨርሲቲ” Google displays a list of web URLs which are related to the user query, as shown in Figure 5.8.

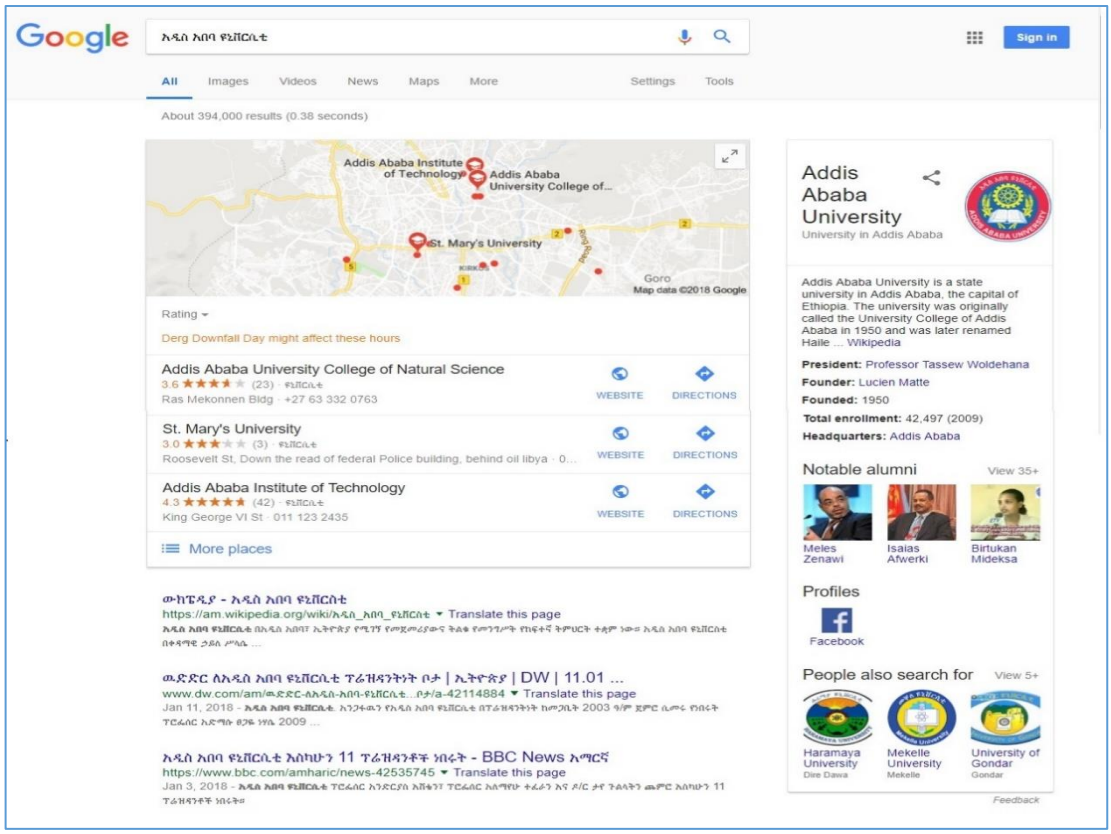

Figure 5.8: Google's search result for "አዲስ አበባ ዩኒቨርሲቲ"

Figure 5.9 details that Google plotted different universities and colleges located in Addis Ababa on a map with their respective locations for the query አዲስ አበባ ዩኒቨርሲቲ.

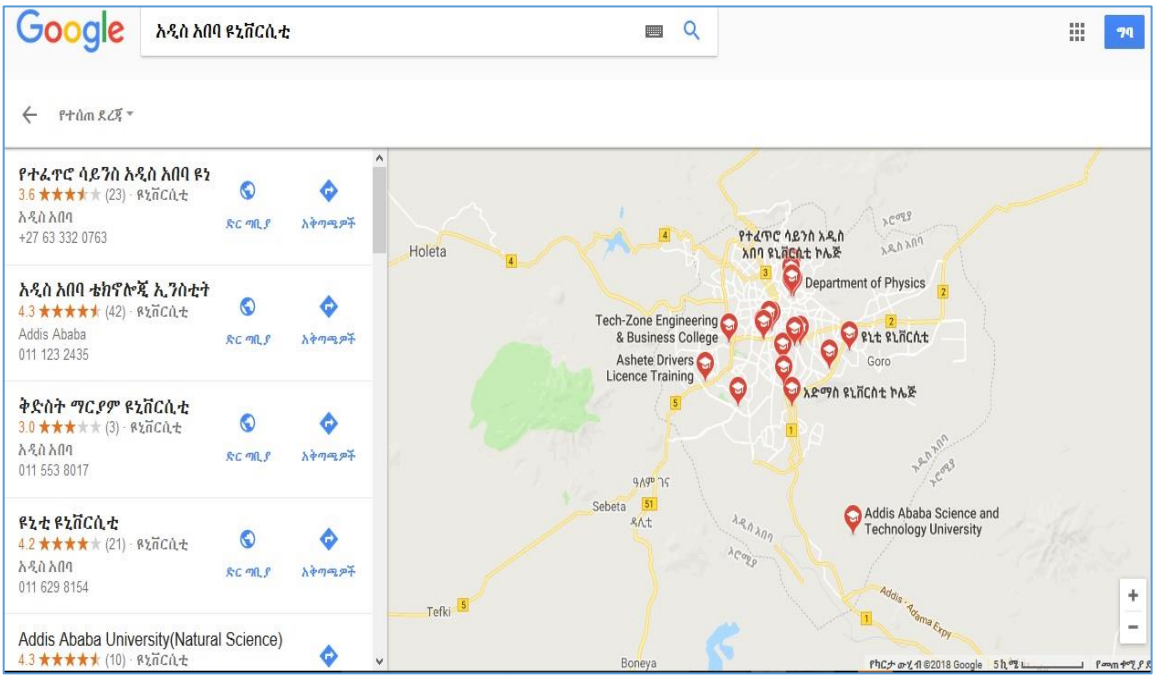

Figure 5.9: Google map's plot for "አዲስ አበባ ዩኒቨርሲቲ"

As illustrated in Figure 5.10, and Figure 5.11, Yahoo and Bing return a list of web sites which are related to the user query "አዲስ አበባ ዩኒቨርሲቲ" in one or another way. Both of the search engines didn't provide any information about the locations or spatial features of the query string "አዲስ አበባ ዩኒቨርሲቲ".

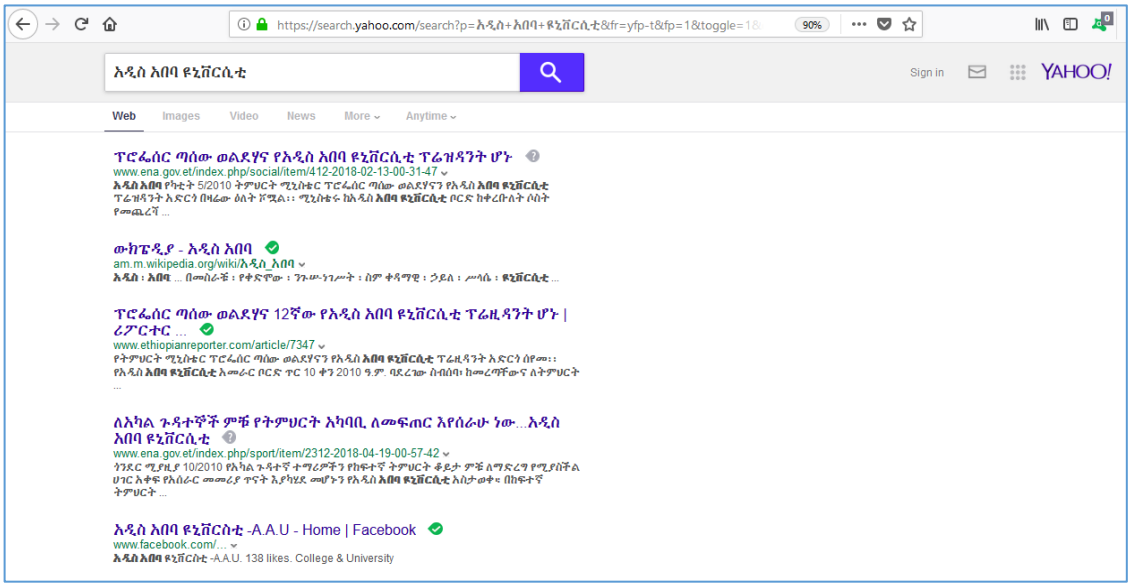

Figure 5.10: Yahoo's search result "አዲስ አበባ ዩኒቨርሲቲ"

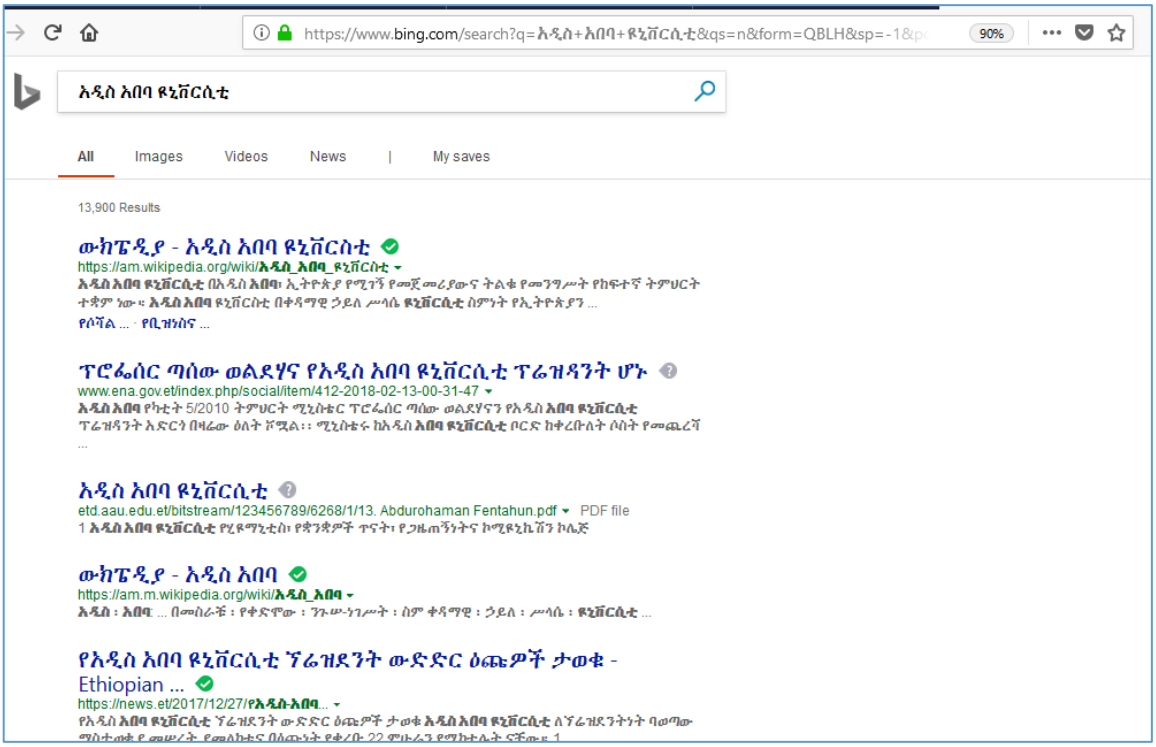

Figure 5.11 : Bing's search result for "አዲስ አበባ ዩኒቨርሲቲ"

For the same user query "**አዲስ አበባ ዩኒቨርሲቲ**" our model displays the result as shown below in Figure 5.12.

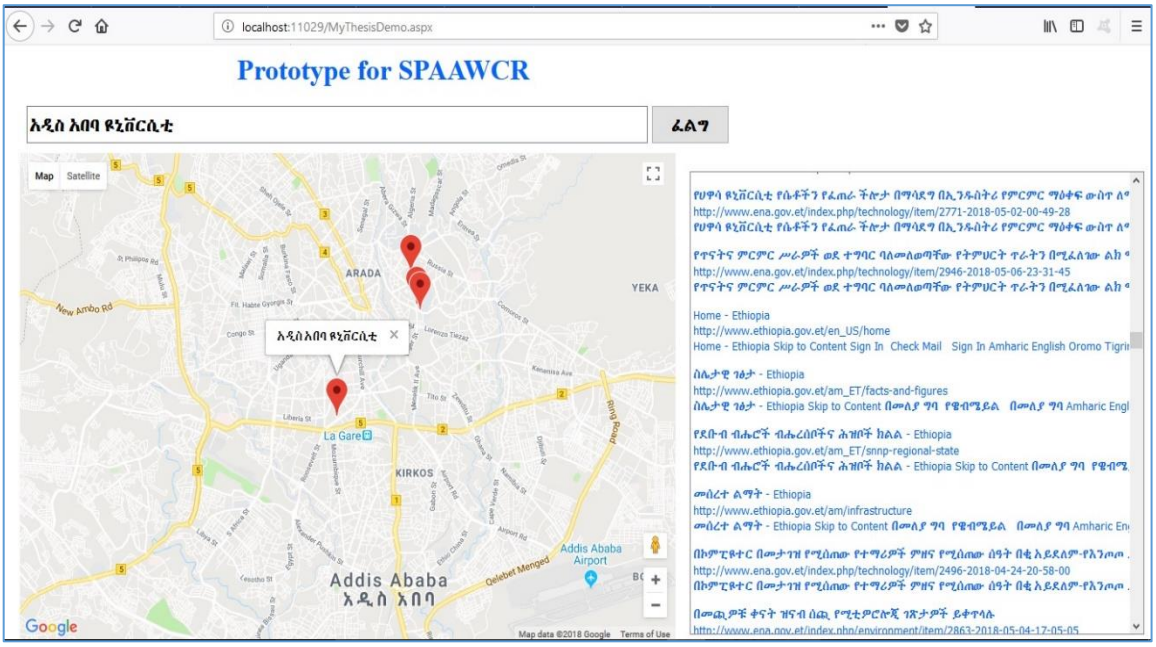

Figure 5.12: Our Prototype's search result for "አዲስ አበባ ዩኒቨርሲቲ"

- **Evaluation with a Query containing a Place Name with a Spatial Relationship to a geographic Feature Type**

Looking the result google displayed in Figure 5.8 and Figure 5.9, someone may say that Google is doing well with spatially aware Amharic web content retrieval. But it is not. To evaluate that we provide google with another search string. We tested Google, Yahoo, and Bing whether they consider search string as a geographic instance or not, in other words their spatial awareness for Amharic content retrieval. We submit a query string which consists not only the place name but also a spatial relationship "በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች". This query is formulated as < Place Name, Spatial Relationship, Geographic Feature Type >, which consists of a place name "አዲስ አበባ", a spatial relationship "የሚገኙ", and a geographic feature type "ዩኒቨርሲቲዎች".

Figure 5.13 : Google's search result for "በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች"

If Google was able to understand the spatial relationship for Amharic language, the result set should be different. It would plot universities which are located in Addis Ababa on a map, as it did in the previous query. Instead it took each word as a search key word and displays a list of web document URLs for each word as shown in Figures 5.13. The same is true for the other two search engines, which are Yahoo and Bing. It is illustrated in Figure 5.14, and Figure 5.15 what Yahoo and Bing respond to the query "**በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች**". None of them give a clue about the geographic features or feature types as Google did for the first query string.

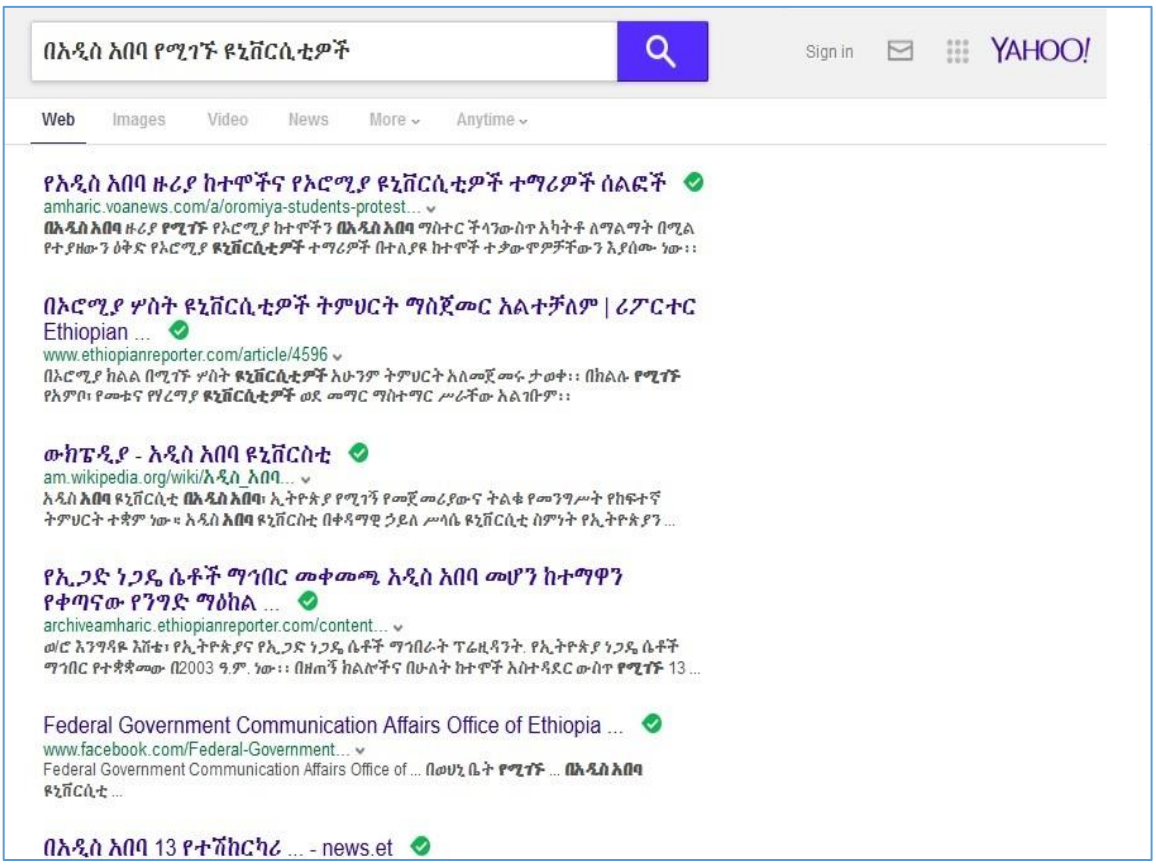

Figure 5.14:Yahoo's search result for "በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች"

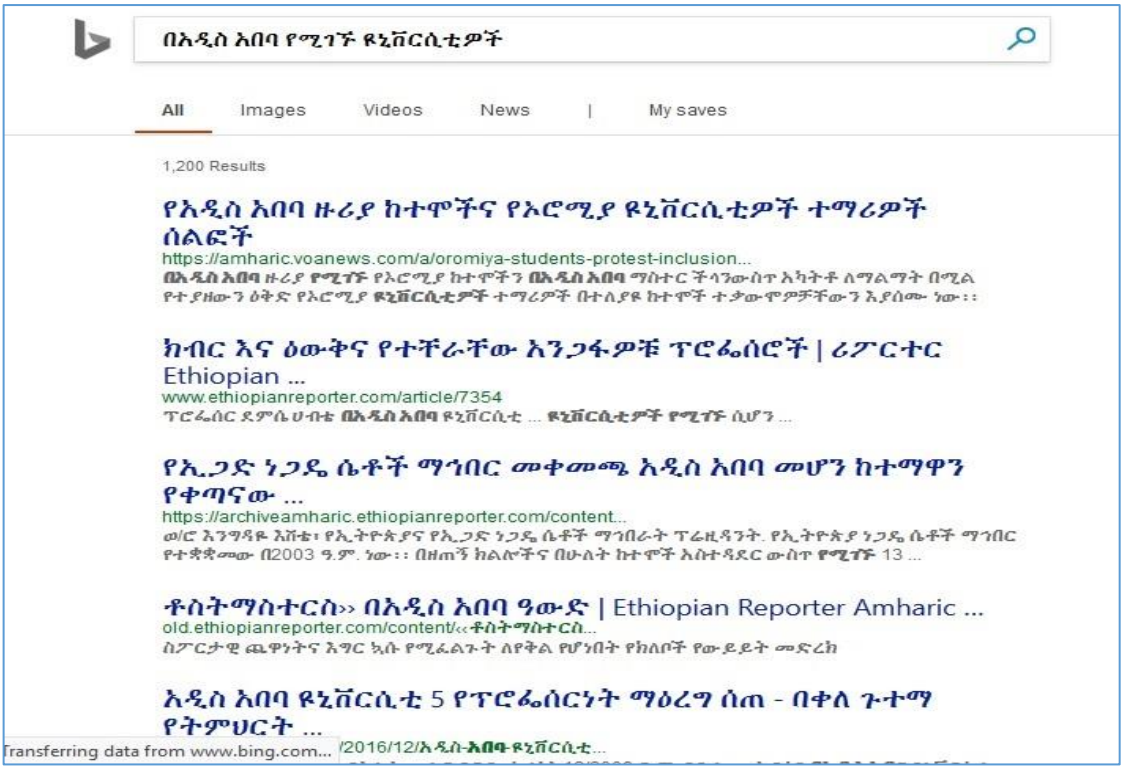

Figure 5.15: Bing's search result for "በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች"

Finally, we evaluated our prototype application with the same query formulation <Place Name, Spatial Relationship, Geographic Feature Type> as we did before. And we provided the string "በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች" as a query string to our prototype application. It returns list of web URLs that are related to the requested "በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች" and plot their location on Google map, as shown in Figure 5.16.

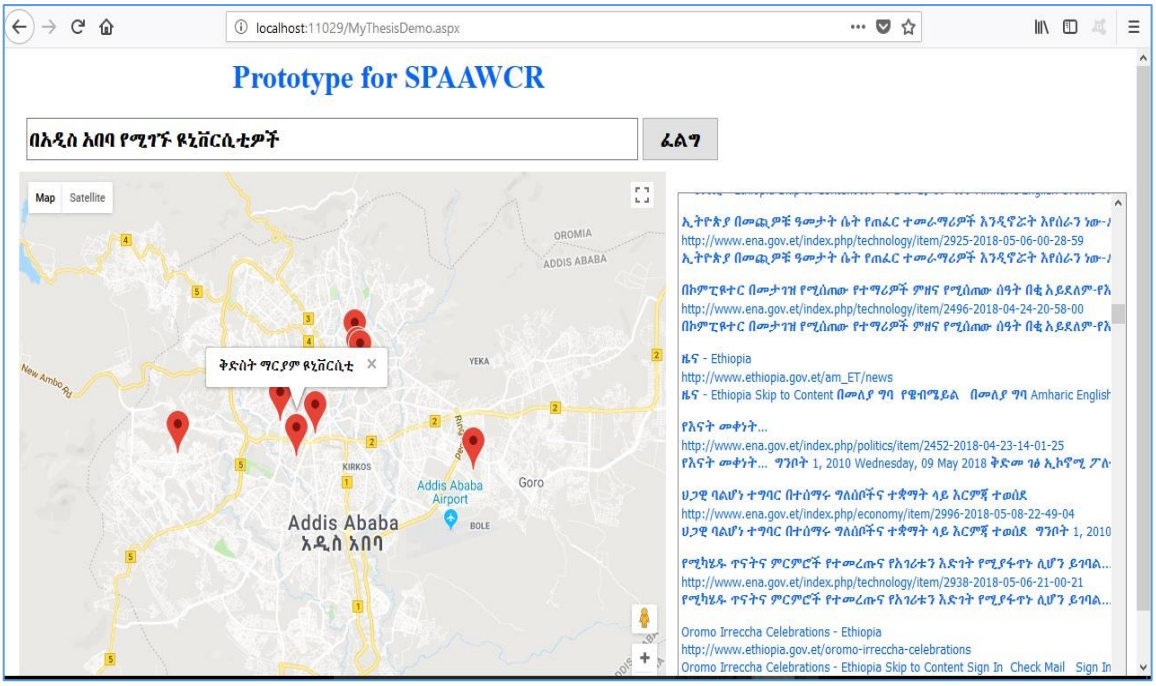

Figure 5.16: Result of our prototype for query "በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች"

- **Evaluation with a Query containing only Feature Type**

As we have been doing in the previous two scenarios, we submit our query, which is formulated with the geographic feature type only, to the three search engines Google, Yahoo, and Bing. We provide a query string "የኢትዮጵያ ከተሞች" for them and to our prototype application as well. The results each of the three search engines deliver are illustrated below.

Figure 5.17 shows what Google displays for the query string "የኢትዮጵያ ከተሞች". Yahoo's and Bing's result sets for the same user query are illustrated in Figure 5.18 and Figure 5.19 respectively.

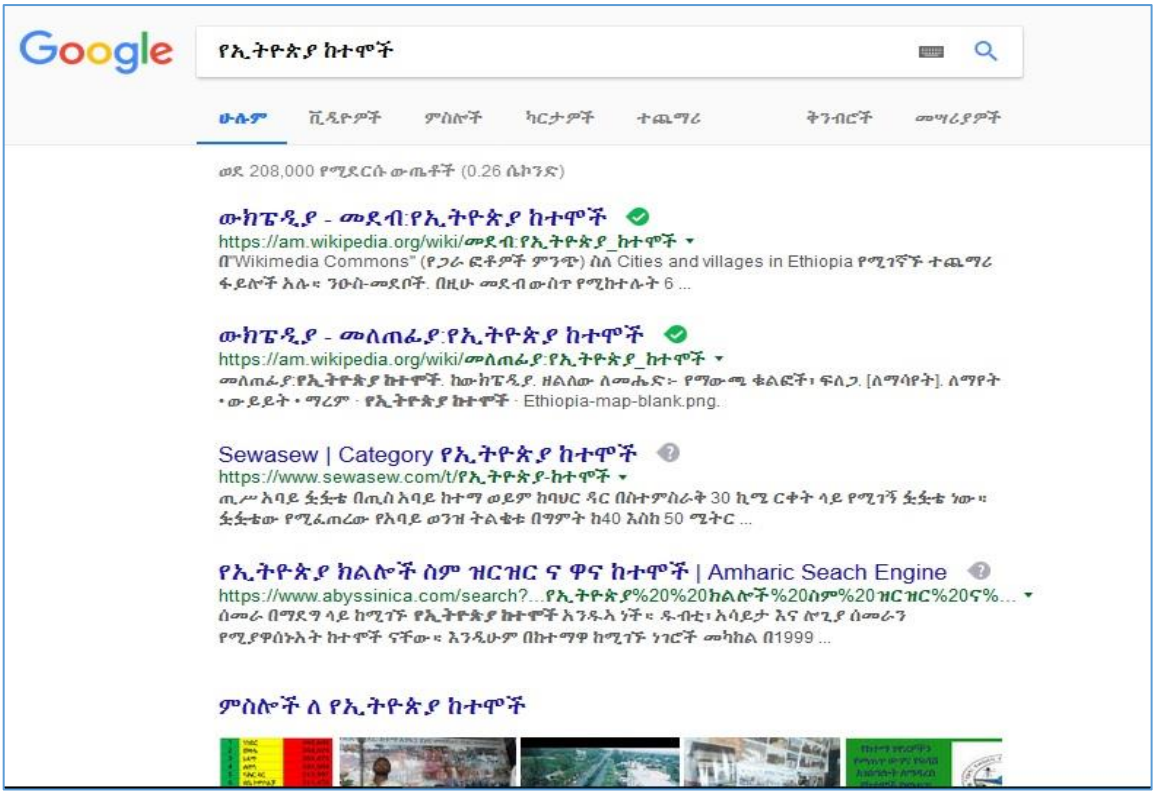

Figure 5.17: Google's result for query "የኢትዮጵያ ከተሞች"

Figure 5.18: Yahoo's result for query "የኢትዮጵያ ከተሞች"

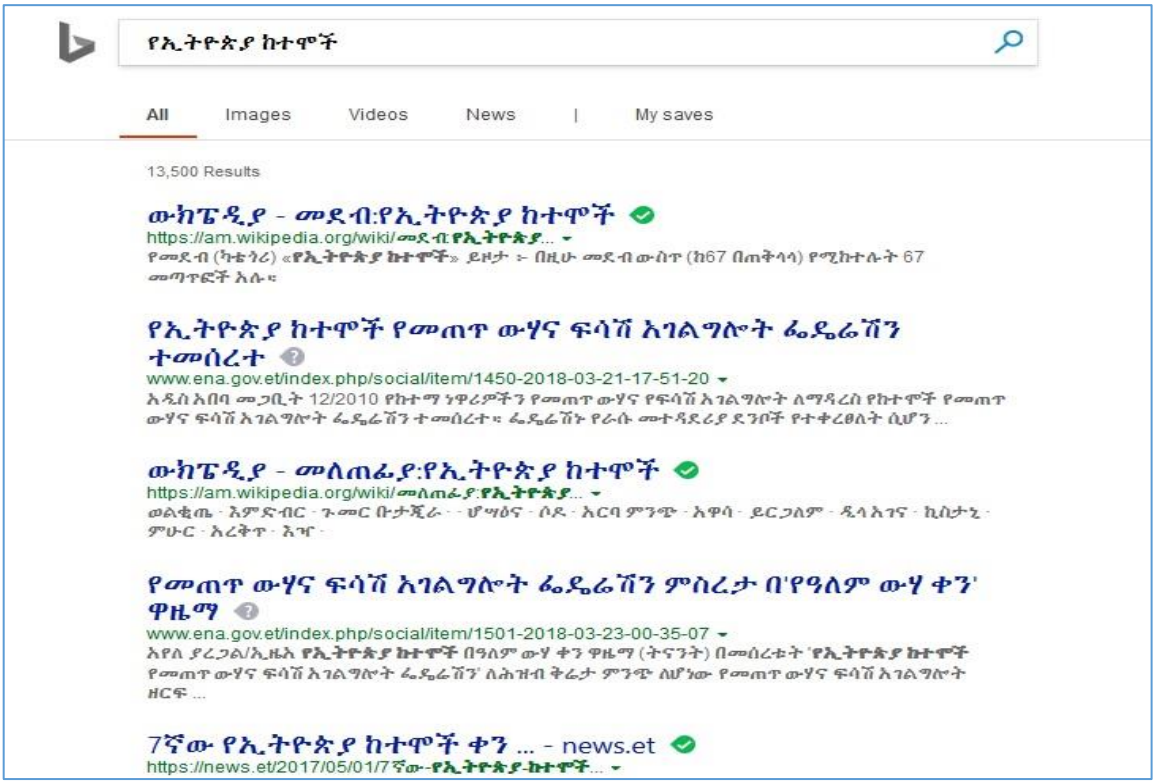

Figure 5.19:Bing's result for query "የኢትዮጵያ ከተሞች"

Then after the same query "**የኢትዮጵያ ከተሞች**" was submitted to our prototype application and what it returns is shown in Figure 5.20.

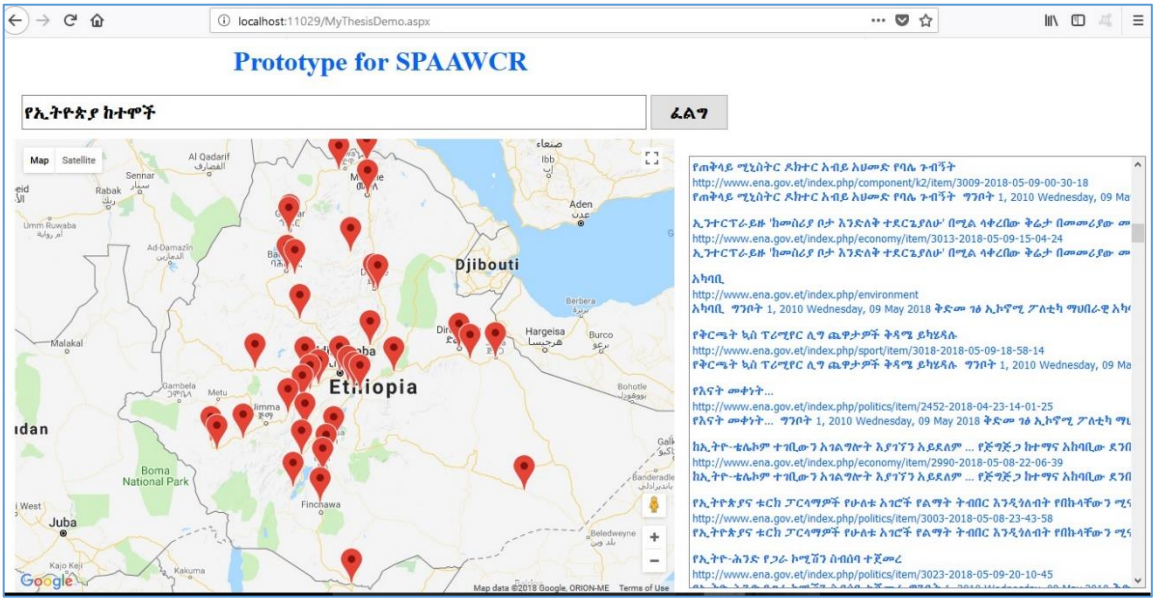

Figure 5.20: Our Model's result for query "የኢትዮጵያ ከተሞች"

Comparison between our model and the three search engines is elaborated in Table 5.8. From the three query formulations Google returns a spatial reference only for the <Place Name Only> query formulation. It didn't provide any spatial information for the other two query formulations. Yahoo and Bing didn't retrieve any spatial data at all. Unlike those three search engines our model maintains spatial awareness and plot the points on the digital map which are relevant to the query.

Table 5.8: Comparison of search results

Query Formulation	**Total Number of Points Plotted on the map**				**Number of Relevant Points Plotted on the map**			
	Google	**Yahoo**	**Bing**	**Our Model**	**Google**	**Yahoo**	**Bing**	**Our Model**
<Place Name Only> አዲስ አበባ ዩኒቨርሲቲ	3	0	0	4	3	0	0	4
<Place Name, Spatial Relationship, Feature Type> በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች	0	0	0	9	0	0	0	8
<Feature Type Only> የኢትዮጵያ ከተሞች	0	0	0	33	0	0	0	33

5.5 Precision and Recall

In order to evaluate the performance of our model we used the commonly used information retrieval evaluation measures which are Recall, Precision, and F-measure. Recall is the ratio of the number of documents retrieved correctly to the total number of relevant documents in the document collection. It is the number of correct results divided by the number of results that should have been returned.

$$Recall = \frac{Number\ of\ Relevant\ Documents\ Retrieved}{Number\ of\ Relevant\ Documents\ in\ the\ Collection} \quad (1)$$

Whereas precision is the ratio of the number of documents retrieved correctly to the total number of documents retrieved. In other words, it is the number of correct results divided by the number of all returned results.

$$Precision = \frac{Number\ of\ Relevant\ Documents\ Retrieved}{Number\ of\ Documents\ Retrieved} \quad (2)$$

F-measure is a standard measure that tests the accuracy of the evaluation. It considers recall and precision techniques to get the harmonic mean of the two.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

We conducted the evaluation by consulting experts who are geo-system engineers working on the development of location-based systems and services, each of them has more than five years of work experience. We measured the accuracy of our model based on the queries these experts generate.

We selected six queries out of the fifteen queries the experts formulated. List of the fifteen queries is attached as Annex C. Those queries are formulated in three types. The first one consists of <Place Name Only>, the other consists of a triplet <Place Name, Spatial Relationship, Feature Type>, and the last one <Feature Type Only>.

The matrix presented in Table 5.9 summarizes the results we found while testing the accuracy of our model via precision and recall. These accuracy measures are expressed in equations equation (1), equation (2), and equation (3).

Table 5.9: Result for Precision and Recall Evaluation

Query Formulation	**Number of Relevant Footprints in the Collection**	**Number of Retrieved Footprints**	**Number of Relevant Footprints Retrieved**	**Precision**	**Recall**	**F-measure**
አዲስ አበባ ዩኒቨርሲቲ	4	4	4	1	1	1
በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች	12	9	8	0.89	0.67	0.76
የኢትዮጵያ ከተሞች	38	33	33	0.87	1	0.93
መስቀል ፍላወር	1	1	1	1	1	1
በኢትዮጵያ ያሉ ኤምባሲዎች	14	9	9	0.64	1	0.78
መንግስታዊ ተቋማት	17	14	14	0.79	0.76	0.78
Average				**0.87**	**0.91**	**0.88**

5.6 Discussion

The results found after the evaluation of our prototype system indicates that our model identifies spatial qualifiers for Amharic language much better than the three well-known search engines. Google identifies the existence of place names in the query and plots their location on the map. Neither Yahoo nor Bing return spatial data other than list of web URLs which consists the query string fully or partially.

Based on the result set of Table 5.9 we can say that our model is 88% accurate. It performs better for the query formation that consists only place names <Place Name Only> than a query that consists of place name, spatial relationship and geographic feature type formulated as <Place Name, Spatial Relationship, Feature Type>.

Chapter Six

Conclusions and Future Work

Conclusion

Nowadays, the usage of geographical data is becoming an area of increased interest and research. It is obvious that what people do happens somewhere in the globe, that means it is geo-referenced and it has location which can be represented by place name, and geographic coordinates. It is also clear that many web resources contain information about peoples' activities, incidents, and occasions that happened somewhere.

There is no doubt for the need of information retrieval which considers the interaction of peoples to location. So that the facts show us the potential benefit of spatially aware web content retrieval. People are diversified throughout the world not only in location but also in culture, language and so on. Since Amharic is one of the languages spoken on earth, there should be a way to gather information about events that happened somewhere and written in Amharic language.

This work focused on filling the gap that exists in previously designed and implemented Amharic web content retrieval systems. In particular, we present a model that introduces spatial awareness for Amharic web content retrieval. We developed a prototype system to demonstrate the implementation of our proposed model. It provides a solution to the problem of Amharic information retrieval in relevance with geographic information.

We build a geo-ontology as a knowledge base which assists in identification of geographical concepts and their spatial relationship. We conceptualize geographical features by classifying them in three categories which are point objects, linear objects, and areal objects. We identified geographical feature types, instantiate individual objects, and identified the spatial relationship between these objects. The geo-ontology provides support for query disambiguation, query expansion via the generation of geometric query footprints and metadata extraction to store the geographical context of the crawled web documents and geo-datasets.

While creating our knowledge base which is the geo-ontology, we grab a spatial data from Open Street Map and populated our dataset with those spatial data. We extract location

data from the dataset, in particular place names and geographic coordinates, and instantiate individual geographic objects.

In addition to the geo-ontology, we introduce geo-coding, geo-parsing, and spatial indexing techniques to the Amharic web content retrieval. The model allows documents to be indexed spatially. It also adds a feature to the ranking of query results according to spatial relevance to the query, and the graphical display of search results on a digital map.

Finally, we built a prototype application and evaluated the accuracy of our model. We formulated queries in three formats which consists <Place Name only>, <A Place Name, Spatial Relationship, Geographic Feature Type>, and <Geographic Feature Type only>. We perform the evaluation using 275 known place names, 38 cities in Ethiopia, and 35 geographical feature types. Because of resource limitation, which is inability of getting boundary data for the polygons and linear objects, we didn't test the model for areal objects or polygons. Rather, we represent areal objects or polygons like cities as a point object.

Contribution of this Work

The main contribution of this work for Amharic web content retrieval is listed below:

- Introducing spatial awareness to the Amharic web content retrieval
- Designing an Amharic geo-ontology and populating geographic objects as individuals with their respective footprint
- Designing an algorithm that disambiguates the query by identifying the existence of place names, geo-spatial features or a spatial relationship
- Implementing geo-coding, geo-parsing, and spatial indexing techniques for the Amharic document

Future Work

In this work, we designed a model for spatially aware Amharic web content retrieval which can be easily integrated with the existing Amharic search engines developed by different researchers. There is still room for improvement. As a recommendation of improvement to our designed model, it will be great if the following features can be incorporated in the future work:

- Integration of the model with a semantic Amharic search engine

- Improving the efficiency of the model in identification of the spatial relationships by incorporating an Amharic morphological analyzer
- Improving the query parser, so that it can handle sophisticated query strings which consists of more than one spatial relationship.

References

[1] "Information Retrieval," Wikepedia, the free encyclopedia , 30 March 2010. [Online]. Available: https://en.wikipedia.org/wiki/Information_retrieval. [Accessed 17/10/2016].

[2] M. Tessema, "Design and Implementation of Amharic Search Engine , Masters' Thesis," Addis Ababa University, Addis Ababa, 2007.

[3] R. S. P. Christopher J. Bones, "Geographical Information Retrieval," *International Journal of Geographical Information Science,* vol. 23, no. 2, pp. 219-228, 2008.

[4] M. E. H. Enas, "World Geographical Ontology Model," *International Journal of Computer Applications,* vol. 120, no. 15, pp. 25-33, 2015.

[5] R. S. Purves, P. Clough and C. B. Jones, "The Design and Implementation of SPIRIT: a SpatiallyAware Search Engine for Information Retrieval on the," *International Journal of Geographical Information Science,* vol. 21, no. 7, pp. 717-745, 2007.

[6] M. Egenhofer, "Toward the Semantic Geospatial Web," in *10th ACM International Symposium In Geographic Information Systems*, New York, 2002.

[7] Q. Li, J. Wang and H. Li, "Using Semantic Wikis as Collaborative Tools for Geo-Ontology," in *Geoinformatics, 18th International Conference*, Beijing, 2010.

[8] N. Guarino, Formal ontology and information systems, Amsterdam: The Netherlands: IOS Press, 1998.

[9] D. Mark, S. B and B. Tversky, "Ontology and Geographic Objects: An Empirical Study of Cognitive Categorization," in *International Conference COSIT*, Berlin, 1999.

[10] A. A. Mortadha, "Semantic Geospatial Search and Ranking in the Context of the Geographical Information System TerraFly," University of Miami, 2010.

[11] S. Arno and T. Klaus, "The Geospatial Web: How Geobrowsers,Social Software and the Web 2.0 are Shaping the Network Society (Advanced Information and Knowledge Processing)," Springer-Verlag New York, Inc., NJ, 2007.

[12] "Reasoning on the Web with Rules and Semantics, 2005," REWERSE , 2005. [Online]. Available: http://www.rewerse.net/publications/rewersepublications.html#REWERSE-DEL-2005-A1-D2. [Accessed 13/01/2017].

[13] Pew Research Center, "World Wide Web Timeline," Pew Research Center, 11 03 2014. [Online]. Available: http://www.pewinternet.org/2014/03/11/world-wide-web-timeline/. [Accessed 25/05/2018].

[14] D. Caitlin, "36 ways the Web has changed us," The Washington Post, 12 03 2014. [Online].Available:https://www.washingtonpost.com/news/arts-and-entertainment/wp/2014/03/12/36-ways-the-web-has-changed-us. [Accessed 25/05/2018].

[15] Wikipedia, "World Wide Web," Wikipedia, the free encyclopedia, 18 May 2018. [Online]. Available: https://en.wikipedia.org/wiki/World_Wide_Web. [Accessed May 2018].

[16] Wikipedia, with help from Bart Pursel, "Search Engines," 2005. [Online]. Available: https://psu.pb.unizin.org/ist110/chapter/2-1-search-engines/. [Accessed 25/05/2018].

[17] Wikipedia,"Web search engine," Wikipedia, the free encyclopedia,23/05/2018. [Online]. Available:https://en.wikipedia.org/wiki/Web_search_engine. [Accessed 25/05/2018].

[18] S. Amit, "Modern Information Retrieval: A Brief Overview," *IEEE Computer Society Technical Committee on Data Engineering,* vol. 24, no. 4, pp. 35-42, 2001.

[19] D. M. Christopher, R. Prabhakar and H. Schütze, An Introduction to Information Retrieval, Cambridge : Cambridge University Press, 2009.

[20] A. d. B. Rolf, A. K. Richard, S. Yuxian, C. E. Martin, K. Menno-Jan and J. C. W. Michael, Principles of Geographic Information Systems, Enschede, The Netherlands: The International Institute for Aerospace Survey and Earth Sciences, 2001.

[21] B. Yaser, "Geospatial Semantic Web: Applications," in *Encyclopedia of GIS*, New York, Springer, 2008, pp. 391-398.

[22] R. Henriksson, T. Kauppinen and E. Hyvönen, "Core Geographical Concepts: Case Finnish Geo-Ontology," in *The first International Workshop on Location and the Web*, New York, 2008.

[23] T. A. Alia, D. S. Philip, B. J. Christopher, F. Gaihua and D. Finch, "A Critical Evaluation of Ontology Languages for Geographic Information Retrieval on the Internet," *Journal of Visual Languages and Computing,* vol. 16, no. 4, pp. 331-358, 2005 .

[24] B. Li, L. Jiping and S. Lihong, "Research on Geo-Ontology Construction based on Spatial Affairs," in *International Conference on Earth Observation Data Processing and Analysis (ICEODPA)*, Bellingham, 2008.

[25] Z. Xue and X. Jun, "Construction of Geo-Ontology Knowledge Base about Spatial Relations," in *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference*, Fuzhou, 2011.

[26] F. M. Z. Oliver, "Implementation of a Spatially-Aware Image Search Engine and Its Evaluation using Crowdsourced Relevance Judgements, Master's Thesis," University of Zurich, Zurich, 2013.

[27] P. Damien, C. Guillaume, S. Christian and H. Gilles, "On the evaluation of Geographic Information Retrieval systems," *International Journal of Digital Library,* vol. 11, pp. 91-109, 2011.

[28] H. Wang, L. Li and P.-c. Song, "Design of Geo-Ontology based on Concept Lattice," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 1-5, 2005.

[29] Y. Kun, W. Jun and P. Shuang-yun, "The Research and Practice of Geo-Ontology Construction," in *ISPRS*, Beijing, 2005.

[30] B. J. Christopher, A. I. Abdelmoty, F. David, G. Fu and S. Vaid, "The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing," in *Third International Conference, GIScience* , Adelphi, 2004.

[31] A. Lars, A. A. Atelach, G. Bjorn, E. A. Samuel and N. H. Lemma, "Classifying Amharic Webnews," *Sringer*, vol. 12, no. 3, p. 416–435, 2009.

[32] M. Tessema, H. Redwan and A. Solomon, "Searching the Web for Amharic Content," *Journal of Multimedia Processing and Technologies*, vol. 1, pp. 16-28, 2010.

[33] A. Nega and W. Peter, "Stemming of Amharic Words for Information Retrieval," *Literary and Linguistic Computing*, vol. 17, no. 1, pp. 1-17, 2002.

[34] M. K. Gobena, "Implementing an Open Source Amharic Resource Grammar in GF," University of Gothenburg , Masters Thesis, Göteborg, Sweden, 2010.

[35] R. Hassen, "Enhanced Design of Amharic Search Engine, Masters' Thesis," Addis Ababa University, Addis Ababa, 2008.

[36] Smart, P; Abdelmoty, A I; CB, Jones, "An Evaluation of Ontology Representation Languages for Supporting Web Retrieval of Geographical Information," in *www.geo-spirit.org/publications*, Norwich, 2004.

[37] J. Albrecht, B. Derman and L. Raubramanian, "Geo-ontology Tools: The Missing Link," *Transactions in GIS*, vol. 12, no. 4, pp. 409-424, 2008.

[38] S. A. Mequannint Munye, "Amharic-English Bilingual Web Search Engine," in *International Conference on Management of Emergent Digital EcoSystems*, New York, 2012.

[39] A. I. A. G. F. Christopher B. Jones, "Maintaining Ontologies for Geographical Information Retrieval on the Web," in *OTM Confederated International Conferences*, atania, Sicily, Italy, 2003.

[40] P. Ross S., C. Paul, J. Christopher B., A. Avi, B. Benedicte, F. David, F. Gaihua, J. Hideo, S. Awase Khirni, V. Subodh and Y. Bisheng, "The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet," *International Journal of Geographical Information Science*, vol. 21, no. 17, pp. 717-745, 2007.

[41] M. Fernández López, "Overview Of Methodologies For Building Ontologies," in *Ontologies and Problem-Solving Methods (KRR5)*, Stockholm, 1999.

[42] E. Tomai and M. Spanaki, "From ontology design to ontology implementation : A web tool for building geographic ontologies," in *The 7th AGILE Conference on Geographic Information Science*, Heraklion, 2004.

[43] M. Gasser, "HORNMORPHO 2.5 User's Guide," Indiana University, School of Informatics and Computing, 2012.

[44] G. Anuradha, C. Cindy X., C. Kajal T. and U.-S. Rosario, "From Ontology to Relational Databases," in *Conceptual Modeling for Advanced Application Domains*, Springer, Berlin, Heidelberg, 2004.

[45] G. Bumans, "Mapping between Relational Databases and OWL Ontologies: an example," University of Ltivia, 2010.

[46] Rick,Byham;Craig Guyer, "Spatial Indexes Overview," Microsift, 12 09 2016. [Online]. Available: https://docs.microsoft.com/en-us/sql/relational-databases/spatial/spatial-indexes-overview. [Accessed 03/06/2017].

[47] Rick, Byham;Craig Guyer, "CREATE SPATIAL INDEX (Transact-SQL)," Microsoft, 11 04 2017. [Online]. Available: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-spatial-index-transact-sql. [Accessed 03 06 2017].

[48] "Tesselation" ESRI/GIS, [Online]. Available:http://wiki.gis.com/wiki/index.php/Tessellation. [Accessed 03/06/2017].

[49] C. Kumar, "Relevance and Ranking in Geographic Information Retrieval," in *FDIA'11 Proceedings of the Fourth BCS-IRSG conference on Future Directions in Information Access*, Koblenz, Germany, 2011.

[50] Apache, "Nutch Wiki," The Apache Software Foundation, [Online]. Available: https://wiki.apache.org/nutch/FrontPage. [Accessed 26/04/2018].

[51] Apache, "Solr," Apache Software Foundation, [Online]. Available: http://lucene.apache.org/solr/#intro. [Accessed 24/04/2018].

[52] Google,"Google Maps Platform,",[Online]. Available: https://developers.google.com/maps/ documentation/javascript/tutorial. [Accessed 14/03/2017].

[53] S. Wolfram and R. Slaven, "Bbbike," OpenStreetMap , 2018. [Online]. Available: https://download.bbbike.org/osm/extract/planet_38.581,8.783_39.022,9.08.osm.shp.zip. [Accessed 02/06/2017].

[54] OpenStreetMap, " OpenStreetMap powers map data on thousands of web sites, mobile apps, and hardware devices," [Online]. Available: https://www.openstreetmap.org/about. [Accessed 02/06/2017].

[55] H. Florian, "Nutch – How It Works," 04 03 2012. [Online]. Available: https://florianhartl.com/nutch-how-it-works.html. [Accessed 28/05/2018].

[56] C. B. Jones, R. S. Purves, P. D. Clough and H. Joho, "Modelling Vague Places with Knowledge from the Web," *International Journal of Geographical Information Science,* vol. 22, no. 10, pp. 1045-1065, 2008.

[57] P. Smart, A. Abdelmoty, B. El-Geresy and C. Jones, "A Framework for combining Rules and Geo-ontologies," *Web Reasoning and Rule Systems,* vol. 4524, pp. 133-147, 2007.

[58] P. Damien, D. Curdin and S. P. Ross, "Development and evaluation of a geographic information retrieval system using fine grained toponyms," *GOSIS,* vol. 11, no. 193, pp. 1-29, 2015.

[59] P. Gimenez, A. Tanaka and F. Baião, "A geo-ontology to support the semantic integration of geo-information from the National Spatial Data Infrastructure," in *GEOINFO* , Campos do Jordan, 2013.

[60] R. Xiao, W. Liu, Y. Du and Y. He, "An Intelligent Bay Geo-information Retrieval Approach based on Geo-ontology," in *17th International Conference on Geoinformatics*, Fairfax, VA, 2009.

[61] K. Mei and F. Bian, "An Ontology-Based Approach for Geographic Information Retrieval on the Web," in *International Conference on Wireless Communications*, Shangai, 2007 .

[62] J. Vote, "Search engines with ASP.NET," Developer Fusion, 06 2011. [Online]. Available: http://www.developerfusion.com/article/84415/search-engines-with-aspnet/. [Accessed 12/03/2017].

Annexes

Annex A: Source code (OWL) for the Geo-ontology

```
<?xml version="1.0"?>
<Ontology xmlns=http://www.w3.org/2002/07/owl#
  xml:base=http://www.semanticweb.org/geo-ontology
  xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
  xmlns:xml=http://www.w3.org/XML/1998/namespace
  xmlns:xsd=http://www.w3.org/2001/XMLSchema#
  xmlns:rdfs=http://www.w3.org/2000/01/rdf-schema#
  ontologyIRI=http://www.semanticweb.org/geo-ontology >
  <Prefix name="" IRI=http://www.semanticweb.org/geo-
ontology />
 <Prefix name="owl" IRI=http://www.w3.org/2002/07/owl# />
 <Prefix name="rdf" IRI=http://www.w3.org/1999/02/22-rdf-
syntax-ns# />
<Prefix name="xml"IRI=http://www.w3.org/XML/1998/
namespace />
<Prefix name="xsd" IRI=http://www.w3.org/2001/XMLSchema#
/>
<Prefix name="rdfs" IRI=http://www.w3.org/2000/01/ rdf-
schema# />
    <Declaration>
        <Class IRI="#yefaynans_teqWamat"/>
    </Declaration>
    <Declaration>
        <NamedIndividual
IRI="#?&apos;aqaqi_qaliti_kfle_ketema"/>
    </Declaration>
    <Declaration>
        <NamedIndividual IRI="#?qirqos_kfle_ketema"/>
    </Declaration>
    <Declaration>
```

```
        <Class IRI="#menged"/>
    </Declaration>
    <Declaration>
        <ObjectProperty IRI="#near"/>
    </Declaration>
    <Declaration>
        <ObjectProperty IRI="#partOf"/>
    </Declaration>
    <Declaration>
        <ObjectProperty IRI="#overlaps"/>
    </Declaration>
    <Declaration>
        <Class IRI="#hayq"/>
    </Declaration>
    <Declaration>
        <Class IRI="#menaheriya"/>
    </Declaration>
    <Declaration>
        <ObjectProperty IRI="#adjacentTo"/>
    </Declaration>
    <Declaration>
        <DataProperty IRI="#sIfat"/>
    </Declaration>
    <Declaration>
        <Class IRI="#kolEj"/>
    </Declaration>
    <Declaration>
        <Class IRI="#kIlIl"/>
    </Declaration>
        <Class IRI="#wenz"/>
        <Class IRI="#LinearObject"/>
    </SubClassOf>
```

```
    <SubClassOf>
        <Class IRI="#wereda"/>
        <Class IRI="#ArealObject"/>
    </SubClassOf>
    <SubClassOf>
        <Class IRI="#yeTEna_teqWamat"/>
        <Class IRI="#PointObject"/>
    </SubClassOf>
    <DataPropertyDomain>
        <DataProperty IRI="#meley_a"/>
        <Class IRI="#ArealObject"/>
    </DataPropertyDomain>
    <DataPropertyDomain>
        <DataProperty IRI="#meley_a"/>
        <Class IRI="#LinearObject"/>
    </DataPropertyDomain>
    <DataPropertyDomain>
        <DataProperty IRI="#sIm"/>
        <Class IRI="#LinearObject"/>
    </DataPropertyDomain>
    <DataPropertyDomain>
        <DataProperty IRI="#sIm"/>
        <Class IRI="#PointObject"/>
    </DataPropertyDomain>
    <DataPropertyDomain>
        <DataProperty IRI="#?rIzmet"/>
        <Class IRI="#LinearObject"/>
    </DataPropertyDomain>
    <DataPropertyDomain>
        <DataProperty IRI="#?ma&apos;kelawi_neTb"/>
        <Class IRI="#PointObject"/>
    </DataPropertyDomain>
```

```
    <DataPropertyRange>
        <DataProperty IRI="#meley_a"/>
        <Datatype abbreviatedIRI="xsd:integer"/>
    </DataPropertyRange>
    <DataPropertyRange>
        <DataProperty IRI="#sIfat"/>
        <Datatype abbreviatedIRI="xsd:float"/>
    </DataPropertyRange>
    <DataPropertyRange>
        <DataProperty IRI="#sIm"/>
        <Datatype abbreviatedIRI="xsd:string"/>
    </DataPropertyRange>
    <DataPropertyRange>
        <DataProperty IRI="#?rIzmet"/>
        <Datatype abbreviatedIRI="xsd:float"/>
    </DataPropertyRange>
    <DataPropertyRange>
        <DataProperty IRI="#?ma&apos;kelawi_neTb"/>
        <Datatype abbreviatedIRI="xsd:float"/>
    </DataPropertyRange>
</Ontology>

<!-- Generated by the OWL API (version 4.2.1.20160306-
0033) https://github.com/owlcs/owlapi -->
```

Annex B: SQL Server Syntax to Create a Spatial Index

```
-- SQL Server Syntax
CREATE SPATIAL INDEX index_geoontology
  ON <GeoOntology> ( Centeroid )
   {
     <geography_tessellation>
   }
   [ ON { filegroup_geo | "default" } ] [;]
  <object>::=  GeoOntology. PointObject ]
  <geography_tessellation> ::=
   {
     <geography_automatic_grid_tessellation>
      | <geography_manual_grid_tessellation>
   }
 <geography_automatic_grid_tessellation> ::=
  {
    [ USING GEOGRAPHY_AUTO_GRID ]
    [ WITH (
      [ [,] <tessellation_cells_per_object> [ 12] ]
      [ [,] <spatial_index_option> ]
     ) ]
   }
 <geography_manual_grid_tessellation> ::=
  {
   [ USING GEOGRAPHY_GRID ]
     WITH (
           [ <tessellation_grid> [ ,…n] ]
           [ [,] <tessellation_cells_per_object>[ 64] ]
           [ [,] <spatial_index_option> [ ,…n] ]
        ) ]
   }
 <bounding_box> ::=
   {
     BOUNDING_BOX = ( {
      xmin, ymin, xmax, ymax
       |<named_bb_coordinate>,<named_bb_coordinate>,
<named_bb_coordinate>, <named_bb_coordinate>
```

```
  } )
  }
<named_bb_coordinate> ::= { XMIN = xmin | YMIN = ymin | XMAX =
xmax | YMAX=ymax }
<tesselation_grid> ::=
{
    GRIDS = ( { <grid_level> [ ,...n ] | <grid_size>,
<grid_size>, <grid_size>, <grid_size> }
        )
}
<tesseallation_cells_per_object> ::=
{
   CELLS_PER_OBJECT = n
}
<grid_level> ::=
{
     LEVEL_1 = <grid_size> | LEVEL_2 = <grid_size>
     |LEVEL_3 = <grid_size> | LEVEL_4 = <grid_size>
}
<grid_size> ::= { LOW | MEDIUM | HIGH }
<spatial_index_option> ::=
{
    PAD_INDEX = { ON | OFF }
  | FILLFACTOR = fillfactor
  | SORT_IN_TEMPDB = { ON | OFF }
  | IGNORE_DUP_KEY = OFF
  | STATISTICS_NORECOMPUTE = { ON | OFF }
  | DROP_EXISTING = { ON | OFF }
  | ONLINE = OFF
  | ALLOW_ROW_LOCKS = { ON | OFF }
  | ALLOW_PAGE_LOCKS = { ON | OFF }
  | MAXDOP = max_degree_of_parallelism
    | DATA_COMPRESSION = { NONE | ROW | PAGE }  }
```

Annex C: Queries formulated by Experts

Query Number	Expert Evaluated	Query
Query 1	Expert 1	አዲስ አበባ ዩኒቨርሲቲ
Query 2	Expert 2	በአዲስ አበባ የሚገኙ ዩኒቨርሲቲዎች
Query 3	Expert 3	የኢትዮጵያ ከተሞች
Query 4	Expert 4	መስቀል ፍላወር
Query 5	Expert 5	በኢትዮጵያ ያሉ ኤምባሲዎች
Query 6	Expert 1	መንግስታዊ ተቋማት
Query 7	Expert 2	ቂርቆስ ክፍለ ከተማ
Query 8	Expert 3	ቦሌ አካባቢ የሚገኙ ሆቴሎች
Query 9	Expert 4	ቅዱስ ጊዮርጊስ ቤተክርስቲያን
Query 10	Expert 5	በጉለሌ ክፍለ ከተማ የሚገኙ ባንኮች
Query 11	Expert 1	መስቀል አደባባይ
Query 12	Expert 2	ክፍለ ከተሞች
Query 13	Expert 3	ጥቁር አንበሳ ሆስፒታል
Query 14	Expert 4	የኢትዮጵያ ንግድ ባንክ
Query 15	Expert 5	ሀውልቶች በአዲስ አበባ

Annex D: Sample Place Names with their Coordinated

Place Name	Latitude	Longitude
የድል ሐውልት	38.7634	9.03291
ኮልፌ ቀራንዮ	38.6875	9.00496
ካዛንቺስ	38.7712	9.01593
ደሴ	39.63498	11.1226
ጃፓን ኤምባሲ	38.7648	9.00902
አዲስ አበባ ዩኒቨርሲቲ	38.7643	9.03281
ቅዱስ ኡራኤል	38.775	9.01094
አዲግራት	39.46182	14.28031
ፓርላማ	38.7646	9.03024
ዩኒቲ ዩኒቨርሲቲ	38.807803	9.001129
ሚዛን ተፈሪ	35.5765	6.99213
ፊሊፕስ	38.741	9.01035
አመዴ	38.7352	9.0328
አዲሱ ገበያ	38.7373	9.05959
ሩፋኤል	38.7274	9.05888
ቦስተን	38.7838	8.99056
መርካቶ	38.7376	9.03129
መስቀል ፍላወር	38.7643	8.98801
ብራስ ሆስፒታል	38.7937	8.99022
አቡነ ጴጥሮስ	38.7496	9.0354
አዲስ ከተማ ክፍለ ከተማ	38.7276	9.03575

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: Masreshaye Worku Edo_

Signature: _____

Date: _____

Confirmed by advisor:

Name: Dr. Solomon Atnafu_

Signature: _____

Date: _____