Addis Ababa University

College of Nature Sciences

IMPLEMENTATION OF AFAAN OROMO ONTOLOGY ON SPORT DOMAIN

HAIMANOT KEBEDE

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

Addis Ababa, Ethiopia

August, 2017

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCES

IMPLEMENTATION OF AFAAN OROMO ONTOLOGY ON SPORT DOMAIN

Haimanot Kebede

ADVISOR: Yaregal Assabie (PHD)

This is to certify that the thesis prepared by Haimanot Kebede, titled *Implementation of Afaan Oromo Ontology on Sport Domain* and submitted in partial fulfilment of the requirements for the Degree of Master of Science in Computer Science complies with the regulation of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Name	Signature	Date
1. Advisor _____	_____	_____
2. _____	_____	_____
3. _____	_____	_____

Abstract

Keyword based information retrieval system is affected by factors like synonym, polysemy and the manner in which queries are delivered to the retrieval system. Ontology is one of the semantic technologies employed in the information retrieval system to overcome the problems related to variations in document and query representation. It is achieved through explicitly specifying concepts and their relationships so that different components have common understanding of concepts to work in liaison.

In the Implementation of Afaan Oromo Ontology on Sport Domain, ontology is developed on sport domain. The ontology was developed using Protégé 4.3 in which 317 classes, 1748 individuals, 7747 axioms, 5455 logical axioms, 191 object properties, and 17 data properties were identified. OWL Ontology that uses Pellet reasoner for information classification and consistency checking is developed.

Implementation of the developed system is evaluated on the information retrieval of Afaan Oromo text using precision of the retrieved documents. Comparison is made between ontology aided Afaan Oromo text retrieval system and retrieval executed directly by using the queries as they are. Afaan Oromo information retrieval using the developed ontology on query operation stage has recorded a precision of 81.61% against 32.16% precision when the Ontology is not used. Queries used to test performance of the system have descriptive nature in which the description of the information needed is delivered to the system than the direct keywords by which the required items can be retrieved.

Keyword: Ontology, Text Retrieval, Afaan Oromo text retrieval, Afaan Oromo Sport Ontology, Implementation of Afaan Oromo Ontology on Sport Domain

Acknowledgement

My special thanks go to my Advisor Doctor Yaregal Assabie who has coached and gave me an advice and support that exceed beyond student and lecturer relationships during my entire thesis work. I would like to thank you for encouragement you gave me and boundless patient you showed during my challenges. I learned a lot from this work because you had a belief in me I could do it.

Table of Contents

List of Tables

List of Figures

Acronyms and Abbreviations

AO – Afaan Oromo

AOO – Afaan Oromo Ontology

AOSO – Afaan Oromo Sport Ontology

API – Application Programming Interface

CEL – Classifier for $\varepsilon\mathcal{L}^{+}$

CLIR – Cross Lingual Information Retrieval

CNN – Cables News Network

DARPA – Defense Advanced Research Projects Agency

DAML – DARPA Agent Markup Language

DL – Description Logic

DOC – Document

DOR_Bara-2016 – Dorgommii_Olompiikii_Riyoodejaaneroo_Bara_2016

FE – Free Edition

HTML – Hyper Text Markup Language

IBM – International Business Machine

IDF – Inverse Document Frequency

IR – Information Retrieval

IRS – Information Retrieval System

KIF – Knowledge Interchange Format

LSI - Latent Semantic Indexing

OBAOTRS – Ontology-Based Afaan Oromo Text Retrieval System

OCMC – Operational Conceptual Modeling Language

OIL – Ontology Interface Language

OWL – Web Ontology Language

QO – Query Operation

RACER – Renamed ABoxes and Concept Expression Reasoner

RDF – Resource Description Framework

RDFS – Resource Description Framework Schema

SE – Standard Edition

SHOE – Simple HTML Ontology Extensions

SPARQL – Simple Protocol and RDF Query Language

SWRL – Semantic Web Rule Language

TBC – Top Braid Composer

TF – Term Frequency

TF/IDF -Term Frequency/ Inverted Term Frequency

TOVE - TOronto Virtual Enterprise project

WWW – World Wide Web

W3C – World Wide Web Consortium

XML – eXtended Markup Language

XOL – XML based Ontology Exchange Language

Chapter One: Introduction

1.1 Background

Modern era is marked by the invention of highly sophisticated communication technologies that are exploited by mankind to serve information of any sort anywhere within a fraction of seconds. Getting the right quality and quantity of information is very crucial for the success of any mission. Without having the right and accurate information, it is hardly possible to accomplish anything at all. Without information, it is like moving without guidance. It is one of fundamental assets that plays an important role in the life of human being.

Information can be preserved in different forms. We can have information in oral form that has traversed generations in spoken forms. We can also have well organized information that is conserved in digital or electronics form or as a hard copy written on papers or other forms.

Digital documents or information can be stored on web servers found at any corner of the world or it can be kept on our personal computers, mobile phones, handheld digital devices and so on. In the same way we access information from digital devices around us, it is possible to access information found in a remote machine that is interconnected through computer network.

Ability to transfer and acquire digital information from computers interconnected through computer network takes us to the inception of the concept information retrieval. Information retrieval [1] is finding material (usually documents) of unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). Or it can generally be described as the science of locating those documents that fulfil specified information need of the users from large document collections [2, 3, 4]. Information retrieval can also be described as a discipline within a computer science that studies the representation, storage, organization of, access to the

information items in which the representation and organization of the information items should provide the user with easy access to the information [5].

Rapid developments in Information and Communication Technologies are making available of huge amount of data and information on the web. According to [6], the web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. In spite of these benefits, the web has also introduced new challenges of its own. Finding useful information on the web is frequently a tedious and becomes difficult task [7]. In such circumstances, the information seekers need to navigate the space of the web links looking for the required information. However, since the search space is vast, such navigation task is usually inefficient. For naïve users in particular, the problem becomes even harder. Hence, the availability or presence of multitude documents on the web strongly demands the need for the system that assists users in the retrieval process so that the right documents are found with ease and simplicity.

1.2 Motivation

We are now on a time when a huge amount of information written in Afaan Oromo language on different topics is available on the web for a considerable number of users of the language. However, due to variations in culture, level of understanding of the domain there is inability to formulate the right query that enables to retrieve the required information. As a result, users may not retrieve the right quality of information to the extent they are entitled to. In addition to this, variations in the manner of information representation demands users to know the way it is represented so as to formulate the query that can access the required documents. To reconcile the gap on possible variation of information need and document representation of Afaan Oromo language, a need arise to formalize domain knowledge that will work in liaison with different applications.

In a system that functions in formalized domain knowledge, the knowledge base or ontology will be contacted before the actual document retrieval process commences in order to convert users query into formalized domain concept. In the paradigm of formalized domain knowledge, the meanings of concepts that are formalized are the ones

that are shared among members of the communities. If the meaning of something is shared among parties involving in the communication process, common understanding is promoted and hence, the response for the information quest will be impacted with better result that will better benefit the information seekers.

With this in mind, to make Afaan Oromo text retrieval system less prone to variations in document and query representation and promote common understanding of parities involving in the process, it is necessary to develop ontology on different domains and integrate it into information retrieval system. Hence, implementation of Afaan Oromo ontology on sport domain is one step forward to have formalized representation of concepts on different domain.

1.3 Statement of the problem

In a classical keyword based Afaan Oromo text retrieval systems [8, 9], a given document is retrieved when its index term contains a query keywords derived from users information input in spite of the fact that the meaning of the word used in the document may differ from what the users' meant to search. That means retrieval is executed when query keywords and document index terms are the same. However, different research works conducted on different languages have revealed that information retrieval systems operating based on mere keyword matching techniques have exhibited limitations with respect to retrieval performance of the system [10, 11, 12]. The fate of current keyword based Afaan Oromo text retrieval systems is not different from what has been revealed in other languages as it is operating based on word matching principle.

Retrieval quality of text retrieval system based on word matching algorithm is highly affected by the presence or association of a word with multiple meanings. These words are characterized by having different meanings based on the context [13, 14, 15] of the text in which they are represented. There is high possibility for these terms in the query to be matched with words in irrelevant documents and could be the reason for some of the search result to be too broads [16]. This scenario forces users to select documents they need from multitude of documents delivered by the search engine that could have little or no relevance to the information needed.

Similarly, documents indexed or represented with keywords different from query keyword but represent related content to the users information need may not be retrieved by the keyword matching retrieval system even though these words have equivalent meaning and refer to the same things.

1.4 Objective

General Objective

The general objective of this research is to develop an implementation of Afaan Oromo ontology on sport domain.

Specific Objectives

- To understand the basics of the Afaan Oromo language.
- To review and understand the basics of classical information retrieval system on different languages
- To review existing Afaan Oromo text retrieval system
- To review ontology-based information retrieval system.
- To collect and organize Afaan Oromo text corpus written on sport domain.
- To extract concepts related to sport domain from Afaan Oromo corpus.
- To develop Afaan Oromo ontology on sport domain from the extracted concepts.
- To design an architecture of implementation of Afaan Oromo ontology on sport domain
- To develop prototype of ontology-based Afaan Oromo text retrieval system
- To evaluate performance of ontology aided information retrieval system

1.5 Scope and Limitation

Scope

Ontology is developed for concepts extracted from Afaan Oromo (AO) sport news. Sport related information collected from online sources written in Afaan Oromo Text that cover Athletics, football, Horse race, Bicycle, Swimming, Hockey are included in the Ontology.

The morphological analysis and other details related to Afaan Oromo language are excluded and are out of the scope of this study.

Limitation

- The knowledge base captured covers only limited area of sport domain for the demonstration purpose.

1.6 Methodology

Literature review

A thorough review of literatures on information retrieval, ontology development and previous related research works done on the key word and ontology based information retrieval system of both Ethiopian and non-Ethiopian languages will be done for better understanding of the area of the study.

Tools

Several tools will be used for the development and implementation of Afaan Oromo Sport Ontology (AOSO). Protégé OWL ontology editor will be used to capture the concepts, individuals and the relationships between concepts and individuals. Pellet reasoner will be integrated into Protégé to have the capability of inferring additional information that is not explicitly given. Graviz-2.38 will also be included into Protégé library to enable graphical display of classes, individuals and their relationships. SPARQL query along with Jena will be used to query for the knowledge base. Netbeans IDE will be used to develop Java program for the implementation of the system. Lucene indexing and searching library will be used for indexing and searching of Afaan Oromo

documents. Detailed discussion of tools employed for this work will be presented in Section 5.2.1

Corpus collection

Corpus written in Afaan Oromo text on sport domain will be collected from various sources. The corpora will be employed as a source to extract key concepts of the domain and conduct test performance of the developed system on Afaan Oromo information retrieval system.

Prototype Development

Prototype of ontology aided Afaan Oromo text retrieval system will be developed and tested to verify performance ontology based information retrieval system as compared to keyword based retrieval system.

1.7 Application of the Results

Developing ontology on different domains of Afaan Oromo will help in the realization of semantic based information retrieval system of the language. Afaan Oromo ontology will be employed during the query processing stage to retrieve users' query equivalent of domain concepts that can better be understood by the retrieval system. The developed ontology can be integrated into Afaan Oromo text retrieval system to enhance the retrieval performance of the system. Similarly, documents can be annotated using concepts derived from ontology to aid and improve performance of retrieval system.

1.8 Organization of the Rest of Thesis

The rest of this thesis report is organized as follows:

In Chapter Two, literature review of research works conducted by different authors focusing on the area of information retrieval, ontology, and ontology based information retrieval will be done. In Chapter Three, a brief discussion of related works done on keyword based information retrieval, ontology based information retrieval of Ethiopian and English languages will be done to grasp better understanding of the status of current researches so that it will be used as a takeoff reference for this work. In Chapter Four, implementation of Afaan Oromo ontology on sport domain including the architecture of

the proposed system that shows an overview of the framework of the developed system will be discussed. Chapter Five is about the experimental work done, the implementation of the system, evaluation of the developed system and discussion on the test result of the system will be done. Finally, conclusion and recommendation for the future work will be given in Chapter Six.

Chapter Two: Literature Review

2.1 Introduction

In this chapter, a literature review will be conducted to have a brief overview of Afaan Oromo languages, have an understanding on key concepts residing in information retrieval system which includes information retrieval process that shows simplified representation of steps involved in information retrieval process, a brief discussion on the steps involved in a keyword information retrieval system, models that is used as a blue print in the information retrieval process, term weighting and similarity measuring mechanism to judge and indentify the usefulness of a given word in representing a document, performance evaluation metrics to grasp understanding on how to evaluate performance of a given information retrieval system, ontology which is a means of concept representation, types of ontology, its development process, development methodology and specification language and finally, how ontology will be used in information retrieval process to alleviate drawbacks of keyword information retrieval systems.

Afaan Oromo Language

Afaan Oromo is one of Ethiopian languages. It is categorized under Cushitic language, which is a family of Afro Asiatic languages. Most native speakers of this language are people living in Ethiopia, Kenya, and Somalia [8]. Besides the native speakers, a considerable number of people from other ethnicities who have a direct contact with Oromo people speak the language. For example, the Omotic speaking Bambassi and the Nilo-Saharan-speaking Kwama in northwestern Oromia region are among the people of other ethnicities who speak the language [17]. And hence, it can be considered as one of the most widely spoken Cushitic language family [18].

Afaan Oromo is a working language and medium of instruction in elementary and junior secondary schools of Oromia regional state. It has been serving quite a large number of communities as a language of communications transmitted via television, radio, newspaper, and other media published or aired from both local and foreign countries. Afaan Oromo is a phonetic language, which is spoken directly in the way it is written. That means there is no

character that is skipped or unpronounced during reading of Afaan Oromo word unlike English or other Latin languages [17]. It uses 28 Latin characters (Roman alphabet) called "qubee" adopted in 1991 [18].

2.2 Information Retrieval

An information retrieval system is a software programme that stores and manages information on documents, often textual documents but possibly multimedia. The system assists users in finding the information they need [19]. Users connected to the network can query whatever they need and get response with the assistance of search engine within a short period of time. The search engine retrieved ranked documents as a response to user's information quest, and it is the responsibility of the users to read and select documents that best suit their interest. The information retrieval system does not explicitly return information that gives complete answers to users' query delivered to the system. Instead, it informs on the existence and location of documents that have potential answer and might contain the desired information. Some suggested documents will hopefully satisfy the user's information need [19].

The information retrieval systems' retrieval performance can be evaluated by recall and precision. Recall and precision are used as a tool to assess performance or success of a given retrieval system. Different retrieval models or designs can have different performance. The focus of quality of Information Retrieval design is in evaluating both retrieval effectiveness and efficiency [20]. Efficiency is about optimizing computing resources such as the needed storage space and time complexity while effectiveness is concerned with relevancy of document retrieved that satisfies users' information need [8]. The relevancy of the retrieved documents to the user query or information need is one of crucial measures for the quality of the IR system.

Generally, documents that satisfy users information query is called relevant where as those documents that do not satisfy users' information need are considered to be irrelevant. Researches show that there is no perfect information retrieval system that retrieves all relevant documents and rejects all irrelevant documents [21]. Perfect retrieval systems do not exist because search statements are necessarily incomplete and relevance depends on the

subjective opinion of users [19]. Different users who deliver the same query to the retrieval system may judge the same retrieval result differently. As a result of this subjectivity and user opinion dependent nature of information retrieval, it is not possible to judge that certain retrieval systems are perfect and can perfectly satisfy users' information need. Hence, information retrieval systems are designed on the principle of best match and the accuracy of the system is not expected to be perfect.

Similarly, Hersh [22] described that relevance depends not only on the query and the collection but also on the context, e.g., the user's personal needs, preferences, knowledge, expertise, language, etc. According to Baeza-Yates and Ribeiro-Neto [6], finding text that satisfies users' information need is not a simple task since user's information need is a vague concept, that doesn't remain constant. Documents that are retrieved and satisfy user's information need on one day may not be considered as relevant on other day. Even though a user may query the web with the same keyword, what the user desires to see today may not be similar to the other day. A user whose search keyword is about "American presidential election debate" may not be satisfied with the retrieval of the same documents on different days.

Users may decide to search for information when they want to locate certain information or feel that they have knowledge gap about an issue at hand even without having a clear picture of what they are looking for [6]. This indicates the possibility for the users' to have limitation in selecting the appropriate keywords that enable them query and come up with relevant solution for their information need. According to Ofer *et al* [23], the keywords chosen by users were often different from those used by the authors of the relevant documents, and hence, lowering the system's retrieval quality. This shows that deficiency of knowledge in the subject area is one of the barriers that hinder to generate query keywords that enable to adequately describe the context of the information for better performance of the retrieval system.

2.2.1 Information Retrieval Process

In simplified overview of Figure 2.1, information retrieval process has components like query interface where users deliver their information need to the system. Matching is

performed by a software program that accepts query from an input system, search documents from document repository and retrieves documents that best match the query [2]. In the end, the retrieved documents are displayed to the disposition of the user.

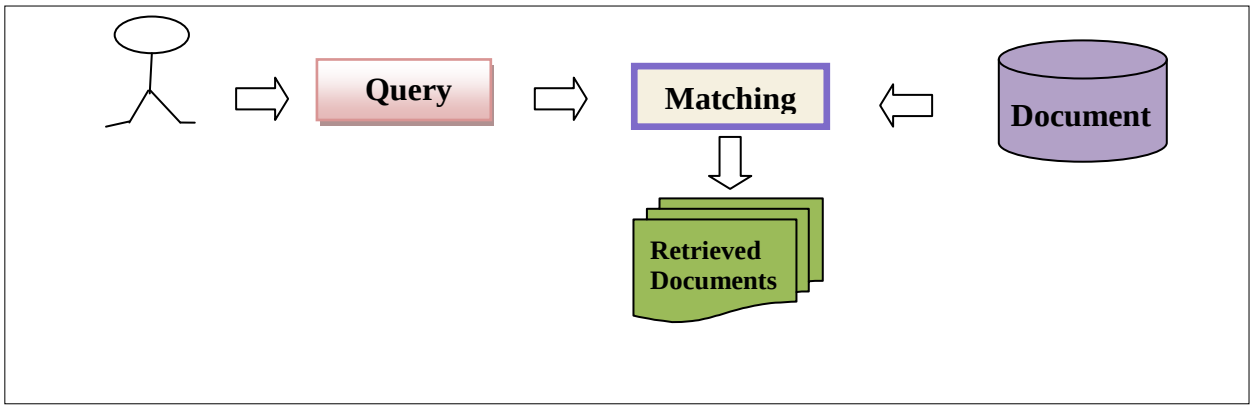


Figure 2.1 Simplified representation of Information Retrieval process

When the query is entered to the user interface and the searching process is triggered by user, information retrieval system computes similarity between query keywords and document keywords and returns those documents that are expected to contain a pieces of information mentioned in the users' information need [12].

The retrieved documents may be presented to the user in a manner that reflects the importance of the documents. The system of delivering retrieved documents to the users sorted based on the perceived importance of the document to a given query is called ranking. For two retrieved documents, one is ranked on top of the other means that it is more likely to satisfy the query than the other [24].

Whether documents returned by a given retrieval system are delivered to users ranked or not is determined by the information retrieval model under consideration. Some information retrieval systems are designed in a way that the retrieved documents are not ranked. On the other hand, there are information retrieval models that incorporate ranking algorithm into their retrieval systems and conduct the necessary computations to deliver documents ranked based on their importance to the query [8]. Ranking has the advantage of sorting documents based on their relevance and helps users to focus only on documents that are retrieved on top of the list.

Nowadays, conventional search engines are the widely used systems for searching and retrieving information on the web. They are designed mainly to function based on keyword matching techniques, the principles of which is based on matching of query terms to keyword with which the document is indexed [11]. Figure 2.2 shows the common architecture of classical information retrieval system from query delivery to obtaining a list of retrieved documents. A formal query keyword is generated from user query after the necessary query pre-processing steps have been undertaken. Query pre-processing steps involve removal of non-content bearing words and other transformations like case folding and stemming. A similarity measurement between the query and the document index has been undertaken by the system to decide on the selection of documents that have potential answer to the query [11]. Based on the result of this evaluation, documents are sorted and delivered to the user.

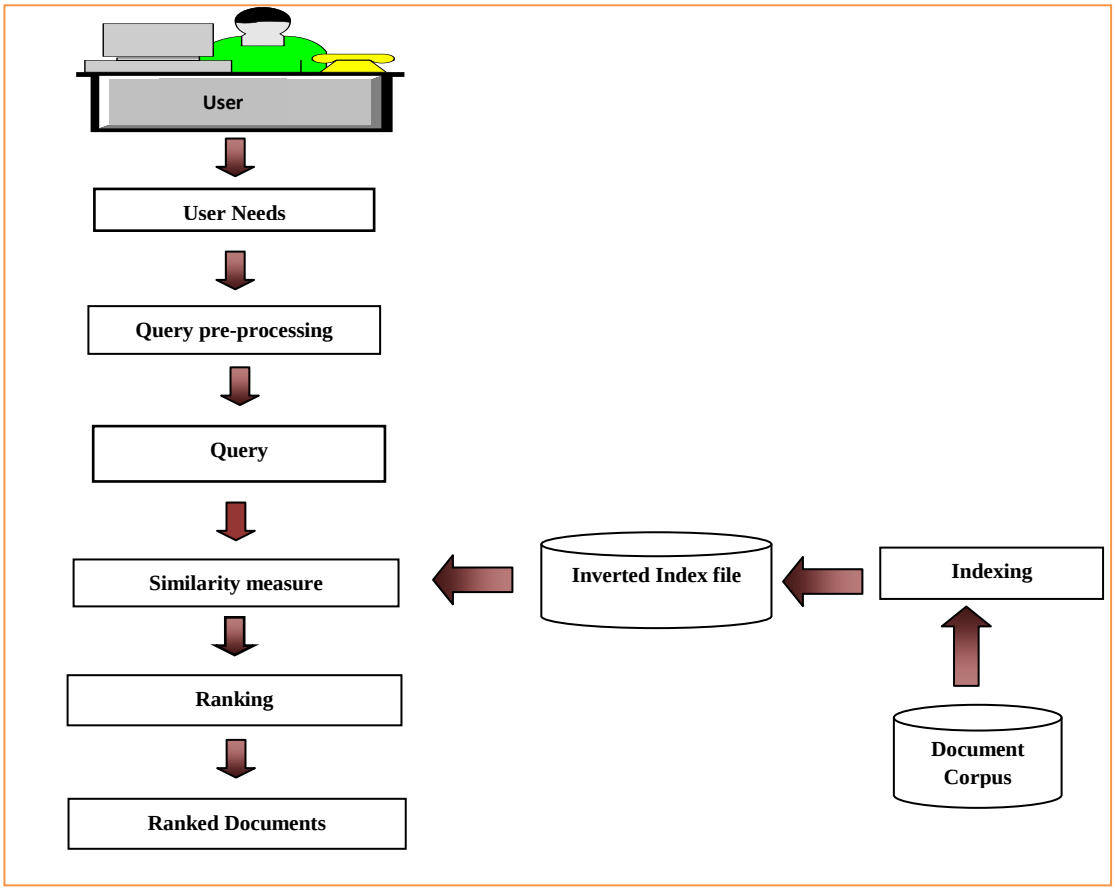


Figure 2.2 keyword-based Information Retrieval Architecture

User Needs - is the intent of the information seeker expressed in textual form. Information seekers input their information need into the user interface.

Query Pre-Processing: is the process of manipulating users' query input to generate query terms that facilitate the comparison of query terms to the document representation. It is the preliminary stage of an IR system that involves applying a set of well-known techniques to convert users' input to a more refined and concise format for the comparison stage [25]. The query terms are generally pre-processed by the same algorithms as document pre-processing. It involves stop words removal, tokenization, case-folding/ normalization and stemming.

Stop words are those words that are common to all documents and have little or no importance in discriminating one document from another. During query and document pre-processing, these extremely common words which appear to be of little value in helping to documents matching are excluded from the vocabulary [20].

Tokenization is the process of chopping a sequence of characters into word pieces called tokens usually based on white spaces. Meanwhile, certain characters such as punctuation marks are also removed [26].

Normalization is the process of converting all texts into similar cases; for example into lower case or upper cases. Token normalization is the process of generating tokens so that matches can occur despite superficial differences in the character sequence of the tokens [20, 26].

Stemming is the process of deriving root words by removing inflectional and derivational morphology. It transforms inflated words into their most basic or root forms [26]. Stemming has both advantages and disadvantages. Its advantage is that it increases the retrieval effectiveness by enabling the retrieval of words of similar stems together. The disadvantage is that, some terms might be over stemmed and changes the meaning of the terms in the document. That means, different terms might be reduced to the same stem which enforce the system to retrieve non relevant documents [27].

Query- Query is a pre-processed users' information need. Query terms should be generated prior to the searching operation is started. It is a string of texts that is representative of the

information that the user is seeking for. Each word of the query is called a search term or query keyword [11].

Similarity Measure - is the process of matching query terms with the representation of the documents or document keywords. It is the process of looking for documents that contain the specified information from within the search space [8, 11, 26]. The result of this stage is a set of hits.

2.2.2 Information Retrieval Models

Information retrieval models are blue prints or guidelines that give directions in the actual implementation or realization process of the information retrieval system [28]. An information retrieval model predicts and explains what a user will find relevant, given the user query [21]. It characterizes the comparison and retrieval mechanism of query delivered by users with bag of words representing the documents.

There are different types of Information retrieval models which are used to specify details of document representation, query representation, matching function [8] and abstraction of the retrieval task [29] as a whole. The most widely employed information retrieval models are the Boolean model, Statistical model (Vector space model, Probabilistic model), and the Knowledge-based (ontology-based) model [8, 11, 12]. Each of these models has its own advantage and disadvantage.

a) Boolean Model

Boolean retrieval model is a model for information retrieval in which we can create query in the form of a Boolean expression of terms, that is, in which terms are combined with the operators AND, OR & NOT [20]. AND represents logical product, OR represents logical sum and NOT represents logical difference.

Documents in the Boolean model are represented as index terms, and queries are represented as a Boolean expression of terms [11]. Its retrieval technique is based on the assumption of exact matching principles in which documents are considered as either relevant or not relevant. According to Gudiva *et al* [29], documents are considered as relevant when the Retrieval Status Value (RSV) is evaluated to 1. RSV is a measurement whose value shows

the status of query-document similarity. The document is considered to be irrelevant for the given query, when the RSV is evaluated to 0. There is no partial matching in Boolean model, relevant documents are retrieved and those found irrelevant are not.

As compared to other retrieval models, the Boolean model has granted experts and users with the opportunity to manoeuvre the retrieval process [21]. They have the authority to decide what should be or shouldn't be retrieved by the system. It is very precise and users get what is specified. It is still widely employed in small searches like searching emails, files from local drives and mid-sized library. In contrary to this, when compared to other retrieval models, this model misses many relevant texts whose representation mismatch the query only partially, there is no ranking of retrieved documents, and it requires complicated query logic formulation [30]. Above all, users may retrieve very few documents or not at all even if there are documents that could potentially satisfy users' information need. In contrary, users may receive a junk of unranked documents that could create difficulty in managing the retrieved documents. In addition to this, managing retrieval process of Boolean model requires good domain knowledge to form good queries [13].

b) Statistical Model

The delivery of unranked documents to users in Boolean model is alleviated by the invention of Statistical retrieval model. Luhn [31], suggested the need of first preparing a document representation similar to the document collection in order to search for the document. The degree of similarity between the query representation and the document searched is used to rank the retrieved list of documents. The more similar the two representations are, the higher the chance for them to represent similar information.

These models employ statistical information of term frequencies of documents to judge the relevance of documents returned as a response to user information need. Documents returned by the system are delivered ranked so that users can prioritize which documents to choose based on the position in the list. Ranked retrieval will hopefully put relevant documents toward the top of the list and hence, minimizing the time the user has to invest in reading the document [21]. Statistical model comprises of Vector Space Model (VSM) and Probabilistic models.

Vector Space Model

The Vector Space Model (VSM) transforms any given text (article, query, portion of an article) into a vector in a high dimensional vector space. The main power of this model comes from its ability to measure the proximity between any two vectors, i.e. the closeness between any two vectors [32].

Vector Space Model (VSM) represents each textual document as a set of terms for the purpose of document representation and indexing. These terms are extracted from the documents after text pre-processing has been performed on the terms. A weight is assigned to each of these terms to determine their relative importance with respect to their discriminatory power to distinguish one document from another. It is an estimate of its usefulness for distinguishing the given document from other documents in the same collection [33]. A given term may also receive different weight in different documents and hence, may better discriminate one document than another. A term receives a weight of zero if it doesn't appear in the document; otherwise it is assigned a non-negative numeric value.

Probabilistic Model

In information retrieval (IR), probabilistic modeling is the use of a model that ranks documents in decreasing order of their evaluated probability of relevance to a user's information needs [33]. It is based on the principle which takes into account that there is uncertainty in the representation of the information need of the users and the document. The probabilistic model of information retrieval starts by guessing the probability that an index terms in a query will show up in a set of retrieved documents and then it uses a recursive process on the retrieved documents so as to improve the result. Retrieved documents are ordered and presented to users in a decreasing rank which were initially ranked based on the estimated probability of relevance to a given information request of the users.

The probabilistic models of information retrieval in its pure form have been implemented only in small scale information retrieval tasks like library catalogue search [20]. Generally, when its performance is analysed, models like vector based model out-performs probabilistic model in large scale information retrieval tasks like web search [6].

According to Baeza-Yates and Ribeiro-Neto [6], the main reasons for the low retrieval performance of probabilistic model as compared to vector space model are probabilistic IR in its pure form incorporates too many assumptions many of which could have side effects. Some of the assumptions upon which the probabilistic model of IR are built include the binary independence model which considers the presence or absence of terms assumption in its probability judgement process, the exclusion of term frequency from its assumption, consideration of each document as a bag of index terms which (i) disregards the sequence in which terms appear in the document, (ii) independence of all index terms from each other and (iii) ranking of some documents does not affect the relevance judgement of other documents are some of the assumptions.

c) Knowledge-based Model

As described above, in classical keyword based text retrieval system, the searching algorithm is based on the matching of keywords. This method of information retrieval has limitations on capturing all relevant documents and also there a probability of retrieving unrelated documents which might not have relevance with the required information.

The contents and services of the current Web use formats such as HTML that can be understood by humans, but not by machines [11]. Some of the data presented on the web is only understood and interpreted by humans and computers cannot comprehend the context of the information represented in it. A software program can assist human being in pointing to a specific Web site and may build, transport, process and provide information, but it does not know what this information means, and therefore, its ability to perform autonomous actions is very limited [11]. From this argument, it is possible to conclude that the current web system has drawbacks in comprehending the semantics of information given in a certain documents.

These and the other limitations of the current web can be resolved by the inclusion of semantically enhanced information retrieval mechanism into the document retrieval process. This means, the challenge of the current web can be solved by indexing documents according to context rather than keywords [34] which can be implemented by an explicit

description of information items, and the overall structure of contents and services of the web.

2.2.3 Term weighting and Measuring Similarity

Term Weighting

Term-weighting schemes assign weights to keywords based on how useful they are in identifying the topic of a document. It is one of the most crucial aspects in relation to the performance of Information Retrieval systems [25]. The usefulness of the document for a given query is hence predicted based on the measurement of similarity of words. For a given word to be considered as a good predictor, it has to first fulfil the criteria of the predictor word. Not all words in the query and documents are equally important to identify how much a given document is likely to answer a certain information need. During retrieval process, documents with more occurrences of meaning bearing terms have got higher score. The score of these terms are summed up to give the score of the documents. Documents are delivered to users ranked based on the aggregate value of the score of the terms for each document. The importance of terms for representing documents is judged based on the evaluation of parameters like term frequency, inverse document frequency and length normalization [35].

Term Frequency (TF)

Term frequency implies the number of times a given term appears or mentioned in the document. A document in which the query term appears or mentioned more times is assumed to be more meaningful and relevant to the users' information need than documents in which it appears less number of times [20].

Inverse Document Frequency (IDF)

Common words that appear frequently in most documents are less useful in discriminating documents as compared to words that appear in few documents. Hence, matching query and document entirely based on the term frequency is not sufficient for good information retrieval system [8]. In addition to term frequency, we need to employ other parameters such as inverse document frequency (idf), to predict documents that suit the required information.

Inverse document frequency aims to measure how significant is the presence versus the absence of a given term to discriminate documents from each other in the collection [11]. It is a mechanism for attenuating the effect of terms that occur too often in the collection to be meaningful for relevance determination [20]. It is employed to scale down the weight of terms with high collection frequency.

Inverse document frequency (idf) is computed as (1):

$$\text{idf} = \log \frac{N}{n} \qquad (1)$$

Where:

idf -Inverse document frequency of a given term

N – Number of documents in the entire document collection

n – Number of documents in which the term t appears.

From the above calculation, a term that appears in all documents will have an inverse document frequency value of zero. Similarly, the value of terms which appear in most documents will have low value.

Term Frequency-Inverse Document Frequency (tf_idf) weighting

To cancel out the effect of common words that appear frequently in almost all documents from not affecting the retrieval effectiveness, term frequency is combined with inverse document frequency, which proposes that terms that appear in almost all documents have little significance to discriminate documents [8]. Tf_idf is calculated using (2) as follows:

$$\text{tf\_idf} = \text{tf}_{t,d} \times \text{idf}_t \qquad (2)$$

Where:

tf_idf - term frequency-inverse document frequency

$\text{tf}_{t,d}$ - term frequency of a given term t in a document d

idf_t - inverse document frequency of a term t

Normalization of Term Vectors

Term frequency is taken into consideration to determine the weight of terms that represent a document and aids in judging the relevance of a document that could potentially satisfies

users' query. Frequency of a given term could be high in large documents and low in small-sized documents. This means that a potential keyword could be mentioned many times in large documents as compared to the smaller ones. Even if a given query keyword is mentioned more number of times in larger size of documents as compared to the shorter one, it doesn't necessarily mean that it better satisfies the information need of the user. According to Singhal [32], long documents usually use the same terms repeatedly. As a result, the term frequency factors may be large for long documents. This increases the average weight of terms in the long documents, which in turn increases the contribution of long document's individual matches (to a query) towards the query-document similarity resulting in a high overall similarity. As a result of this, longer documents could be ranked higher than small-sized documents irrespective of their importance to the given query.

Different normalization techniques are employed in the information retrieval process to cancel out the effect of length variations. Length or maximum term frequency normalization is the most commonly used normalization techniques [21]. Following this criteria, the weight of a term i in a document j is defined as in (3) as follows:

$$w_{i,j} = tf_{i,j} \times idf_i = \frac{freq_{i,j}}{\max freq_{i,j}} \times \log \frac{N}{n_i} \qquad (3)$$

Where:

Wi,j – is weight of a term i in a document j

N - total number of documents in the system

ni - number of documents where the term ti appears

$freq_{i,j}$ - frequency of term ti in the document dj

max $freq_{i,j}$ - maximum frequency of any term ti in the document dj.

Similarity Measure

Similarity between two texts increases with the increase in the number and importance of matching terms. The more word match between two texts the more similar they are. According to Singhal [32], two texts are semantically related if they share same vocabulary; the more vocabulary they share, the stronger is their relationships. This suggests that the measure of closeness should increase with the number of word matches between two texts.

Similarity between the query and the document is judged by computing numeric similarity between them. In a ranked set of documents, the highest ranking document is the one that is most similar to users' query or information need and the similarity decreases accordingly. The similarity between two documents is a function of the angle between their vectors in the term vector space [36] as shown in (4)

$$\text{sim}(\text{di}, \text{q}) = \cos(\theta) = \frac{\text{di}.\,\text{q}}{|\text{di}||\text{q}|} = \frac{\sum_{\text{j}=1}^{\text{n}} \text{w}_{\text{i,j}}\, \text{w}_{\text{q,j}}}{\sqrt{\sum_{\text{j}} \text{w}_{\text{i,j}}^2}\ \sqrt{\sum_{\text{j}} \text{w}_{\text{q,j}}^2}} \qquad (4)$$

Where:

di - is document di in the document space

q - is query

wi,j - weight of term j in document i

wq – weight of query

θ - is the angle between q and di in the document vector space

sim(di,q) = 1 means the two documents are exactly the same and

sim(di,q) = 0 means the two documents doesn't share anything in common

2.3 Evaluation of Information Retrieval System

There are variations among different information retrieval systems on the effectiveness of retrieving relevant documents. Performance of information retrieval system is judged by how much the system is good in delivering relevant documents. Relevance is an inherently subjective concept in the sense that satisfaction of the human need is the ultimate goal, and hence the judgment of the human user as to how well retrieved documents satisfy their needs is the ultimate criterion of relevance [4, 22].

Baeza Yates and Ribeiro Neto [6] classify evaluation of information retrieval system into three categories. Evaluation that deals with the functionalities of the IR system is called functional evaluation; the second evaluation category is concerned with the system's performance evaluation that evaluates the system against its response time and space requirements. The shorter the response time and the smaller the space needed is considered

as a better system. The third category is about the retrieval performance of the system. Retrieval performance of the system analyses the degree to which the users' information request is satisfied by the retrieval system.

The most commonly used retrieval performance evaluation metrics are Precision and Recall [37]. Precision is defined as the ratio of the number of relevant documents that are retrieved to the number of retrieved documents or the probability of retrieved items to be relevant (5). Whereas recall is the ratio of the number of relevant documents that are retrieved to the total number of relevant documents in whole collection of the document space or the probability of relevant items to be retrieved (11). Ideal IR system would like to achieve maximum precision and recall. In reality, since it is not possible to have such system, one must strike a compromise. Indexing terms that are specific yields higher precision at the expense of recall. Similarly, indexing terms that are broad yields higher recall at the cost of precision. For this reason, most of IR systems effectiveness is measured by the precision parameter at various recall levels [29].

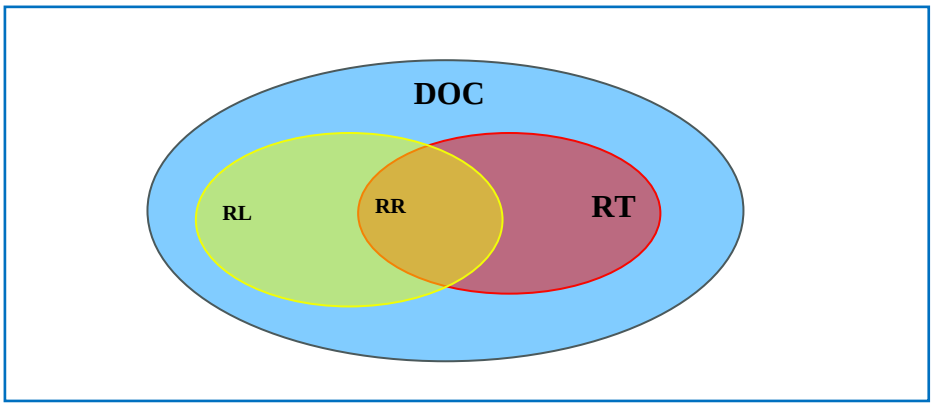


Figure 2.3 Diagrammatic Representation of document retrieval outcome

Where:

DOC - Collection of Documents in the entire search space

RT - Retrieved documents but not relevant

RR - Relevant and Retrieved documents

RL - Relevant to the given query, but not retrieved

If one retrieves all the documents found in the collection, all the relevant documents are retrieved, and hence the recall becomes perfect. Even though retrieving all the documents in the collection maximizes recall, it might not be helpful for users to find what they are looking for. In the situation where only a small proportion of the documents in a collection are relevant, retrieving everything will give a very low precision.

Precision determines the effectiveness with which the system achieves accuracy. Precision is 100% if no irrelevant document is retrieved. The usual reasonable assumption is that the user wants the best achievable combination of good precision and good recall, i.e., ideally the user would like to retrieve all the relevant documents and no non-relevant documents [25].

It is possible to measure the performance of an IR system by both recall and precision for each documents returned. This computation can be done by Precision – Recall graph. In the Precision – Recall graph, precision is usually indicated on the y-axis and the recall is on the x-axis. In general, there is an inverse correlation between values of precision and recall. As precision goes up recall goes down, and vice-versa, and hence precision-recall curves are expected to have a downwardly decreasing slope or have a negative slope [37]. It can be employed for the comparison of different retrieval systems. Retrieval system whose precision-recall curve is above the other or lies to the right as compared to the other is considered to have a better retrieval performance.

$$Precesion = \frac{|RR|}{|RT| + |RR|} \quad (5)$$

$$Recal = \frac{|RR|}{|RL| + |RR|} \quad (6)$$

Some users attach greater importance to precision than recall and vise-versa. There is a situation in which some information seekers target to retrieve as much relevant documents as possible; there are also some users who want to have only some of the relevant document. Hence, the relative preference of precision over recall and vise-versa is user dependent. With this in mind, Rijsbergen [38] came up with information retrieval performance evaluation metrics that takes into consideration both precision and recall as shown in (7).

$$E \approx 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\frac{1}{R}} \quad (7)$$

Where:

E- stands for Effectiveness,

R - Stands for Recall and

P - Stands for Precision,

α - is a parameter which varies from zero (when user attaches no importance to precision), through half (user attaches equal importance to precision and recall), to one (user attaches no importance to recall).

Measuring precision is simpler as compared to recall. If the size of retrieved documents is small, a competent user who can judge relevance of documents for the information need can compute precision of the retrieval system by taking ratio of the number of relevant documents that are retrieved to the total number of documents retrieved. However, for the case of recall, it is much more difficult. It depends on knowing the number of relevant documents in the entire collection, which means that all the documents in the entire collection must be assessed [39].

When talking about information retrieval, users are now experiencing huge difficulties in finding precisely what they are looking for among tons of documents available [40]. Under this circumstance, users have to pick up some documents and go through it in order to locate the information they are looking for and judge for its usefulness. The user will get a document or a set of documents and will have to analyze the documents to find the desired information if it exists [41].

Two terms can be lexicographically different but can have the same meaning. That means they could be synonyms and represent the same content or represent semantically similar concepts. Similarly, different words may represent the same class in some hierarchical categorization and it could be more meaningful for users to retrieve concepts of the same class than totally unrelated items captured through term comparison. On the other hand, totally different concepts may be represented by similar words and may lead to the retrieval of documents that have no correlation with the information need of the users [37]. This leads to the conclusion that information retrieval modeled by mere keyword matching techniques

is not complete enough to deliver adequate information that satisfies users' information need. Retrieval problems associated with keyword matching technique is solved through the introduction of semantic technologies like ontology into information retrieval system.

2.4 Ontology

Ontology is one of the semantic technologies employed to overcome the limitations of word matching retrieval techniques by introducing an explicit descriptions of concepts, individuals and the relationships between them. As of 2000, there is a tremendous increase in the size and complexity of knowledge bases, computing systems and especially the Internet have necessitated the invention of a mechanism that facilitates communication among heterogeneous components. This has paved the way for the application of ontologies in many disciplines of computer and information science including artificial intelligence and database theory [42].

In ontology, since concepts and relationship between concepts of a certain domain are explicitly described, they facilitate interoperability of the heterogeneous components and easy manipulation of the resources on the web. They are used mainly for knowledge representation, knowledge sharing, information retrieval, and knowledge management, and hence, adopted as a central part of the Semantic Web [42].

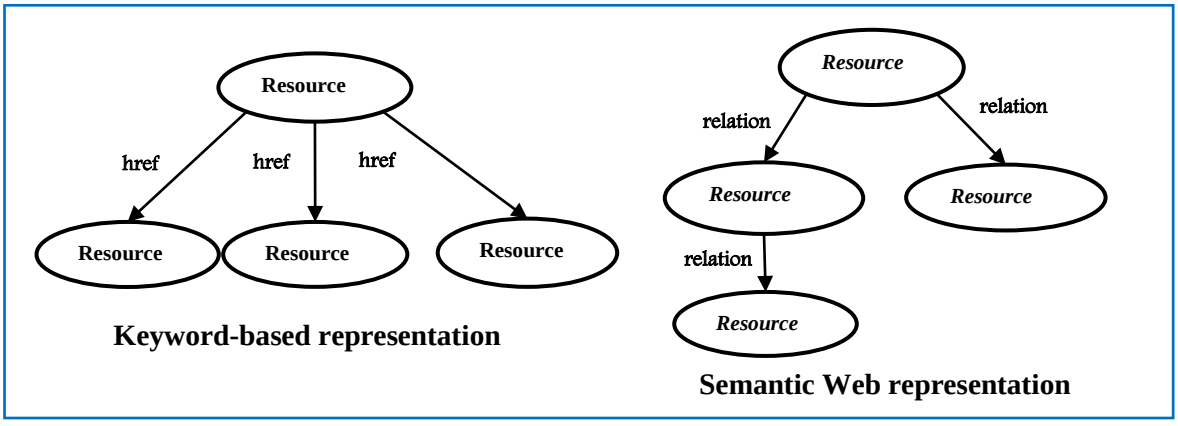


Figure 2.4 Representation of keyword and Semantic Web

Figure 2.4 shows the comparison between keyword (current) and the semantic web pages. In the keyword based web page, the web page content is represented as nodes and the link between the nodes is represented by the hyperlinks. In this representation style, there is no clear description for the computer to identify them [11]. However, in the semantic web representation scheme, the nodes or resources and the relationships between nodes are explicitly stated so that it is simpler for computers to make differentiation among resources and hence enables semantic retrieval [43, 44].

In the context of computer science, different definitions have been given for the concept of ontology. Guarino and Giaretta [45] compiled the definitions of ontology as:

- Ontology is a specification of conceptualization,
- Ontology is a representation of conceptual system via a logical theory,
- Ontology is a vocabulary used by a logical theory,
- Ontology is a meta-level or specification of a logical theory.

The most commonly quoted definition of ontology is from Gruber [46], who defined ontology as an explicit specification of a conceptualisation. Explicitness in this and other definitions of ontology stands to denote that the types of concepts and constraints are thoroughly defined in order to facilitate communication processes especially between humans and computer. Conceptualisation refers to an abstract, simplified view of the world that we wish to represent for some purpose.

Borst [47] also defined ontology as a formal specification of a shared conceptualisation. Formal means that the ontology should be machine-readable and should encompass computer-useable definitions of basic concepts and relationships among them. Shared refers to the fact that the knowledge that is accepted by a group of community or a commonly accepted understanding.

Studer *et al* [48] also defines ontology as a specific, formal representation of a shared conceptualization of a domain. It is specific in the sense that it is used to describe concepts, relations, instances and axioms relevant to a given domain. Formality is to denote machine readability and interpretability of the developed ontology. Shared conceptualization refers that concepts or knowledge have consensus of the members of the community. The

developed ontology provides a common vocabulary, and defines the meaning of terms and relationships between them. The main objective of ontology is to enable communication between humans and computer [49, 50]. It represents concepts in a manner in which the communicating agents can understand the meaning of the contents of the documents.

Ontology is also defined by Uschold and Gruninger [51] as a shared understanding of some domain of interest which may be used as a unifying framework. It is a semantic technology that is mainly exploited to define domain specific knowledge. The purpose of ontology is not to model the whole world, but rather a part of it called domain. It describes the concept that it is created to form some kind of general understanding of the domain in order to bring interoperability between software agents.

Ontology is closely related to concepts such as knowledge base and database schema. However, Van Nguyen [44] states that ontology can be distinguished from knowledge base in that it is a conceptual structure of a domain while a knowledge base is a particular state of a domain. Ontology also separates itself from a database schema in that ontology is sharable and reusable while database schema tends to be specific to the domain and is context dependent.

Basically, ontology consists of a set of classes and relationships that prevail between different objects of classes called instances. Some authors categorize instances under ontology while others do not. For example, Weller [52] states that instances are not necessary considered as elements of ontologies. Ontologies together with the objects or instances of classes form knowledge base.

Classes in ontology are set of objects with certain characteristics in common. It is possible to define a class Person, Student, Instructors, Course etc. Individuals that fall under certain concept have common characteristics, behaviours, or natures that enable us to categorize them together.Class Student for example contains individuals that share common characteristics that are peculiar for individuals categorized as students. Similarly, class Instructor comprises of individuals that have common characteristics and make them unique from another set of individuals.

Relationship is a link that relates or connects one instance of a class with other. Most ontologies include is-a relationship between concepts. They are usually denoted by verbs. For example, in Figure 2.5, “teaches” is a relationship that relates class “Instructor” with class “Course”. “takes” is a relation that relates class “Student” with class “Course”. Similarly, “teaches-At” relation relates the class “Instructor” with the class “University” etc.,

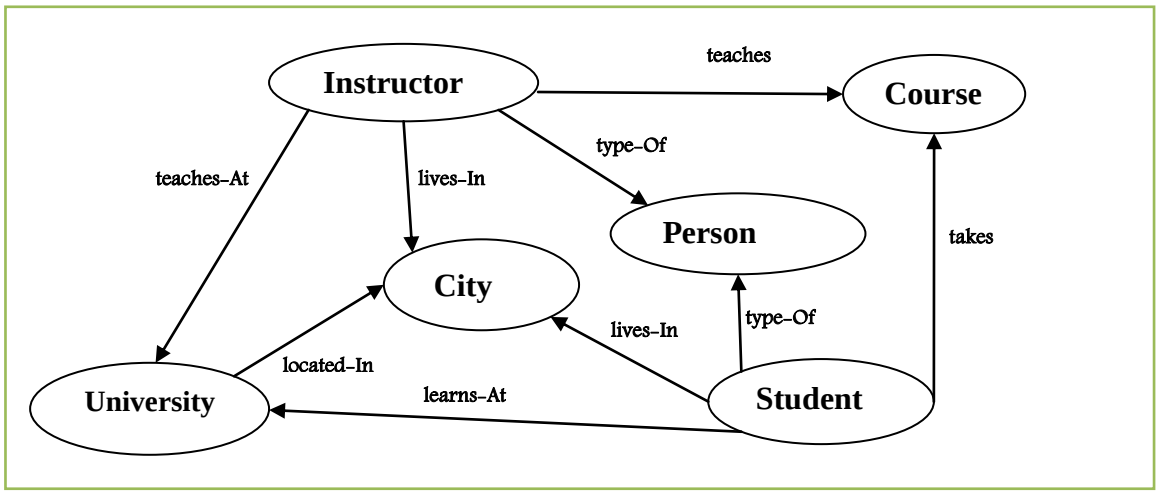


Figure 2.5 Concepts and Relations representation in Semantic web

2.4.1 Types of Ontologies

We can identify different types of ontology based on different characterization or classification parameters. Ontologies can be classified based on granularity, formality, generality and computational capability [53].

Granularity refers to the level of conceptualization of the domain under consideration. They differ on the level of expressive power of the ontology. Based on the granularity, ontology can be classified as Coarse-grained and Fine-grained ontologies [44]. Coarse-grained ontologies facilitate the conceptualisation of a domain at the macro-level, and are typically represented in a language of minimal expressivity. Fine-grained ontologies, on the other hand, allow the conceptualization of a domain at the micro-level, and tend to be represented in a language of significant expressivity.

Formality of ontologies refers to the level at which the contents described in the ontology is formalized. Based on formality, Uschold and Gruninger [51] states that ontologies may be

classified as being highly informal, semi-informal or rigorously formal. Highly informal ontologies are those ontologies that are expressed in natural language while rigorously formal ontologies are those that are defined in a language with formal semantics with a desirable computational property such as soundness and completeness. Semi-informal ontologies are those with an intermediate formality between the informal and rigorously formal ontologies.

Generality with respect to the classification of types of ontology refers to the extent to which the captured conceptualization is described in a detailed or generalized manner. Guarino [54] classifies ontology based on the generality as upper-level ontologies, domain ontologies, task ontologies, and application ontologies.

Upper-level ontologies from the generality categorization of ontologies are ontologies that describe very general concepts like space, time, matter, object, events, actions, etc. Upper level ontologies are also called top level ontologies or foundation ontologies. They are high level ontologies and are not domain dependent [54].

Domain ontologies describe concepts related to the general domain by specializing of the concepts provided under upper level ontologies. Domain specific concepts like sport, Olympic, cricket, education, agriculture, construction, legal, etc., can be dealt with domain ontologies.

Task ontologies deal with the vocabulary describing certain tasks or activities. For example activities like teaching, running, brewing, etc.,

Application ontologies are ontologies that are utilized for specific applications. They utilize both domain and task ontologies, and encompass all the necessary concepts to model the required application.

2.4.2 Ontologies and Other Knowledge Structure

There are various knowledge representation schemes that could be employed to represent concepts in a certain domain, which according to Carola [37], vary in their level of formalization and semantic expressiveness. These are taxonomy, thesaurus, topic map and ontology.

Taxonomy

Taxonomy is one of knowledge classification systems applicable in a certain domain. It consists of categories which are hierarchically ordered in a linear structure. This means that each category is assigned to maximally one super ordinate category, thereby forming a tree structure of category hierarchies [37]. Daconta *et al* [55] define taxonomy as the classification of information entities in the form of a hierarchy; according to the presumed relationships of the real-world entities that they represent. However, no matter how taxonomy classifies concepts in some form of hierarchical relationships, the classification is only based on the type-of, is-a, etc., and doesn't show cause-effect, part-of, association and location relationship. There is also no definition of attributes to terms in taxonomy.

Thesaurus

Thesaurus can be taken as taxonomy with additional features such as equivalence, association, inverse, etc, which shows semantic relationships between concepts. It is a controlled vocabulary developed for document indexing and retrieval. It is a tool designed to support effective information retrieval by guiding indexers and searchers to consistently choose the same terms for expressing a given concept or combination of concepts [56].

Topic Map

Topic Map [37] has more semantic expressive power as compared to taxonomy and thesaurus. It shows association between topics, differentiates abstract subjects, their representation and occurrences. When compared to ontologies, however, the topic maps standard defines only an abstract data model and does not specify its formal representation. The standard hence, does not allow for drawing inferences on the represented knowledge and in contrast to ontologies, no formal query languages are available for querying topic maps based knowledge representations.

2.4.3 Ontology Development Process

The development process of ontology has to pass through a certain general steps. There are seven activities involved in ontology development processes [57, 58].

i) Determine the domain and scope of the ontology

It is the stage of ontology development process at which answers to certain basic competency questions have to be given in order to acquire sufficient knowledge that will help to understand the domain and delineate the scope of the ontology to be determined. We have to answer questions like, for what purpose we are going to use the ontology, what types of questions are dealt with our ontology, who will use and maintain the ontology etc., are some of the basic competency questions one has to give answer and focus on in order to determine the domain and scope of the ontology to be developed [59].

ii) Reusing Existing Ontology

It will save time, energy and other expenses, if it is possible to reuse an already existing ontology that has been developed by others. We can reuse an already existing ontology by extending or customizing to adopt it to our domain of interest. Rather than developing ontology from scratch, it is almost always worth considering what someone else has done and checking if we can refine and extend existing sources for our particular domain and task [49].

iii) Enumerate Terms in the Ontology

It is the list of terms we are going to explain to the users. It is a glossary of terms. It is where one makes a list or graphs of all the nouns and verbs that are going to be considered in the ontology to be developed [60].

iv) Define classes and the class hierarchy

It is the idea of classifying concepts in a hierarchical manner. Some nouns are types of other nouns and can be classified under the other called subclass and super class respectively. A certain concept C1 is said to be subclass of another concept C2 if and only if every instances of concept C1 is an instance of concept C2. But not all instances of concept C2 can be categorized under concept C1. There are several approaches that can be employed for developing class hierarchy [58]. These are top-down, bottom-up and combination approaches.

- **Top-down approach:** It is a type of concept hierarchy development strategy that starts with the general concept and goes down to the specific concepts.

- **Bottom-up approach**: it is a type of concept classification that starts from specific concepts and works up to the general concepts.
- **Combination approach:** it is a type of concept hierarchy development approach that utilizes the combination of bottom-up and top-down ontology development approach

v) Define the properties of classes

Properties define relationships that prevail between different concepts. It includes defining of the relation in detail by giving a name, source concept, target concept, cardinality (how many instances of a concept are related with how many of others), inverse name (you can read from A to B, also from B to A) [60].

vi) Define the facets of the slot or add constraints or allowed values of the properties

These are the value types, allowed values, the number of the values (cardinality) and other features of the values the slot can take [50].

vii) Create instances

An instance is an individual of a class. You can describe in detail relevant instances that may appear by giving them a name, concept to which they are related, attribute names and values [60].

2.4.4 Ontology Development Methodology

An ontological development methodology deals with the methodological aspects that are employed or followed for the realization of ontology. These methodologies provide a set of guidelines, procedures and activities to be followed and undertaken for the creation of domain ontology under consideration. Each development team usually follows its own set of principles, design criteria and phases in the ontology development process. Quite a number of research groups are building ontologies, but it is less clear what design decisions and activities are taken and how they contribute to the success (or failure) of the ontologies developed. The absence of agreed guidelines and methods hinders the development of shared and consensual ontologies within and between teams [61].

So far we have been discussing about the importance of constructing ontologies for reuse and sharing of knowledge etc. The prevailing reality resemble that the development of

ontologies is very much an art rather than a science. This situation needs to be changed, and will be changed only through an understanding of how to go about constructing ontologies. In short what is needed is a good methodology for developing ontologies [62].

In order to support the development of ontologies, several methodologies have been proposed to date, facilitating the process of ontology development or ontology engineering [63]. Depending on the status of the task to be dealt with, one can select one methodology out of the several available methodologies. Despite the fact that quite a number of ontology engineering methodologies have been proposed, still the field lacks widely accepted and mature methodologies. They vary in steps and tasks that they propose an ontology developer should perform when building ontology [64]. In addition to variations in the steps followed, these ontology development methodologies also lack sufficient details of techniques and activities employed in them [63].

The selection of methodology depends on the domain of the ontology going to be developed and application sought. The methodology employed to develop ontology for information retrieval may not be similar to the methodology followed for ontology used in Artificial Intelligence. However, there are some fundamental rules that can help ontology developers to make design decisions in many cases [50].

- There is no one correct way to model a domain. There are always viable alternatives. The best solution almost always depends on the application that you have in mind and the extensions that you anticipate.
- Ontology development is necessarily an iterative process.
- Concepts in the ontology should be close to objects (physical or logical) and relationships in your domain of interest.

The following are the most common ontology development methodologies agreed upon by scholars conducting research in the area.

Uschold and King Methodology

This methodology is based on the experience for developing enterprise ontology [65]. It is applicable to model activities which are typical to companies or a group of companies which are involved in similar activities or business. They divide the ontology development process

into 6 stages which includes identifying the purpose, ontology capturing, ontology coding, integrating existing ontologies, evaluation and documentation.

a) Identifying the purpose is grasping the reason why the ontology is to be built.
b) Ontology capturing is the process of capturing the key concepts and relationships in the domain of interest, generating precise and unambiguous vocabulary of terms.
c) Coding is representing the knowledge captured in a formal ontology specification language
d) Integration refers to reusing the existing ontology
e) Evaluation is to make technical judgement of the ontologies, their associated software environment, and documentation with respect to frame of reference [66].
f) Documentation refers to guidelines for documenting the ontology.

Gruninger and Fox

Gruninger *et al* [59] propose an ontology creation methodology based on the business domain. It was proposed based on the experience of creating TOVE project ontology [67]. The methodology first focuses to capture the ontology requirements by means of informal description. This informal description is transformed to formal language, which is a computable model expressed in first order language. Alike the Uschold and King's methodology, this methodology has limitation in describing activities and techniques used [61].

Methontology

Methontology is an ontology creation methodology that advocates the development of domain ontology. It is employed for building of ontology either from scratch, reusing other ontologies as they are, or by reengineering them [68]. In this methodology, unlike the two methodologies discussed above, it describes the technique used and activities undertaken in detail. It includes development oriented activities like specification, conceptualization, formalization, integration and implementation [63]. In Methontology ontology development methodology, there are activities that are considered as support activities in addition to those developments oriented activities. They include activities like knowledge acquisition, evaluation, integration, and documentation.

CYC Methodology

Lenant and Guha [69] describe that the CYC Methodology is emerged from the experience of developing the CYC knowledge base. According Fernandez-Lopez [61], the CYC methodology is based on three phases. The first phase requires manual coding, the second phase proposes knowledge codification aided by tools and the third phase relies majorly on the tools for work requiring little human intervention.

SENSUS

SENSUS ontology development follows a top down approach for deriving domain specific ontologies from huge ontologies, with 1) a series of terms are taken as seed, 2) these seed terms are linked by hand to SENSUS, 3) all the concepts in the path from the seed terms to the root of SENSUS are included, and 4) for nodes with the large number of paths, the entire sub tree under the node is added sometimes are some of the steps followed [70] in this methodology.

2.4.5 Ontology Specification Languages and Tools

Ontology specification languages

For ontologies to be understood by computers they have to be represented in some form of computer readable and understandable languages. They are modelling web languages that have been developed to represent or express ontology [71]. Ontology languages are used to add semantics to the information we represent on the web.

Ontology specification languages can be distinguished based on their power of expressiveness. Languages based on higher-order logics are more expressive than those languages based on first-order logics, which in turn are more expressive than languages based on description logics [44] Languages with higher level of expressiveness allow more complete representation of knowledge and better reasoning capability as compared to the languages with low level of expressiveness. Since they vary in their expressiveness, the choice of one language from the other is determined by the application the ontology is aimed for. Most of these languages are based on XML Syntax, but they have different

terminologies and expressions. Indeed, some of these languages have the ability to represent certain logical relations which others do not [71].

There are two commonly used knowledge representation formalisms in most of prevailing knowledge representation systems. These commonly used knowledge representation formalisms are named as description-logic based and frame-based systems. These two formalisms are easily distinguished from one another by the languages used to create the ontology. Description-logics are a family of knowledge representation languages that can be used to represent the knowledge of an application domain in a structured and formally well-understood way [72, 73, 74]. Description logic-based systems (DL systems) provide users with highly optimized reasoning procedures that have been designed with the objective of being implemented in automated reasoning system [75]. The adoption of OWL as the ontology specification language for the semantic web is perhaps the most notable success of this particular language paradigm. In the frame-based systems, ontologies are represented by frame-based languages in which the modelling entities are frames and slots [43]. In this system, frames represent concepts whereas slots represent attributes associated with the concepts.

Generally, ontology specification languages can be classified as traditional and web based ontology specification languages [50]. According to Taye [71], traditional ontology specification languages are those languages based on first order predicate logic (KIF, CYCL), frame-based languages (ontolingua, F-logic and Operational Conceptual Modelling Language (OCML)), and description logic based languages (LOOM) etc. Web based ontology specification languages includes OIL, DAML+ OIL, XOL, SHOE, and OWL [76].

eXtended Markup Language (XML) is a mark-up language that separate web content from web presentation. Even though it has overcome the problem of presenting data from its presentation as in HTML, lack of semantics is considered as one of its major draw backs [77].

RDF (Resource Description Framework) is a W3C standard and semantic network based language for describing resources on the web. It has an XML-based syntax and designed to be read and understood by computers, and it is not displayed to people. RDF is becoming a

widely recognized language and a representation formalism that can serve as a worldwide Interlingua for information interchanges [78]. RDF model has three elements and is expressed as triplets. These triplets are resource (subject), the object and the predicate. It is expressed as the <subject> has a property <predicate> valued by <object> [71]. Later, RDF is extended to RDFS to address data type definition and incorporating frame-based primitives [79]. The expressiveness of RDF and RDFS is very limited. RDF is limited to binary ground predicates and RDFS is limited to subclass hierarchies and a property hierarchy, with domain and range definitions and these properties [80].

OIL (Ontology Interchange Language) is a full-fledged web based ontology language that is based on RDFS and was developed by European researchers in the On-To-Knowledge project. It permits semantic interoperability and designed for both describing and exchanging ontology [81].

DAML+ OIL is the result of the combination of American-based DARPA Agent Mark-up Language (DAML) with the European-based Ontology Interchange Language (OIL). By doing so a more efficient web ontology language is created as compared to OIL. It includes more features from description logics where as more frame-based features are excluded which makes DAML+OIL difficult to use with frame-based tools [44].

OWL (Web Ontology Language) is a standard ontology language that is recommended by W3C for expressing ontologies for the semantic web. It is an extension of RDFS and is compatible with the early languages such as SHOE, RDFS, DAML+OIL. It provides more power to express semantics and enables automatic reasoners to derive knowledge through logical inferences [78]. OWL was developed by Web-Ontology working group in 2001 and became W3C recommendation language in 2004. It is built based on RDF to overcome the weakness in RDF/S and DAML+OIL. It provides more affluent integration and interoperability of data between communities and domains [71]. There are three species or sublanguages of OWL ontologies distinguished by their power of expressiveness. They are named as OWL-LITE, OWL-DL and OWL FULL.

OWL-LITE is designed for easy implementation and has the least expressive power. It is syntactically the simplest sub-language and intended to be used in situations where only a simple class hierarchy and simple constraints are needed [44, 82].

OWL-DL is much more expressive than OWL-LITE. It is based on Description Logic to represent the relations between objects and their properties. It is, therefore, possible to automatically compute the classification hierarchy and check for inconsistency in an ontology that conforms to OWL-DL [82]. OWL-LITE is a sublanguage of OWL-DL [71].

OWL FULL provides the highest level of expressiveness as compared to OWL-LITE and OWL-DL. It is intended to be used in situations where very high expressiveness is more important than being able to guarantee the decidability or computational completeness of the language [82]. OWL-DL is a sublanguage of OWL-FULL [71].

Ontology development Tools

There are a number of ontology development tools that are used for the creation and manipulation of ontology. They are used to create new ontology from scratch or by reusing existing ones. The ontology editors are tools that allow users to visually manipulate, inspect, browse code, and support the ontology development and maintenance task [83]. These tools can be applied at all stages of the ontology life cycle including the creation, population, implementation and maintenance of the ontology [80]. The most prominent ontology development tools are:

Protégé is the most popular ontology development tool. It is free, Java-based, open source ontology editor developed by Stanford Medical Informatics. Protégé aids users to develop domain ontology and offers two approaches for modelling ontologies; a traditional frame-based approach (via Protégé-Frame) and using OWL (via Protégé-OWL) [44]. Protégé ontologies can be stored in a variety of different formats, including RDF/RDFS, OWL and XML Schema formats. It can be extended via plug-in architecture and Java application programming interface (API) [84].

Apollo is an ontology editor that is implemented in Java and allows a user to model the ontology with basic primitives such as classes, instances, function, relations and so on. The knowledge base of Apollo consists of a hierarchical organization of ontologies and has a

frame-based internal model. It doesn't support graph view, collaborative (multi-user processing) etc., [80].

OilEd is a simple ontology editor that supports the construction of OIL-based ontologies. It is less flexible as compared to other ontology editors like protégé. It lacks many features that enable it to be categorized as a full-fledged ontology editor [85].

OntoStudio is based on IBM Eclipse framework and it is an ontology engineering environment by using graphical means. It is based on client-server architecture in which the ontology is stored on central server and clients can access and modify it from the client end. It supports multilingual and collaborative development of ontology and its knowledge model is related to frame-based language. It uses OntoBroker inference engine that enables it to exploit the strength of F-logic and represent expressive rules [83].

Top Braid Composer (TBC) comes in three editions; Free Edition (FE), Standard Edition (SE), and Maestro Edition (ME). Free edition (FE) is an introductory version with only a core set of feature. SE includes all features of FE plus graphical views, import facilities etc [44]. In Maestro edition there are all features of the SE and having some extra features like SPARQL Motion, Top Braid Live, Ensemble etc., [85]. TBC is based on Eclipse platform and Jena API [44, 83]. TBC is a comprehensive editor for RDF(S) and OWL ontologies and offers a plug-in architecture and has pellet as its built-in reasoner.

Swoop is open source, web-based OWL ontology editor and browser. It has reasoning (RDFS lie and Pellet) support (OWL inference engine) and provides multiple ontology environment, by which entities and relationships across various ontologies can be compared, edited and merged seamlessly. Navigation could be simple and easy due to the hyperlinked capabilities in the interface of the swoop. It doesn't follow a methodology for ontology construction [83].

Ontology Reasoning Tools

A reasoner is a program that infers logical consequences from a set of explicitly asserted facts or axioms and typically provides automated support for reasoning tasks such as classification, debugging, and querying [86]. The most common ontology reasoners are Pellet, RACER, FACT++, Snorocket, Hermit, CEL, ELK, SWRL-IQ, TrOWL and others

which have various inferencing algorithm, supporting logic, degree of completeness of reasoning, implementation language etc., [87].

2.5 Application of ontology for Information Retrieval

As a response to the problems emanated from the drawbacks of classical information retrieval system, many research works have been undertaken to devise a mechanism on how to enrich the information available on the web and make it more informative and enables a better understanding of content of information by computers. This proposed information retrieval paradigm is the one that ensures the delivery of minimal irrelevant information (high precision) and at the same time, guarantee relevant information is not overlooked. This relentless demand for the improved information retrieval techniques led to the inception of the concept of ontology based information retrieval system.

The concept of semantic technology deals with indexing documents according to meanings, although this will entail a way of converting words to meanings. Ontology provides a means of describing word meanings and relationships between them. It is introduced in Information Retrieval System for the purpose of solving the problem of semantic understanding [88]. In addition to semantic indexing, concepts that represent a given domain and encoded in the ontology are utilized during concept mapping between query and document instances. Ontologies allow definition of class hierarchies, object properties and relation rules. Using this knowledge, it is possible to define instances of classes, to associate them with documents, and to make inferences about them [89].

According to Wu and Wang [88], ontology has rich and formal logic-based for specifying meaning of terms and hence, the use of ontology for IR is an efficient method that can be superior to others in both precision and relevance. Through ontology, computers can understand exactly the information need of the user. Due to semantic annotation and comprehension, the semantic understanding of the document is realized. In ontology, the subsequent specification of the concept will create definitive meaning of concept that is understood by computers irrespective of the terminology used [90]. As a result, the search process of ontology-based IR system has excellent reasoning capability which makes the

whole search process intelligent, such as expanding relevant concepts and filtering irrelevant concepts.

Ontology that is employed in information retrieval process should first be developed from the domain of interest by extracting concepts that represent the domain under consideration. It is created using suitable ontology development tools and integrated into information retrieval system. Figure 2.6 shows simplified flow diagram of ontology creation and its usage in information retrieval. In information retrieval process, the ontology could be used during query processing or annotation of documents with the concepts derived from the ontology.

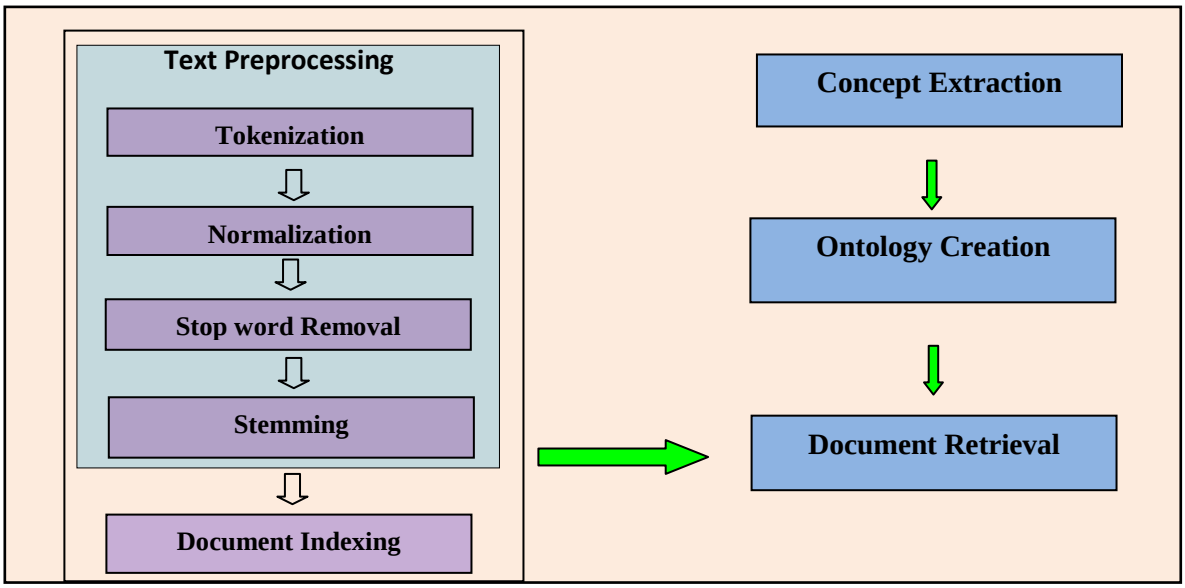


Figure 2.6 Simplified Representation of text pre-processing, ontology creation and inclusion in text retrieval system

Text Pre-processing

It involves tokenization, case folding, stop word removal and stemming. Tokenization converts character streams into tokens [8], case folding converts cases into similar case, stop word removal involves removing non-content bearing words, stemming refers to reducing a word to its root word [91].

Concept Extraction

Ontology formally describes a list of terms which represent important concepts. During ontology development process, extraction of these concepts deals with the identification of terms describing set of objects, events, circumstances [46].

Ontology creation

Ontology creation is the process of populating ontology by the concepts identified at the concept extraction stage. Concepts are captured as a class, and the relationships between them are captured as object or data properties.

2.6 Summary

The objective of information retrieval technology is to enable the retrieval system to retrieve as many relevant documents as possible and to exclude as many irrelevant documents as possible from the list of retrieved documents. The retrieval effectiveness of a given information retrieval system is measured by precision and recall. Precision and recall is affected by factors like presence of synonym, polysemy and other conceptually related words. Various research works show that quality of information retrieval system can be improved through explicitly defining concepts, relationships and an instances of a concept and integrating in to the retrieval system. This is possible through the development of ontology.

Ontology is developed through the application of ontology development tools like Protégé, TBC, and Apollo etc. Ontology development process could be executed in a top-down, bottom-up or combination approaches depending on the concept under consideration. Different Ontology development languages are used to capture concepts, individuals and relationships between them. The choice of the ontology development tools and languages is dictated by the application the ontology is aimed for.

Chapter Three: Related Work

3.1 Introduction

In the review part of related works, review of different research works conducted by different authors on Ethiopian languages like Afaan Oromo, Amharic languages and non-Ethiopian languages mainly English language will be done. Emphasis will be given to information retrieval works done on keyword and ontology based systems and other related works like information selection, latent semantic indexing and concept based document categorization.

Few of the research works done on Ethiopian languages focus on cross lingual information retrieval of Afaan Oromo-English languages [92, 93], cross lingual information retrieval of Afaan Oromo-Amharic [94], keyword based Afaan Oromo text retrieval system [8], Afaan Oromo search engine [9], concept-based automatic Amharic document categorization [95], latent semantic indexing for Amharic text retrieval [16]. Similar research works were done to show the benefit of semantic technology for the retrieval effectiveness of non-Ethiopian languages like English language. Some researches compare performance of ontology-based retrieval system with keyword based, and latent semantic indexing [96]. Comparison of performance between keyword and concept based information selection for broad queries, and narrow queries [97], performance of ontology in query expansion [37], comparison of performance of keyword based text retrieval system with text retrieval system based on ontology developed on different domains [11, 98, 100, 101].

3.2. Information Retrieval of Ethiopian Languages

Kula Kekeba *et al* [92] have conducted a research on a bilingual information retrieval system that involves Afaan Oromo and English texts. In this work, the researchers developed the first Afaan Oromo-English Cross Lingual Information Retrieval (CLIR) system that is based on dictionary for query translation. The objective of the research work was to design and develop Afaan Oromo-English CLIR system with an intention to enable speakers of Afaan Oromo language access and retrieve text documents written in English language by using queries written in Afaan Oromo words. That means a query is written in

Afaan Oromo and the information retrieval system translates the query into English language before the actual retrieval process is commenced. This shows an involvement of machine translation process between the query and the retrieval languages. Hence, retrieval performance of this system could suffer from limitations posed as a result of the inclusion of dictionary in the query translation process. There are some words like names and words borrowed from other languages that cannot be handled by dictionary. There is also a need to update the dictionary to incorporate newly created words. The authors conducted three official runs differentiated with the inclusion of title, description and narration to Afaan Oromo query and have found average precision of 22.0%, 25.04% and 24.5% respectively. Lower mean average precision for title run was found due to very short nature of most titles CLIR.

Similarly, attempts were made to develop CLIR systems that are based on query translation of Afaan Oromo queries. Corpus-based Afaan Oromo-English CLIR [93] which enables to retrieve documents written in English language and corpus-based Afaan Oromo-Amharic CLIR [94] which enables to retrieve Amharic documents using queries written in Afaan Oromo language. Publically available parallel documents like legal documents, religious documents, constitution of Oromia regional state and bible chapters were used in the CLIR. GIZA++ is the word aligner used to build the bilingual dictionary. In the Afaan Oromo English CLIR system, 55 pairs of Afaan Oromo and English texts were used in the experiment and have found a maximum average precision of 0.468. Where as, in the Afaan Oromo-Amharic Cross Lingual Information retrieval system, 50 randomly selected publically available Afaan Oromo and Amharic parallel documents were used for the experimentation. The system was tested with two consecutive experiments. The first experiment was tested using queries obtained through one to one translation and the second was done with all possible translations. Maximum average precision value of 0.8 and maximum average recall value of 0.58 was recorded for the monolingual run and a precision of 0.45 and a recall value of 0.38 results were registered for bilingual run (retrieval of Amharic documents using Afaan Oromo queries). In the second experiment, maximum average recall of 0.7 and maximum precision of 0.6 results were registered.

Cross lingual information retrieval has granted users of Afaan Oromo language the alternative of retrieving documents written in a language different from query languages. Query is delivered in Afaan Oromo language and the system performs query translation and retrieves documents written in English or Amharic language. It has advantage for users who have low language proficiency in delivering queries in English or Amharic languages and enables them to access vast information available on the web written in the two languages. However, this method of information retrieval doesn't benefit those users who cannot understand these languages.

From the principles of information retrieval system, it is important to have a system that works for Afaan Oromo language and then better to extend to CLIR [8]. In addition to this, there is also a possibility of encountering problems related with ambiguity as the translated document does not necessarily represent the sense of the original query.

Implementation of retrieval system that retrieves documents written in a language of query language benefits the speakers of the language to get what they need in their own language. In this respect, Gezahegn Gutema [8] had developed keyword-based Afaan Oromo text retrieval system that retrieves Afaan Oromo texts using queries written in the same language. The author employed Vector Space Model to guide searching for relevant documents from Afaan Oromo text corpus. Different document corpuses were collected and prepared by the author from different domains. Text pre-processing techniques like tokenization, normalization, stop word removal and stemming were employed to pre-process afaan Oromo texts before indexing has taken place. The test result of the developed retrieval system has a performance of 0.575(57.5%) precision and 0.6264(62.64%) recall.

Tesfaye Guta [9] has developed Afaan Oromo search engine. Search engine is software that is used to retrieve information on the web. Afaan Oromo search engine is used to locate and retrieve documents written in Afaan Oromo languages. Some of the components of the developed system of Afaan Oromo search engine are searching and indexing components which are implemented using Lucene. The searching process in Lucene is implemented based on the principle of looking for query words in the document index and finding documents that contain those query words. That means matching of user query and the

document index is conducted based on keyword matching principle. Using this technique an average precision of 57% and recall of 93% was obtained.

In Afaan Oromo text retrieval systems discussed above, Cross Lingual Information retrieval [92, 93, 94], Afaan Oromo Text Retrieval [8] and Afaan Oromo Search Engine [9], matching between query and document representation is made through word comparison. Information retrieval systems operating based on keyword matching principle is suffering from the problem of retrieving documents based on the meaning or concepts of the content of the query and document.

Tewodros Hailemeskel [16] employed Latent Semantic Indexing (LSI) technique to develop Amharic Text Retrieval system so as to overcome the problems of the exact term matching techniques. He used a statistical technique called Singular Value Decomposition (SVD) to implement his proposed method. Latent Semantic Indexing method takes into account the dependencies between terms in the representation of documents and queries through the pattern of co-occurrences of words. The proposed, Latent Semantic Indexing (LSI) for Amharic Text retrieval is able to infer the hidden or latent structure of relationships between documents and terms. The major advantage this method has over the keyword matching techniques is that it uses concepts or topics instead of individual words to index and retrieve the desired document. In this research work, the researcher used 206 Amharic documents which enabled him to come up with 9256 unique non-stop word index terms generated by automatic indexing of documents. 25 queries were used to test the proposed method. Finally, he achieved a better retrieval performance reflected in the form of precision, as compared to the standard vector space methods. He obtained an average improvement of 2.42% precision using the proposed LSI method. The average precision obtained for the LSI approach is 71.57% as compared to the 69.13% for standard vector space method.

Research works done on Amharic language by some authors have revealed the role of ontology in improving performance of the retrieval system. Solomon Nega [40] has designed and formalized ontology that models the knowledge contained in Ethiopic church manuscript, art and music. The study identified 8 ontology groups namely Number, Date, Creature and Creator, Feast, Manuscript, Music and Art. TopBraid Composer (TBC) is a

tool employed for specification, formalization and implementation of ontologies. Apache Nutch is used for indexing and searching components. Jena provides models for interacting with the defined ontologies and also got facilities for SPARQL manipulation and reasoning support. Based on comparison made between ontology aided retrieval and keyword based retrieval systems on top 5 documents returned by the respective systems, the former has got better precision of 43% as compared to 34% of the latter. Performance of ontology aided system further increases as the comparison is made on top 2 query results which have registered a precision of 79% as compared to 43% of keyword based retrieval.

Ontology can also be exploited in the area of text categorization. Text categorization is the process of classifying a given document to a given category. Automatic text categorization is performed automatically through similarity measure that takes into consideration keywords or concepts of the document. In a keyword based categorization, for a given document to be categorized under a certain topic, the document should contain specific keyword that matches the words under which it will be categorized. Whereas in a concept based text categorization, semantic relationships between concepts is employed to classify a document based on the meaning it conveys. Meron Sahlemariam [95] conducted a research on concept-based automatic text categorization of Amharic News. The research was performed on Sport, Economics, Environmental preservation and weather conditions, Science and Technology and Accidents domains. Ontology was developed on Amharic news domain to represent knowledge that would help to categorize a given Amharic document into predefined category. TopBraid Composer was used as an ontology development platform and OWL was the ontology development language. During the ontology development process, glossary of terms were identified with the involvement of domain experts and used to build concept hierarchy and the prevailing relationships. Jena semantic framework was used to implement the reasoning capability of the developed system. During document classification process, the knowledge base module is involved between the document to be classified and the classification module. The knowledge base module returns a list of concepts to the classification module from the document to be classified. Depending on the weight of the concept, the classifier determines the actual category of a given document. The developed system was evaluated on 975 Amharic news and the result of the

test shows that on average 92.9% of the documents were correctly categorized under their respective concept.

From the works done on concept based text retrieval and categorization of Amharic documents, it is possible to understand how those systems that are supported by ontology improve their performance. Ontologies play key roles in describing the semantics of information exchange. They provide a shared and common understanding of concepts of a domain that can be communicated across people and application systems, and thus facilitates knowledge sharing and understanding between parties involving in the communication process.

3.3. Information Retrieval of non-Ethiopian Languages

Paralic and Kostial [96] made a comparison on the retrieval effectiveness of full text search (vector model with tf-idf weight schema), latent semantic indexing model and ontology-based retrieval model. The comparison was made on a well known Cystic Fibrosis collection of Medical scientific papers. Based on the experiment made on 1239 files, the researchers concluded that ontology-based retrieval system has better precision and recall as compared to the LSI and full text approach. The drawback of this research work was that manual assignment of queries to their respective concept was used and nothing was mentioned about the ontology used.

Similarly, Latifur [97] implement a concept-based model using domain dependent ontology for ontology based information extraction. Comparison was made for keyword and ontology-based retrieval system for broad queries, narrow queries, and context queries and the results of all the queries for the ontology-based model has shown a high level of precision and recall as compared to keyword-based search. The comparison was made on the working ontology which has 7000 concepts for the sport domains from CNN broadcast and Fox sport audio on 2481 audio clips and the result of the prototype test shows 90.5% of the objects are successfully associated with the concepts from the ontology. Based on the nature of the queries, the recall and precision of the ontology-based model outperform the keyword based model for broad and context queries.

Aung and Naing [98] presented the sport domain ontology architecture and sport information retrieval with semantic relationships between the concepts. The authors developed ontology on sport domain with emphasis given to football league that provides a means to retrieve sport related news from website. The system implementation architecture involves ontology creation and information retrieval systems. Sport domain ontology was built using Protégé-OWL editor and Jena SPARQL to handle RDF and OWL documents and getting SPARQL facilities for retrieving sport information from OWL ontology. The ontology creation involves identification of domain knowledge, collection of relevant terms and defining inference rules. The developed domain ontology consists of 32 classes, 20 object properties, and 9 data type properties. In the information retrieval system, user query terms are matched with the developed ontology in order to reformulate the query into domain concepts to improve retrieval effectiveness. Here, the users query keywords are converted to concepts of the domain for the semantic retrieval of documents. The semantic relationships are used for mapping concepts with predefined First Order Logic (FOL) and retrieve the results according to the mapping concept. The result of this research work shows that information retrieval system using semantic relationships based has better retrieval performance.

Lakshmi *et al* [99] discussed that the keyword approach results in a poor precision and recall. This is because in conventional search engine, information retrieval is based on a set of keywords or natural language query. According to the authors of this paper, in keyword retrieval, the system cannot understand what really is user needed, so it retrieves large number of documents which have poor semantic relationships and create difficulty for the users to navigate and select the appropriate documents. To overcome this problem, the researcher proposed the need for semantic based information retrieval system and developed ontology based information retrieval system using protégé OWL ontology development tool. Finally, conclusion was drawn which stated that the drawbacks of the conventional search engine can be rectified with the application of semantic web technologies like OWL ontology to develop efficient retrieval of documents. In ontology based text retrieval system, in which the meaning of information and services on the web is defined making it possible for the web to understand and satisfy the request of the user to use the web content.

In addition to the exploitation of ontologies for the semantic web where they mainly serve as a source of knowledge for semantic annotations and structuring of web data, they can also be exploited in the information retrieval system in the form of ontology-based query expansion. Carola [37] employed ontologies as a potential source of knowledge to reformulate users query as initial queries are considered to be naïve and often do not lead to results that immediately satisfy user's information need. According to this research work, queries can be automatically expanded with semantically related terms to increase either the exactness or the comprehensiveness of the matching process so as to come up with a better retrieval result. The test conducted on different semantic relations revealed that query expansion has significant impact on the result obtained for both automatic and interactive expansion tests done.

Fernandez [11] conducted a research on the exploitation of domain knowledge to support semantic search capabilities in large document repositories stressing the use of full-fledged ontologies for the retrieval of unstructured documents. The author incorporated vector space model into the proposed ontology-based retrieval system so that the retrieved documents are delivered ranked to the users' interface based on the importance of the documents for the delivered query. The proposed retrieval architecture involves the reformation or mapping of users query into query instances based on the knowledge base at hand. These query instances that are returned from the mapping via query processing module are matched against the weighted annotation or the semantic index of the document. In addition to this, to cope up with the possibility of semantic knowledge incompleteness in heterogeneous web environment, the author also proposed and incorporated keyword based retrieval scheme that complements the pure semantic retrieval model. The system retained the keyword matching capability in case of scarce or unavailable semantic information in the knowledge base. From this, conclusion was drawn as qualitative improvement can be achieved over keyword based retrieval paradigm by integrating and exploiting fine grained domain ontologies into information retrieval system.

Gaihua *et al* [100] reported on how the ontologies developed in the EU semantics web project SPIRIT are exploited to retrieve documents that have spatial information. Spatial terms are encoded in the geographic ontology and non-spatial information is encoded in the domain ontology. Query expansion method is executed on both the domain and geographical ontologies to supplement and widen the representation of users' queries and resolve the problems of vague spatial relationships. The experimental results show that the proposed method can considerably improve search result when a query involves a fuzzy spatial relationship.

Swathi *et al* [101] developed ontology for semantic information retrieval using ontology in the university domain. The objective of this work was to design, develop and implement a semantic search engine. The system development involves three major components which include ontology construction, refined query formation and ranking of retrieved links. Concepts related to university domain are constructed using Protégé 4.1 development tools. The query delivered by the user is refined to provide better search result by means of refined query formation system. In this system, the domain keywords that are semantically related to the given user queries are extracted from the ontology. Jena API is employed to extract semantically related and domain specific keywords using SPARQL. These refined keywords are used to fetch more semantically related web links on passing them to Google search API. Re-ranking was performed on the returned web links in the appropriate order of semantic relatedness. Performance evaluation of the developed system shows that it has better average precision and recall of 0.79 and 0.55 as compared to Google which has demonstrated 0.64 and 0.48 respectively.

3.4. Summary

In research works done on text retrieval system of both Ethiopian and non-Ethiopian languages, all text retrieval systems developed based on the semantic technologies demonstrated better performance than their keyword counterparts. Particularly, those research works done on the ontology-based text retrievals have demonstrated better retrieval performance on English and Amharic texts.

The research works done so far on Afaan Oromo text retrieval are keyword based and cross lingual information retrievals. Afaan Oromo search engine developed in [9] functions based on word matching principles. In all cases, the text retrieval technique is based on keyword matching techniques. In text retrieval system that is based on keyword matching principles, information retrieval is based on keyword or set of keywords. A given document is retrieved if it contains keywords specified in the user query words irrespective of the meaning of the word or the content of the documents. As a result, keyword text retrieval system misses important documents that do not have query keywords. At the same time, irrelevant documents that have keywords similar to query words but refer to different concepts are retrieved.

As many research works done on other languages show, ontology is proved to perform well in improving the performance of the retrieval system. It was developed and tested in text retrieval of Ethiopian languages like Amharic and non-Ethiopian languages like English and was found successful in improving the quality of text retrieval system. With ontology, domain concepts the meaning of which is shared by a group of communities, the meaning of information and services is explicitly defined and enhance the understanding of different parties involving in the communication process. This will in turn help the retrieval of documents that better satisfy the request of users.

With this in mind, Afaan Oromo ontology is developed on sport domain and integrated into text retrieval system to test and verify performance of retrieval system aided by ontology.

Chapter Four: Design of Afaan Oromo Ontology on Sport Domain

4.1 Introduction

Design of Afaan Oromo Sport Ontology consists of Afaan Oromo Sport Ontology (AOSO) that is developed from concepts extracted from sport news written in Afaan Oromo (AO) language. In the process of ontology development, terms describing key concepts of the domain are extracted from AO text corpus and organized based on their relatedness. They are used to construct hierarchical tree of concepts that show hierarchical relationship between concepts of the domain. In the concept hierarchy tree, concepts that are part of or specializing characteristics of the bigger ones are categorized under the concepts they specialize. Instances of a given concept are captured as type-of relationship to the concept they belong. Relationships that relate instances of concepts are captured as object relationships and relationships between instances of concepts and data values are captured as data type properties.

Implementation of Afaan Oromo ontology on sport domain is evaluated on a text retrieval system designed to exploit concepts, instances of the concepts and their relationships from the ontology. During the retrieval process that involves the ontology, queries that are delivered in the form of SPARQL query to AOSO via query operation process retrieves entities that fulfill conditions of the query. SPARQL query is created from user query from concepts, relationships and instances of AOSO through form-based user interface. The retrieval output of the ontology will further be used to retrieve Afaan Oromo (AO) documents based on the matching made between the ontology output and AO document index. AO documents are retrieved from document depository and delivered sorted based on their importance for the given query.

This chapter will present an implementation of AOSO which includes architecture of the developed system with simplified overview of key components of the developed system, and discussion will be made on the processes involved in the developed system.

4.2 Architecture of AO Ontology on Sport Domain

The architecture of the proposed system consists of SPARQL query that reflects users' information need, Afaan Oromo Sport ontology which is a repository of sport domain knowledge, Query operation that operates on AOSO based on the information from users' query, semantic entities are an output from query operation phase and is used as an input to the searching process. SPARQL query, AOSO, Query Operation and Semantic Entities are part of a component that deals with Extraction of Semantic Entities.

Extraction of Semantic Entities is added by this thesis work to AO text retrieval system. It is a component that functions based on the domain knowledge acquired from AOSO and delivers domain concept equivalent of user query. It enables Afaan Oromo text retrieval system to convert keyword based queries to domain concepts, individuals and relationships before the document searching process is started. The remaining components of the developed system consists of indexing unit that converts Afaan Oromo document corpus into AO index, searching unit functions between the unit that extracts semantic entities, AO index and ranking unit. The ranking unit sorts documents according to their relevance to the users' query. Detailed description of each components of the proposed system will be presented in Section 4.3

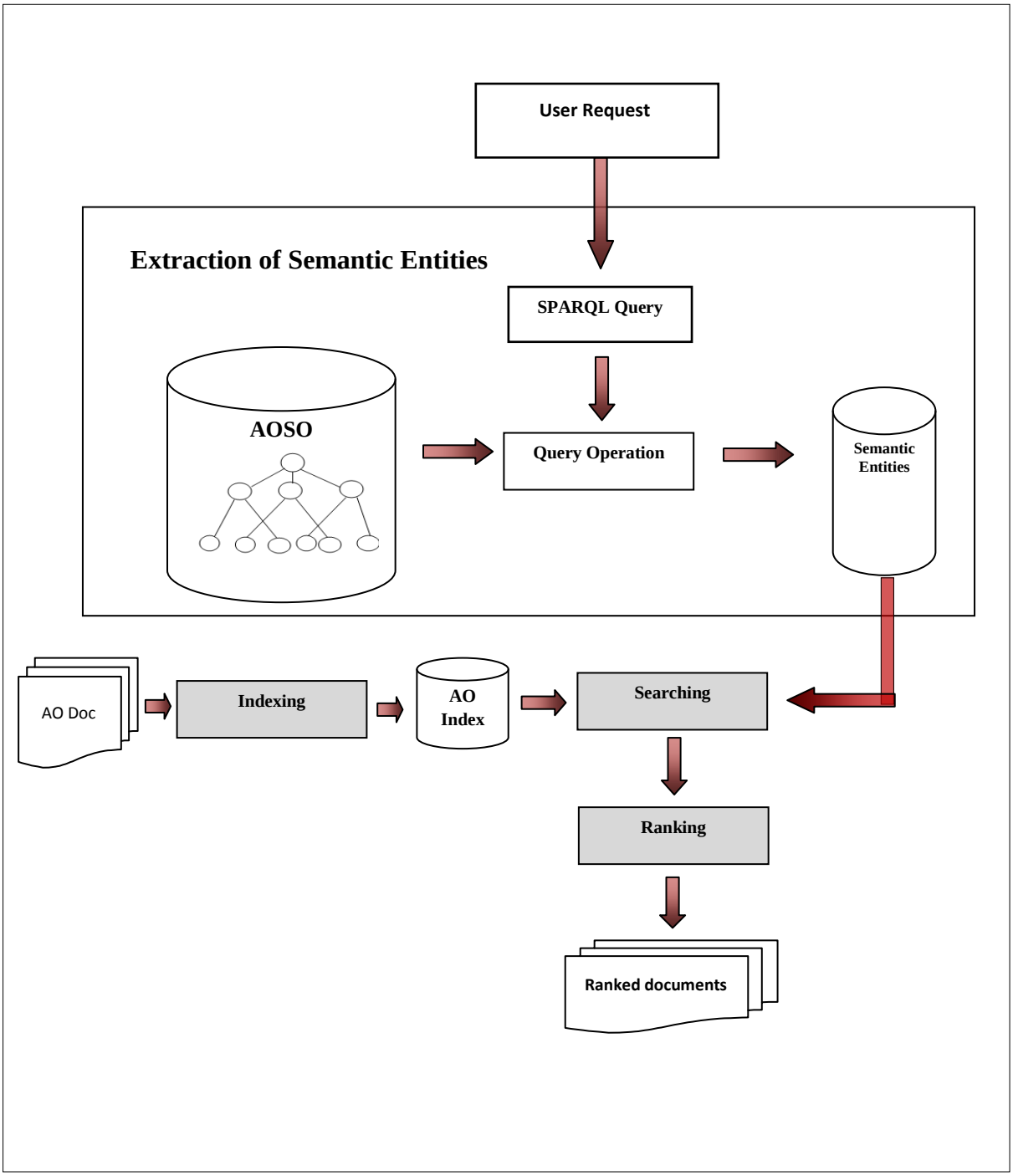


Figure 4.1 Proposed Architecture of AO ontology implementation

4.3 Extraction of Semantic Entities

Extraction of Semantic Entities is a component that deals with extraction of semantic entities from AOSO. It is a component that is responsible to convert SPARQL queries into domain

concepts that fulfill the conditions stipulated in the SPARQL query pattern. It has an ontology against which the SPARQL query is executed to retrieve semantic entities that will further be used in the searching phase. It is a system that is added to the classical keyword based Afaan Oromo text retrieval system to enable the retrieval system exploits information captured in the AOSO and helps contents of the query be expressed in the form of domain knowledge. The components that work as a unit to convert user queries into entities that represent domain concepts are SPARQL query, Query Operation and AO ontology.

4.3.1 SPARQL Query

SPRQL is an acronym that stands for Simple Protocol and RDF Query Language. It is a query language for querying RDF, RDFS and OWL files and is a standard query language adopted by World Wide Web Consortium (W3C). SPARQL has a capacity of extracting the required user information from RDF graphs or triples in the form of URIs, blank nodes, plain and type literals. It is employed to extract information from RDF graph of AOSO which are stored in the form of triples (<subject> <predicate> and <object>). It is a query language of preference over the others because it is easy to implement and more expressive as compared to others. It easily works with and be implemented in Jena and Protégé.

It uses a select-from-where syntax of information extraction that is based on pattern matching. The result of a query is a set of binding for the variable appearing in the “Select” clause. Entities that are retrieved from AOSO are specified in the “Select” clause. The data source from where the target entities retrieved is specified in the prefix. The prefix is combined with the conditions of the query in the query pattern of where clause like “prefix:condition”. Full prefixes used in the SPARQL query are shown in Section 5.3.3 of Chapter Five.

The where clause is used to specify the description of the entities in the form of subjects, relationships and predicates. It describes about the entities or variables indicated in the select clause. Entities that fulfill all the conditions specified in the where clauses are retrieved from AOSO. It is where descriptions of the required entity are given to identify one item from another. It helps to solve the problem of entities having similar naming from different domains as they have different descriptions.

4.3.2 Afaan Oromo Sport Ontology (AOSO)

Afaan Oromo sport ontology is developed from concepts extracted from sport news gathered from online sources. Concepts describing the sport domain are extracted and coded into the ontology. A total of 317 concepts were identified and manually captured in the Afaan_Oromo_Sport_Ontology. The extracted concepts and their hierarchical relationships are validated by two domain experts who have assessed the developed ontology and made their own contribution to correct the ontology as per the prevailing reality of the domain. Sample of graphical representation of concepts of the AOSO is given under the Section 5.2.2 and Annex B

The AOSO is retrieved by SPARQL query before the document matching or searching process is commenced. The ontology holds concepts, relationships, individuals and axioms of the domain and any query that is delivered in the form of SPARQL query will retrieve this concepts, relationships and individuals.

Apart from the asserted facts of the domain, the ontology employs Pellet reasoner to infer or derive new knowledge that is not explicitly stated during Ontology population (Annex D). The knowledge derived through inference of reasoner is also used during Ontology retrieval process.

4.3.3 Query Operation (QO)

Query Operation is the process of transforming users' information need entered in the form of SPARQL query into semantic entities by the operation on the Afaan Oromo Sport Ontology (AOSO). Query operation returns a list of items that fulfill the conditions specified in the query. Users may not explicitly know about the entity they are going to look for. They may deliver queries to information retrieval system in the form of description of the required entities which is not simple for the machine to comprehend and perform document retrieval that satisfy users' information need. Hence, queries of this nature need to be converted to the form that could be understood by computer to perform the required operation. Comprehension of the meaning of queries could be realized through query operation stage that performs matching of query pattern against RDF store from AOSO. Query operation

process is purely a Boolean retrieval process in which an exact match between query pattern and RDF triples stored in the AO ontology is required.

The ultimate objective of developing Afaan_Oromo_Sport_Ontology is to utilize the Ontology in the query operation stage and generate list of semantic entities which will later be used in the document searching phase. How AOSO is manipulated during query operation stage will be shown in Section 5.2.3.

4.3.4 Semantic Entities

Semantic entities are concepts, relationships and individuals returned from the retrieval of knowledge base through the process of Query Operation. They are the result of matching made between SPARQL query pattern and the RDF graphs of Afaan Oromo Sport Ontology. They are the retrieval output of query operation where an item or set of items fulfilling SPARQL query conditions are retrieved from AOSO. These Semantic entities are used as query words with which documents are searched through best matching principles.

4.4 Indexing

Indexing is the process of representing the contents of the documents by terms that are extracted from the documents. It is an off-line (done before the execution of text retrieval system) operation performed for the efficiency of the retrieval operation. It involves the processing of the original data into a highly efficient cross-reference lookup (index) in order to facilitate searching [9]. To search large amounts of text quickly, one has to first index the text and converts it into a format that enables to search it rapidly [102].

Before indexing process of AO documents is started, the following tasks need to be completed first. These are AO document analyzing which involves AO text tokenizer, AO text filter and case normalizer, stop words removal and stemming. According to [102], no search engines indexes texts directly: rather the text must be broken into a series of atomic elements called tokens; that is what happens during document analyze stage. Each token corresponds roughly to a word in the language.

Indexing is language dependent because stop words and stemming operations are language dependents and should be handled according to the rules of the languages. It is not possible

to apply the stemming algorithm developed for English and other languages like porter to Afaan Oromo language due to difference in the pattern of word formation and differences in their morphologies [18].

AO Analyzing

Apache Lucene Analyzer is customized to accommodate Afaan Oromo text pre-processing. Apache Lucene Analyzers are not written to perform text pre-processing of Afaan Oromo texts. Hence, it is not possible to directly apply Apache Lucene Analyzers for Afaan Oromo texts. Even though they use the same English alphabet, there is a complete difference between English (or other languages for which apache lucene is programmed) and Afaan Oromo languages. Apache lucene analyzer involves operations like stop words removal and processing tokens to reduce them to their roots using PorterStemFilter [102]. If it is applied directly to Afaan Oromo text pre-processing, it considers Afaan Oromo word as if it is English or other language word and could lead to improper manipulation that ends up with different outcomes. Hence, it needs to customize Apache Lucene Analyzer to Afaan Oromo language. AO Analyzing is conducted by AfaanOromoTextAnalyzer module. It has Tokenizer, AO Text Filter and Case Normalizer sub parts.

Tokenizing

In the analysis phase of Afaan Oromo text, there are modules that perform interrelated activities. The first module is Afaan Oromo Text Tokenizer which chops Afaan Oromo texts based on white space.

In the following Afaan Oromo text, “Tanaan afreeffachoo! Jarman Arjantiinaa Mootee Waancaa fi doolara Miliyoona 35 Argatte.”. The entire sentence is chopped into [Tanaan] [afreeffachoo!] [Jarman] [Arjantinaa] [Moote] [Waancaa] [fi] [doolara] [Miliyoona] [35] [Argatte.]. The entire sentence is split based on the white space. AO Tokenizing is executed by AfaanOromoTextTokenizer module.

Text Filter

The second module of the Afaan Oromo Analyzer is Afaan Oromo Text Filter where non-alphabetical characters that do not add value for the retrieval process are removed from

Afaan Oromo words. Symbols or special characters likes “% , ^ * (“) [ ] ” etc., are stripped off from the words bearing them. The only exceptional special character that is treated differently in Afaan Oromo text retrieval system is the apostrophe " ’ ". In Afaan Oromo language, it is considered as part of the word and is used to separate consecutive vowels that represent different sounds. Hence, it is left intact after text filter.

In the example above, an exclamation mark and full stop are removed. After application of this module, the above chopped texts will be [Tanaan] [afreeffachoo] [Jarman] [Arjantinaa] [Moote] [Waancaa] [fi] [doolara] [Miliyoona] [35] [Argatte]. It is the responsibility of AfaanOromoTextFilter module to remove these characters from Afaan Oromo words.

Case Normalizer

Case folding or case normalization is one of the pre-processing tasks undertaken to convert the texts into similar cases, mainly into lower case representation. Case normalization is important in order to have all documents and query words be represented with the same case. It is useful in the text retrieval process to avoid superficial variations caused as a result of case differences which could be misleading for the unit that performs matching operation. Semantic entities that are the retrieval output of Query Operation are also converted to lower case before they are employed for document searching purpose. The above example is case folded to [tanaan] [afreeffachoo] [jarman] [arjantinaa] [moote] [waancaa] [fi] [doolara] [miliyoona] [35] [argatte].

AO stop words removal

Stop words are words that do not convey meaning and has no significance in differentiating one AO document from another. Similar to other languages, there are words that have limited significance in aiding Afaan Oromo text retrieval process. They are less important because, they are found in almost all documents and do not help in differentiating one document from the other. Hence, it is wise to get them rid of from the document that is going to be indexed.

Words like “akka”, “ani”, “fi”, “nuti”, “nu”, etc are stop words in Afaan Oromo languages and have no significance in conveying the content of the document. Complete Afaan Oromo

Stop words identified in [8, 9] are listed in Annex A. Afaan Oromo stop words are removed from the document by AfaanOromoStopwordRemover.

Stemming

The core of every suffix stripper is a set of rules which test whether a word ends with certain character sequence and subsequently delete or replace this sequence or modify the stem [103]. Stemming is the process of removing morphological inflation from words to reduce them to their base forms. Stemming algorithm and rules developed in [103] are used to stem Afaan Oromo words to their root words in this work as shown in Algorithm 4.1.

```
1. Read the next word to be stemmed
2. Open stop word file
    Read a word from the file until match occurs or end of
    file reached
    If a word exist in the stop word list
      Go to 5
    Else
      Go to 3
3. If word matches with one of the rules
   Remove the suffix and do the necessary adjustments
   Go back to 3
   Else
   Go to 6
4. Return the word and record it in stem dictionary
5. If end of file not reached
   Go to 1
   Else
   Stop processing
6. If there is no applicable condition and action exist
   Remove vowel and return the result
   Go to 4
```

Algorithm 4.1

AO Index

Raw AO documents need to be converted to representation that is simple to traverse and locate during text retrieval. This process is called Indexing [102]. To generate indexed documents, contents of documents undergo document pre-processing as described above. In this work, the AO index is generated using OromiffaaTextIndexer module. Index files are represented as data structure called Inverted index which is employed for fast full text search. In this work, Afaan Oromo index is generated by using Apache Lucene 3.6.2 IndexWriter and IndexWriterConfig libraries.

4.5 Searching

In the query processing stage, semantic entities are retrieved from AOSO for a given query input in the form of SPARQL query. The searching module performs a comparison between the semantic entities and Afaan Oromo document index to decide documents that best satisfy the information need. Unlike the query processing stage, searching phase operates based on best or approximate matching principle. The purpose of query processing stage is to retrieve entities that are described in the user query which will later be used as input for the document searching phase.

According to [102], some of the Lucene classes that perform basic searching operations are IndexSearcher, Term, Query, TermQuery and TopDocs. IndexSearcher is a class that opens an index in a read-only mode. It is gateway to searching index. It uses a Directory instance to hold the previously created index and offers a number of search methods. Term is the basic unit for searching. It consists of a pair of string elements; the name of the field and the word (text value of that field). Term Query is the most basic type of query supported by Lucene. It is one of Query subclasses and used for matching documents that contain fields with specific values. When a Lucene index is queried, a TopDocs instances containing an ordered array of ScoreDoc is returned. The array is ordered by score by default. Lucene computes a score (a numeric value of relevance) for each document given a query.

Lucene computes a score (a numeric value of relevance) for each document given a query. Documents are retrieved based on the similarity measure conducted between query and document terms. Similarity measure gives the score that reflect how much a document and a

query match each other. A similarity measure with higher score reflects higher similarity between query and document terms than the one with lower score. Similarity measure of the query words and the AO document index is computed according to Apache Lucene similarity computation algorithm, the score of which is computed for each document and matching term [102]. Lucene combines Boolean model of information retrieval with Vector Space Model. Documents approved by Boolean model are scored by vector space model [104].

4.6 Ranking

Ranking is the process of sorting and listing documents returned from the execution of the information retrieval system based on their importance to the given query upon presenting the result of searching operation. The documents retrieved from the searching phase of Afaan Oromo document corpus are listed according to the relevance of the retrieved documents for the given query. The score of the query and document term similarity measure is the base to sort documents in order. The most relevant document will have the highest score, and is presented to the user at top of the list. The Ranking operation for this work is performed by Apache Lucene which sorts the matching documents in descending relevance score order so that the most relevant documents appear first to help users find the required documents in the first few searches [102].

Chapter Five: Prototype Development

5.1 Corpus Collection

Afaan Oromo Corpus from which concepts, relationships and individuals are extracted and used to populate the ontology are collected from different sources. The document corpuses are collected from online sources written in Afaan Oromo texts on sport domain giving emphasis to major international sport events. These documents are downloaded from online websites like VOA Afaan Oromo (http://www.voaafaanoromo.com/), Awash Post (http://www.awashpost.com/), OMN (http://www.oromiamedia.com/), Fana broadcasting (http://www.fanabc.com/), Barisaa (http://www.ethpress.gov.et). A total of 91 documents were collected and used for the ontology development process. They are chosen because they are the only easily available source of information found for our purpose.

5.2 Implementation

5.2.1 Tools Utilized

Java Framework

The system is implemented using Jena (Jena-Core-2.13.0) configured into Netbeans 8.1 Java development IDE. Jena is used for reasoning support and SPARQL manipulation. It also provides an abstract model that interacts with the RDF stores. A model builds on the basic graph to offer rich interactions with semantic web data. Applications read, write, reason and query semantic web data through access to the Jena model [43]. It is the preferred and standard language for the semantic web. The developed ontology is queried using SPARQL query editor. Queries are delivered in the form of SPARQL query to search for the required semantic entities from the knowledge base.

Apache Lucene

Searching and Indexing of AO documents are performed using Apache Lucene libraries. It is an open source library written in java that allows the programmer to build and maintain an index documents and allows querying of this index to find those documents one is interested in [102]. It has software libraries that enable applications to acquire searching and indexing

capability. IndexWriter, Directory, Analyzer, Document, and Field etc., are some of the core Apache Lucne classes employed for indexing while IndexSearcher, QueryParser, Term, Hits and TopDocs etc., are some of key classes involved in the searching of documents.

Protégé

Afaan Oromo Sport Ontology (AOSO) is developed using Protégé 4.3 ontology editor with Pellet reasoning engine configured to perform the inference. Pellet reasoner is used to check for the consistency of the asserted concepts and computes the hierarchical relationships of class [105]. Graviz-2.38 is configured into Protégé to enable the graphical display of concepts, individuals and their relationships. It is an OWLviz plug-in that helps to graphically visualize the asserted and inferred class hierarchy, instances of a class and their relationships. The ontology language chosen is OWL which is a W3C recommendation and built based on logical concepts that enable the use of reasoner to check for consistency of statements and definitions in the ontology.

5.2.2 Ontology Development (OD)

The ontology development process begins with the selection of the right ontology development tools. Protégé is the ontology editor used to capture concepts, individuals and relationship between the concepts. The ontology language used to model the concepts, individuals and their relationships is OWL ontology, which helps to build complex concepts from simpler ones using reasoners . Pellet is the preferred reasoner because it is the first and currently the only, complete OWL-DL consistency checker and has the most coverage of OWL as a whole of any reasoner [105].

During ontology development process of Afaan Oromo sport domain, concepts of the domain are identified and organized into taxonomies (super class – subclass), and the instances of the concept of domain and their relationships are captured in the ontology. Concepts, instances of a concept and their relationships are identified from AO document corpus. All documents are read and key concepts of the domain are captured as classes and relationships between the objects of the classes are captured as object relationships.

The overall metrics depicting axiom, logical axiom count, class count, object property count, data property count and individual count of the developed Afaan Oromo sport ontology is shown in Figure 5.1

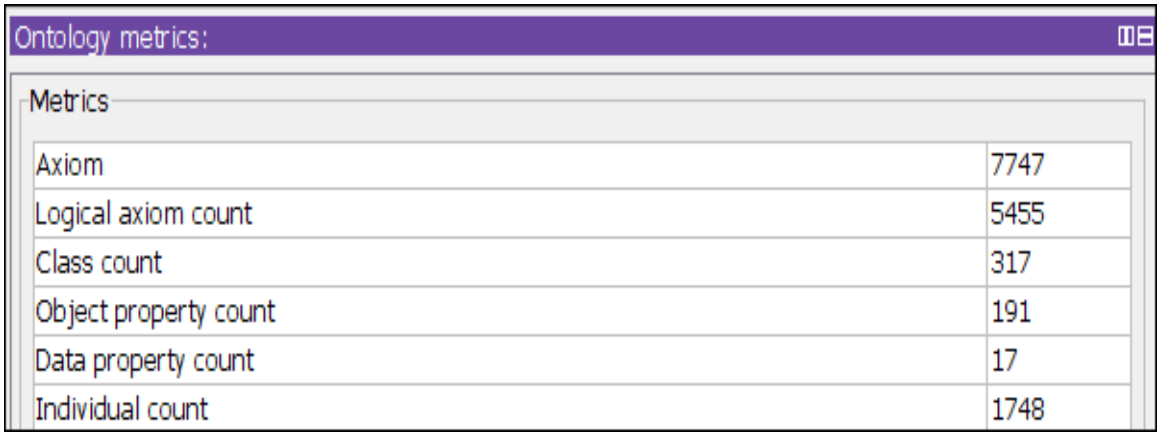


Figure 5.1 AOSO Concepts, Relationships and Individuals Metrics

Axioms and Logical Axioms

Axioms are facts that are asserted in the ontology. For example class A is subClassOf class B is a class axioms. Similarly, all other asserted facts during ontology development are called axioms. They are used to model sentences that are always true. They can be included in the ontology for several purposes, such as constraining the information contained in the ontology, verifying its correctness or deducting new information. Logical axioms are axioms that are inferred by reasoners [106].

Concepts

AO sport concepts which are also called classes are the conceptual representations of sets of entities of the same characteristics which are identified from Afaan Oromo sport domain. They are organized in a hierarchical tree based on subclass-super class, part-of etc., relationships. List of top level concepts identified and captured into AOSO are “Adabbii”, “Baasii”, “Badhaasa”, “Bara”, “Bakkee”, “Golii”, ”Ispoortii”, “Kilabii”, ”Kubbaa”, “Qabxii”, “Rikordii”, “Ummama” and “Waldaa”. The developed AOSO has a total of 13 top level concepts.

Class “Adabbii” is a concept related to punishments imposed in the form of money, point deduction, penalty, card (red, yellow), game embargo, participation embargo etc., for any sporting or administrative infringements committed in sport.

Class “Baasii” is related to expenditures incurred for various sporting activities, events, accommodations, commodities purchase, acquisition, salary, players transfer etc.,

Class “Badhaasa” is related to reward given for winning or achieving the target goal in sporting arena. It could be in the form of monitory, item, trophy, medal and the like.

Class “Bara” is related to time or episode. A period or year on which a certain sport event was undertaken.

Class “Bakkee” is concept related to a place or country where sporting activities are hosted.

Class “Ispoortii” is a concept related to any sport activities.

Class “Golii” is related to goal scored in sport competition.

Class “Kilabii” is related to organization of sport team.

Class “Kubbaa” is a class referring to ball with which sporting activities are accomplished. Football, basketball, volleyball etc., are examples of ball or “Kubbaa”.

Class “Rikordii” is related to records that are recorded for exceptional performance in sport.

Class “Ummama” stands to represent participants or actors (Athletes, Referee, Coaches or Managers, Players, Administrators, Agents, Horses, fans etc.,) of the sport.

Class “Waldaa” is used to represent sport related federations or organizations. FIFA, CAF, Ethiopian Football federation are examples of organizations in the world of sport.

Class “Qaphxii” is another class in the Afaan Oromo sport world used to represent the point gained, accumulated from winning a match, set etc., Three point is rewarded for winning, one point for a draw in football competitions, a point for winning a ball in volleyball, a point, two or three points given for scoring a ball in basketball competition depending on the position from where the point is scored. Figure 5.2 shows top level concepts identified in Afaan Oromo sport ontology.

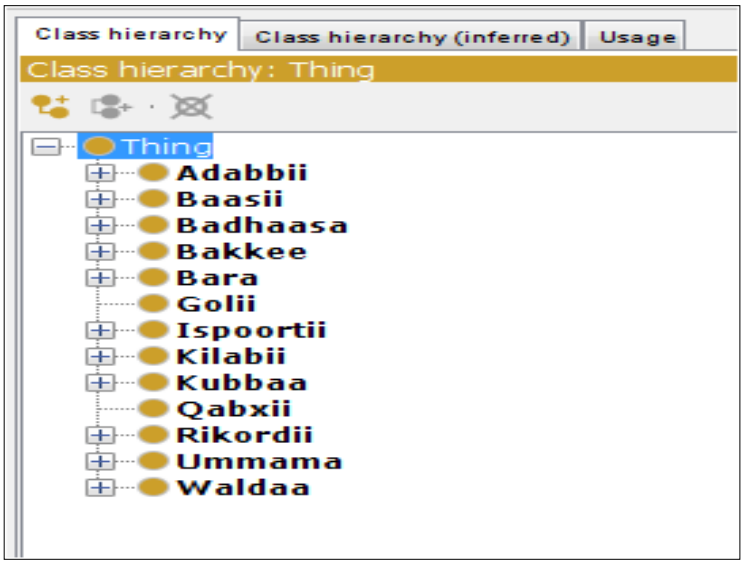


Figure 5.2 Top Level Concepts identified AOSO

Figure 5.3 depicts the graphical representation of top level classes identified in AO sport domain.

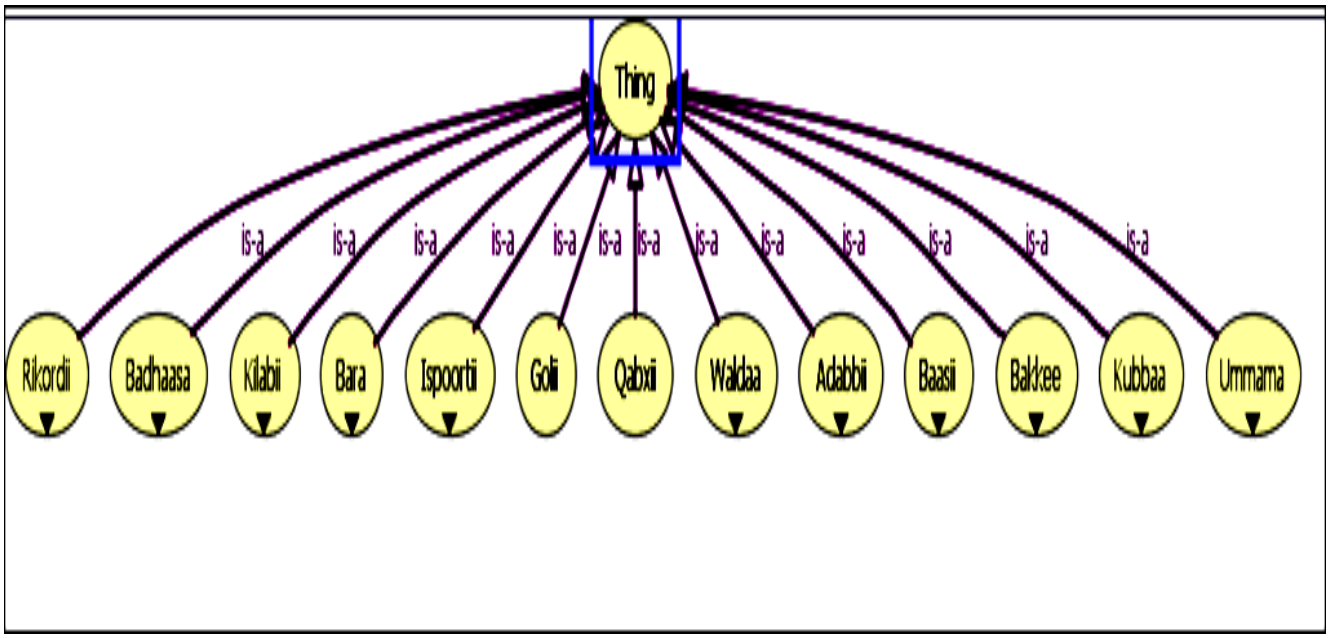


Figure 5.3 Graphical visualization of top level concepts of AOSO

These top level classes are further classified into related subclasses based on the relatedness of the concepts under consideration. “Ispootii” class is classified into different types of subclasses like “Dorgommii” which includes competitions like World and continental cups,

Olympics, World Athletics championships, club competitions, and other major international games etc., Similarly, other related subclasses of “Ispoortii” concepts are “Atileetiksii”, “Bishaan_Daakuu”, “Biskileetii_Oofuu”, “Booksii”, “Daartii”, “Gaaraa_Koruu”, “Dhukaasaa_Ilaamaa”, “Guluffii”, “Isportii_Aadaa”, “Isportii_Kubbaa”, “Jimnaastikii”, “Juudoo”, “Konkolaachisuu”, “Reesilingii”, “Skeetingii”, “Tarkaanfii”, “Ulfina_Kaasuu”, “Hookii” are categorized under “Ispoortii” concept. The hierarchical tree depicting the subclasses of “Ispoortii” concept is shown in Figure 5.4. Top level concepts and some of their sub concepts are displayed in Annex B.

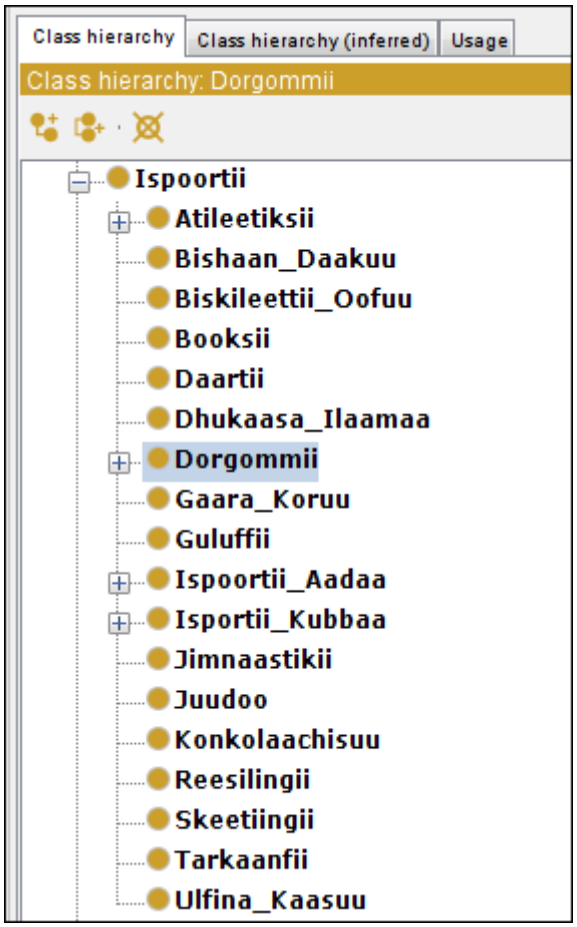


Figure 5.4 Subclasses representation of “Isportii” concept

The AOSO concept development process follows a combination of top-down and bottom-up approaches in the class hierarchy development process. Since the concepts are extracted from documents, for every concept found, the class construction approach could either be an upward to the general concept or specialized downward toward the specific concepts

depending on the nature of the concepts identified during ontology population. If the concept extracted is a general concept like Olympics, “Olompiikii” in AOSO context, a top-down approach is followed to the level of subclass of Olympics and goes upward or bottom-up to its super class “Dorgommii” and “Ispoortii” under which Olympic is categorized. Similarly, if the concept extracted is about a particular world cup, an upward class hierarchy development strategy is followed to include concept under which that particular world cup is categorized, which means a bottom up development approach. For a statement describing about an instance of world football cup, for example “Dorgommii_Kubbaa_Miilaa_Riyoo_Bara_2016”, an upward development strategy is followed to place it under its proper concept “Dorgommii_Kubbaa_Miilaa_Addunyaa”. In Figure 5.5, for information captured on the “Dorgommii_Maaraatonii_Dubartii_Atlaantaa_Bara_1996”, Atlanta 1996 women marathon competition and “Dorgommii_Maaraatonii_Dubartii_Landan_Bara_2012”, London 2012 women marathon competition, a bottom-up approach is followed to the level of “Olompiikii_Gannaa” or summer Olympics, “Dorgommii”, and “Ispoortii”. For information captured on the success of Ethiopian Athletes on Olympic arena, a top-down development approach is followed from the concept of “Dorgommii_Olompiikii” etc., to the instances of specific Olympics.

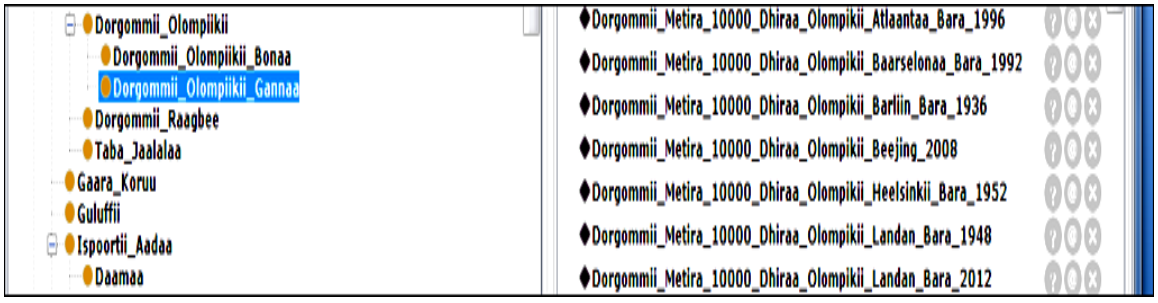


Figure 5.5 Concepts and Individuals representation in AOSO

Individuals

Individuals also called objects are instances of classes or concepts. “Hayilee_Gabrasillaasee”, “Qananiisaa_Baqqalaa”, “Xirunash_Dibaabaa” etc., are instances or real objects of the concept “Atileetii” or Athlete. “Meesii”, “Kiristiyaanoo_Ronaaldoo”, “Suwaarez” etc., are instances of class “Tabataa_Kubbaa_Miilaa” or football player. For

During ontology population, for individuals known by different names, all names can be captured as "Same Individual As" in the individual description tab so that the reasoners can infer as different documents describing them refer to the same thing. For different documents describing about African athletics championships hosted in Addis_Ababaa, Finfinnee, Shaggar in the year 2004, it is possible to infer as all these documents are stating about one event from AOSO. Figure 5.6 shows how individuals known by different names are captured in AOSO so that reasoners can infer as they mean the same thing during information retrieval.

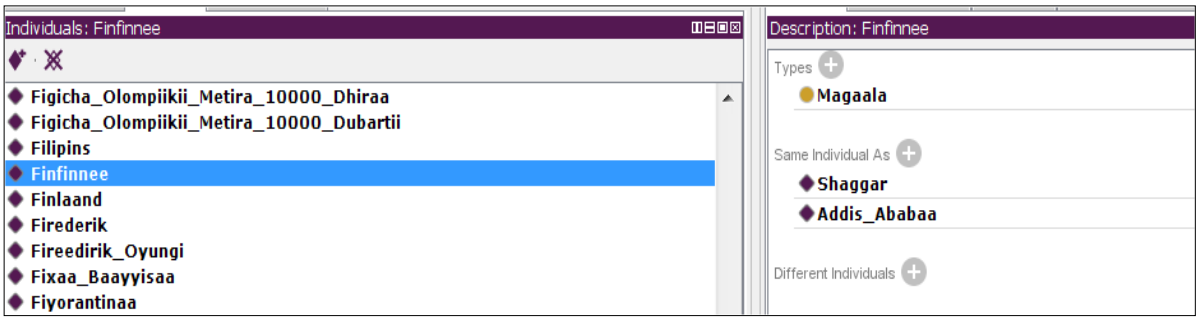


Figure 5.6 Individuals known by different names representation in AOSO

In Figure 5.7 & 5.8, all individuals are categorized under their respective concept type. Instances of Athlete are categorized under concept "Atileetii" and instances of football players are categorized under their respective "Tabataa_Kubbaa_Miilaa" class. They are types of their respective concept and that is the reason why concepts are defined as a collection of instances or objects of the same type.

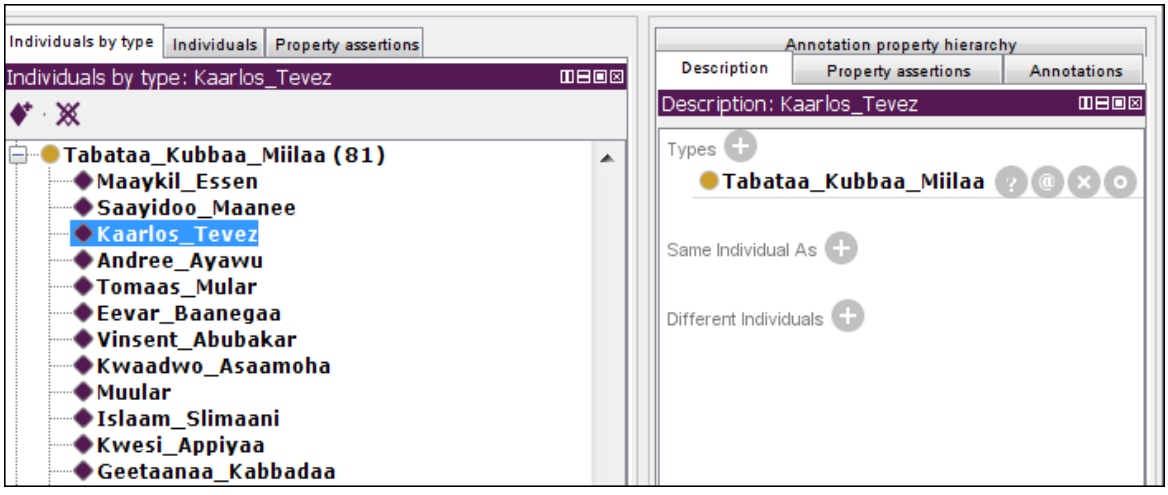


Figure 5.7 Individuals representation for Concept "Tabataa Kubbaa Miilaa".

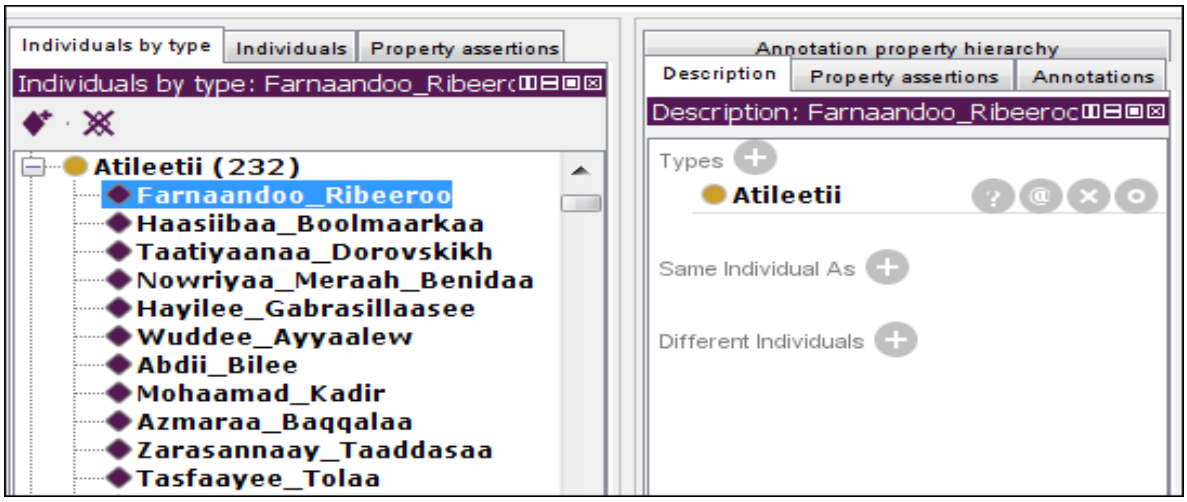


Figure 5.8 Individuals representation for Concept "Atileetii"

Figure 5.9 shows graphical representation of individuals or instances of class "Baaloon_Dor" displayed using OntoGraf. "Baaloon_Dor" is a type of reward for football players given based on their performance demonstrated in a given competition year. It is subclass of concept "Badhaasa" in AOSO. Individuals for this concept are Balon Dors awarded for football players on each year since its inception. "Badhaasa_Baaloon_Dor_Bara_2014", "Badhaasa_Baaloon_Dor_Bara_2015" etc., are individuals of a concept categorized under the concept "Baaloon_Dor".

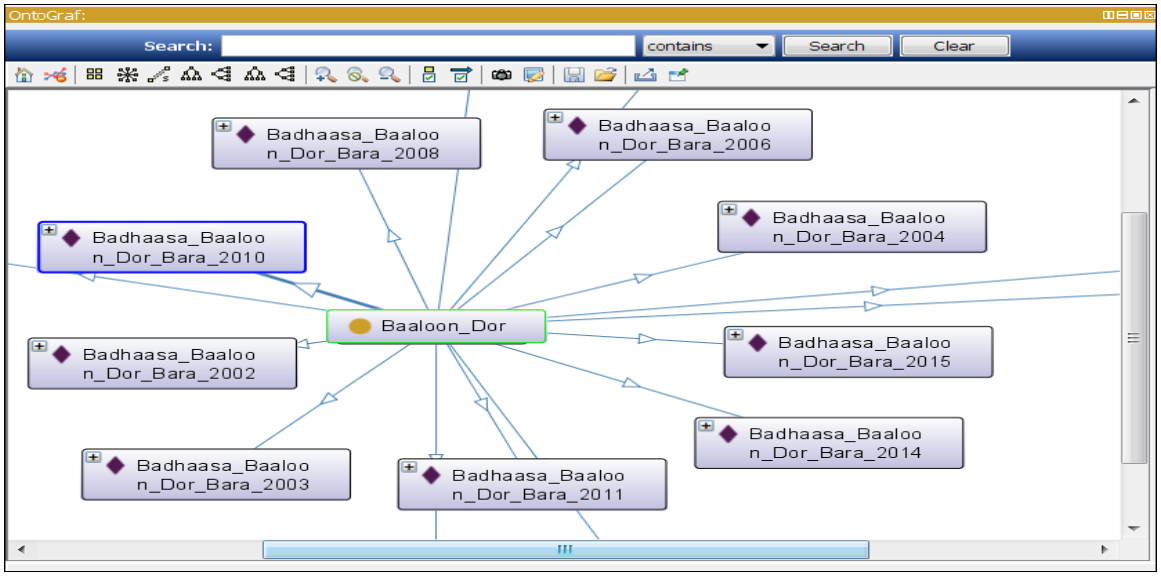


Figure 5.9 Graphical representations of AOSO Individuals for concept "Baaloon_Dor"

From the Ontology metrics displayed in Figure 5.1, the total number of sport related individuals identified and captured in the AOSO knowledge base are 1748. Some of the individuals' categories of AOSO are graphically displayed in Annex C.

Relationships in AOSO

Two types of relationships or properties are modeled in the AOSO. Object property and Data type property. Object properties show the relationship between two individuals. It links one individual with another individual. Data type properties describe relationship between an individual and data value. In the description or RDF statement of "Xirunash atileetii_Biyya Itoophiyaa", "Xirunash" and "Itoophiyaa" are individuals representing instances of concepts "Atileetii" and "Biyya_Afrikaa" respectively. "atileetii_Biyya" is an object property linking the two individuals. Whereas, in the description "Xirunash maqaa_Guutuu Xurunash Dibaabaa Qananii", "maqaa_Guutuu" which means full name of object "Xirunash" is data type property describing or linking object of a class "Atileetii" and data value "Xirunash Dibaabaa Qananii". Properties referring to the same situation can be captured as equivalent properties so that they can be inferred during query operation as seen in Figure 5.10.

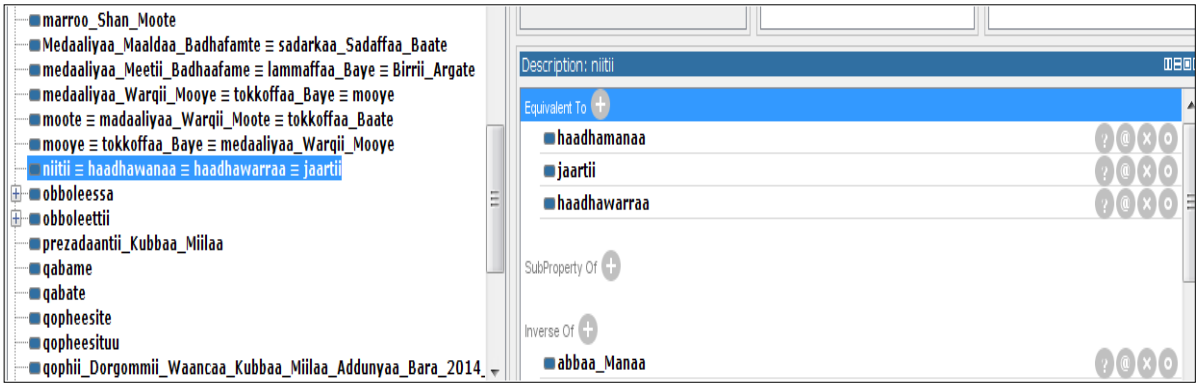


Figure 5.10 Equivalent property representation in AOSO

Relationships Representation in AOSO

Concepts or classes are usually described by nouns whereas properties are verbs that link the individuals. In English language sentences, verbs are usually placed in between nouns. In the sentence, "Brazil hosted 2016 Summer Olympic", "Brazil" and "2016 Summer Olympic" are nouns. They can be captured in the knowledge base as individuals. The verb

hosted that relates the two individuals is placed between subject “Brazil” and the object “2016 Summer Olympic” and can be captured as relationship or property during ontology population.

However, in Afaan Oromo sentence construction, verbs that relate two individuals are not usually placed in between the nouns. In Afaan Oromo case, the sentence “Brazil hosted 2016 summer Olympic” can be translated as “Braazil dorgommii Olompiikii Riyoodejaaneroo bara 2016 qopheesite”. If we conduct similar interpretation as we did for English, “Braazil” and “Dorgommii Olompiikii Riyoodejaanerooo bara 2016” are nouns where as “qopheesite” is a verb that relates the two nouns. When these facts are captured in AOSO, the two nouns are captured as individuals. “Braazil” is an instance of concept “Biyya_Amerikaa_Kibbaa” meaning South American country and “Dorgommii_Olompiikii_Riyoodejaaneroo_Bara_2016” is an instance of concept “Olompiikii_Gannaa”. During ontology creation, properties are placed between individuals they link or relate. For the information, “Brazil hosted 2016 summer Olympic”, the property “hosted” is placed between the individuals “Brazil” and “2016 summer Olympic”. Similarly, in AOSO even though the verb relating the subject and object has different arrangement in the sentence construction as compared to English language, the property will go between the individuals it is relating. In the information “Braazil Dorgommii Olompiikii Riyoodejaaneroo bara 2016 qopheesite”, the property “qopheesite” is rearranged in the ontology development as “Braazil” “qopheesite” “Dorgommii_Olompiikii_Riyoodejaaneroo_Bara_2016” (DOR_Bara_2016). Figure 5.11 and 5.12 show how object property is modeled during Afaan Oromo Ontology development and how it is actually represented in the developed AOSO respectively.

Figure 5.11 Object property modeling in AOSO

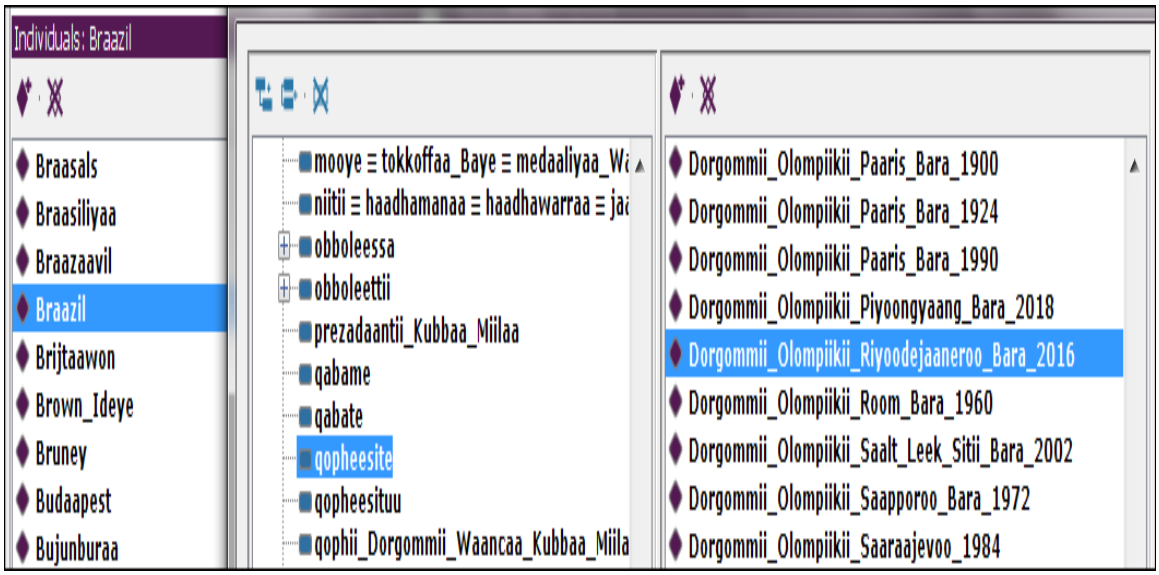


Figure 5.12 Object property representation in AOSO

Data property is also represented in similar way as the object property. Objects are related to the data value representing them through their respective data property. Figure 5.13 and 5.14 show the way the property relating individuals to their data value are modeled and represented in AOSO respectively.

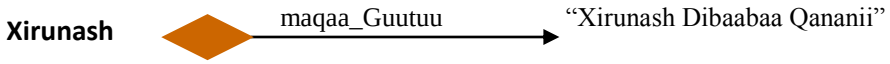


Figure 5.13 Data property Modeling in AOSO

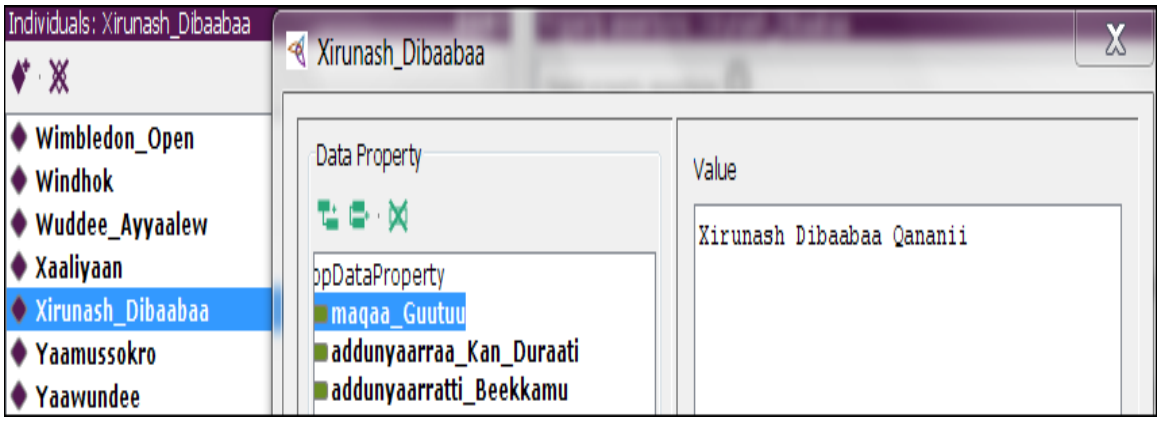


Figure 5.14 Data property Representation in AOSO

From the Figure 5.1, 191 Object properties and 17 data properties were captured during AOSO development process. Graphical representation of properties in AOSO is displayed in Annex D.

5.2.3 Manipulation of AOSO using SPARQL query

Developing ontology is not an end by itself. Ontologies are developed for some purposes. The information captured in the ontology should be manipulated in some way to be useful information. The information in the ontology is accessed through SPARQL query. Concepts, individuals, relationships, axioms and other relevant information derived through inference mechanism using Pellet reasoners are accessed during query operation stage. The required information is specified in the form of variables in the select clause and the where clause describes the conditions of the required information through properties, predicates and objects so that comparison and matching will be done with the RDF statements stored in the ontology. For the ontology to be accessed, it has to be loaded to the Java IDE through Jena Ontology Models. Ontology model is an extension of the Jena RDF model that provides extra capabilities for handling ontology data sources [107]. An RDF Model is a set of statements expressed in the form of subject, predicate and objects. An RDF graph in Jena is called a model [108]. A collection of these RDF graphs gives us a set of statements that can be accessed and retrieved. An ontology model is created through Jena ModelFactory to manipulate information populated in Afaan Oromo Sport ontology using SPARQL query.

```
OntModel model = ModelFactory.createOntologyModel(OntModelSpec.OWL_MEM_MICRO_INF);
```

OntModel is a special type of Jena Model used to derive meaningful relationships that the Model does not express directly. It is Model plus a reasoner to derive additional information through inference mechanism. The OntModelSpec is used to configure an ontology model with the language in use, the reasoner and the means of handling compound documents. OWL_MEM_MICRO_RULE_INF is a specification for OWL models that are stored in memory and use the micro OWL rules inference engine for additional entailments.

```
String filename  = "Afaan_Oromo_Sport_Ontology.owl";
File file = new File(fileName);
FileInputStream reader = new FileInputStream(file);
```

model.read(reader, "RDF/XML-ABBREV");

Ontology created on sport domain is loaded, read in the form of file, serialized in the form of "RDF/XML_ABBREV" and queried using SPARQL query in the form described in Figure 5.15

```
"?x rdf:type myont:Atileetii. "
+ "?y rdf:type myont:Atileetii. "
+ "?g rdf:type myont:Atileetii. "
+ "?x myont:atileetii_Biyya ?z. "
+ "?y myont:atileetii_Biyya ?z. "
+ "?g myont:atileetii_Biyya ?z. "
+ "?x myont:oboleettiin_Ishee ?g. "
+ "?g myont:moote myont:Dorgommii_Metira_1500 _Dubartii_Shaampiyoonaa_Addunyaa_Beejing_Bara_2015. "
+ "?z rdf:type myont:Biyya_Afrikaa. "
+ "?x myont:niitii ?y."
+ "?x ?r ?y. "
+ "}";
```

Figure 5.15 SPARQL Query instance to query AOSO

Figure 5.15 represents SPARQL query code snippet written to retrieve Afaan Oromo Sport ontology to search for information on athletes of African country where one of the required athletes is wife of the other and has athlete sister who won Beijing 2015 world Athletics championships in 1500m race. The equivalent query in Afaan Oromo language is "Atileetii biyya afrikaa dhirsaaf niitii walii ta'anifii biyya tokko kan dorgoman, oboleettiin ishillee dorgommii metira 1500 dubartii shaampiyoonaa addunyaa Beejing bara 2015 mootera".

During the SPARQL operation, the given information is split into relationships, objects or subjects and processed accordingly during the ontology retrieval. The required entity whether it is a subject, relationship or object is expressed in select clause in the form of variables. Additional information which will further specify the required entity is expressed in the where clause. An entity or a set of entities which fulfill all the conditions expressed in the where clauses are retrieved if captured in the ontology. In the SPARQL query specified in Figure 5.15, the conditions specified in all the 11 rows should be satisfied for the retrieval of items from AOSO. Required individuals specified in the form of variables x, y and g should be an athlete, z should be a type of African country and all the three athletes specified as x, y and g are athletes of country z. Athlete x and g are related through object

property “oboleettin_Ishee” (sister_Of) and x and y through “niitii” (wife_Of) and athlete g is specified as “dorgommii metira 1500 dubartii shaampiyoona addunyaa Beejing bara 2015 moote” or a winner of women 1500 metre competition on 2015 Beijing world athletics championships. The output of the ontology retrieval process is displayed in Figure 5.16.

x	r	y
myont:Xirunash	myont:haadhawarraa	myont:Silashii
myont:Xirunash	myont:haadhamanaa	myont:Silashii
myont:Xirunash	myont:jaartii	myont:Silashii
myont:Xirunash	myont:niitii	myont:Silashii

Figure 5.16 Ontology Retrieval Output of SPARQL query

From these semantic entities, we can obtain query words with which the documents will be searched. Individuals like “Xirunash”, “Silashii” and their relationships “haadhawarraa”, “haadhamanaa” “jaartii” and “niitii” are obtained to serve as a query word as shown in Figure 5.17.

```
[Xirunash, haadhawarraa, Silashii, haadhamanaa, jaartii, niitii]
```

Figure 5.17 Query words comprising of AOSO Objects and Relationships

These query words are case folded or normalized before they are used for the required purposes. They are converted to lower cases and used against the document index to search the required document as shown in Figure 5.18.

```
xirunash, haadhawarraa, silashii, haadhamanaa, jaartii, niitii
```

Figure 5.18 Case folded query words retrieved from AOSO

5.2.4 Form-Based Access to Afaan Oromo Sport Ontology

Concepts, relationships and objects of Afaan Oromo sport ontology is delivered to users through form based interface so that it will be simple to select and deal with the required subject under consideration. In the “All Concepts” column in Figure 5.19, concepts of the domain that are captured in the ontology are displayed. Similarly, in the “All Relationships”

and “Individuals” columns, all relationships and individuals of AOSO are displayed respectively. Similarly, upon selection of concepts, individuals that are type of the selected concept are displayed in the “Individuals” column. Individuals that are specific to the selected concept are retrieved and populated the list so that it would be simple to create query.

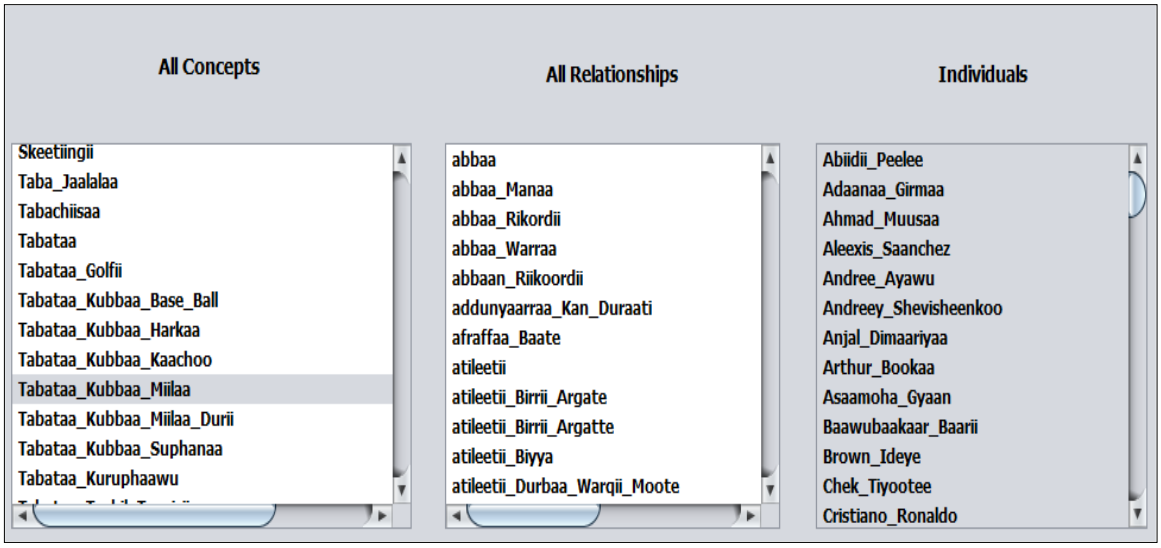


Figure 5.19 Concepts, Relationships and Individuals of AOSO accessed through Form based Interface

For a given concept, it is also possible to select relevant relationships from “All Relationships” column. When a given relationship is selected, the system displays predicates and subjects that are captured in the AOSO together with the selected relationship to help in the construction of query as seen in Figure 5.20. When relationship “dhaloota_Biyya” which means country_Of_Birth is selected, countries which are captured in the ontology with the specified relationship i.e., “dhaloota_Biyya” are displayed to assist in the creation of the query.

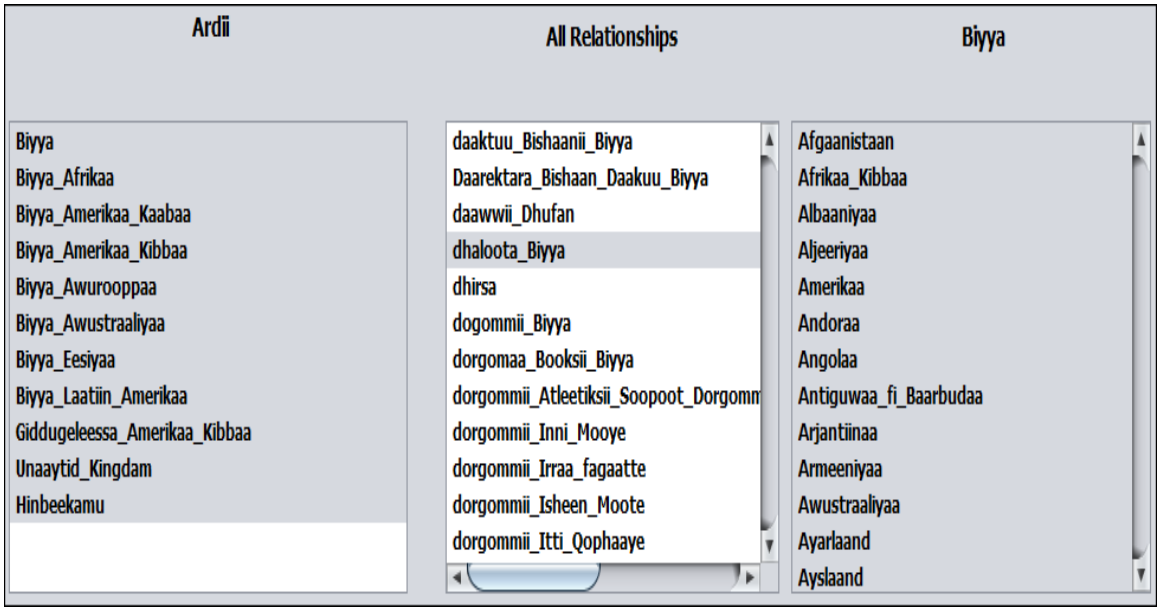


Figure 5.20 Predicates of "dhaloota_Biyya" relationship in AOSO

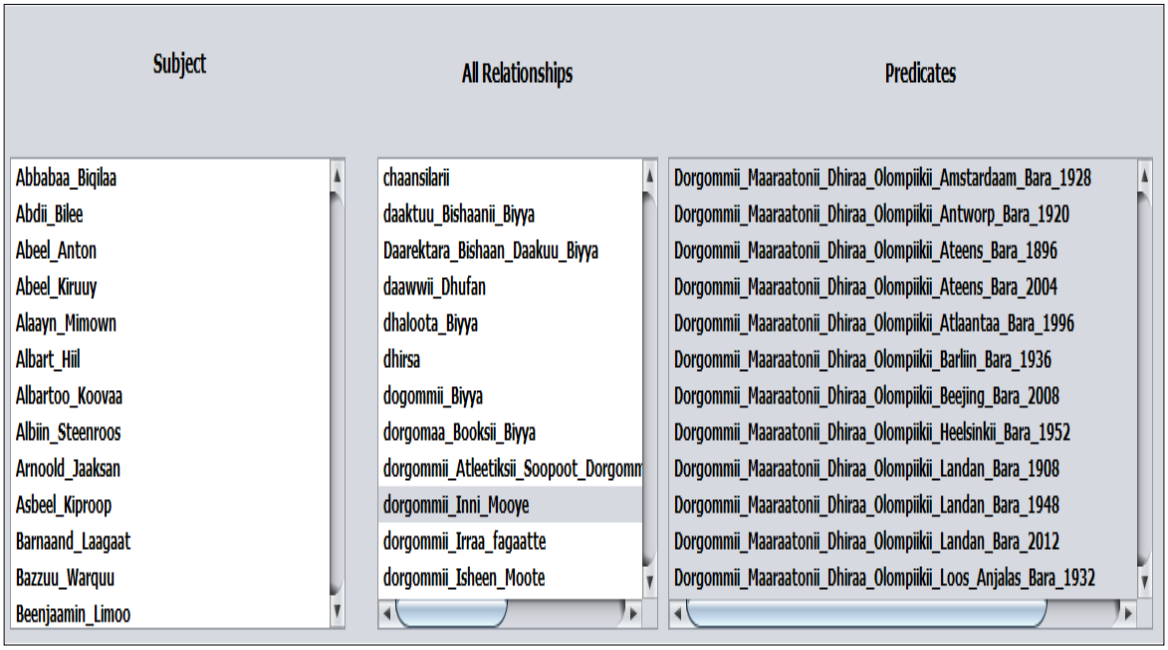


Figure 5.21 Predicates of relationship "dorgommii_Inni_Mooye" in AOSO

Figure 5.21 shows how to access subjects and predicates when a relationship or property is selected. "dorgommii_Inni_Mooye" is an object property that relates two objects. A competitor is displayed in the subject column and a competition that is won by any competitor and captured in AOSO is displayed in the predicate column.

For each concept selected, for example for the concept “Tabataa_Kubbaa_Miilaa”, the system constructed a query which enquires for an entity that is a type of the selected concept, in this case “Tabataa_Kubbaa_Miilaa”. Similarly for a given relationship and predicate selected, simple SPARQL query is constructed as shown in Figure 5.22. In the Plain Text Query part, all the concepts, relationships, predicates selected are displayed for confirmation of the selected item. In the SPARQL_Query part, respective simple SPARQL query is generated for the selected semantic entities.

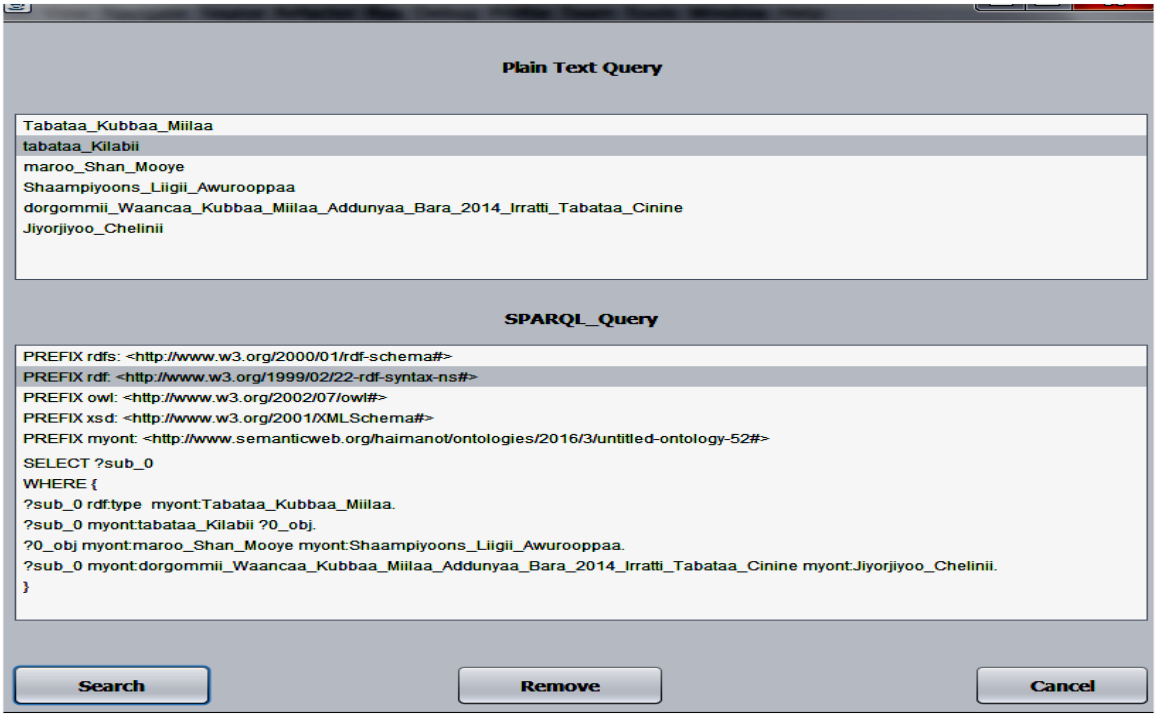


Figure 5.22 Simple Query Creation from Concepts, Relationships and Predicates of AOSO

Through navigation of AOSO concepts, relationships and individuals, Simple SPARQL and plain text queries are formed for users so that it is possible to retrieve the required entities from AOSO that fulfils the specified condition of query pattern. For the selection of every concept, an enquiry of entity of the selected concept type is formed. For the selection of concept “Tabataa_Kubbaa_Miilaa”, respective SPARQL query formed is “?var rdf:type myont:Tabataa_Kubbaa_Miilaa.” Which enquires the system to search for an entity or set of entities of type “Tabataa_Kubbaa_Miilaa” from AOSO. For the selection of relationship “tabataa_Kilabii” for example, entities which are linked via “tabataa_Kilabii” relationship in the AOSO is displayed and upon selection of the displayed entities a query combined of the

selected items in the form of subject, relationship and predicates are created as shown in Figure 5.22.

5.3 Result and Discussion

5.3.1 Evaluation

Evaluation was made on the retrieval effectiveness of the developed system using the document corpus. The queries used are not simple and straight forward keyword queries but they are queries that are description of the required entity. These queries are converted into SPARQL query to generate the semantic entities that are used as query word to search for the document corpus. The retrieval result effectiveness is measured using precision which measures the ratio of relevant documents that are retrieved by the system. Queries employed to test the implementation of ontology in the Afaan Oromo text retrieval process is presented in Table 5.1. These queries are delivered in the form of SPARQL query for the retrieval of equivalent knowledge captured in Afaan Oromo Sport Ontology.

Table 5.1 AO queries for the evaluation of AO text retrieval system with and without the use of Ontology

S.N.	Query
1	Tabataa kubbaa miilaa kilabii biyya Awuroppaa shaampiyons liigii maroo shan mooyeef tabatu, dorgommii waancaa kubbaa miilaa addunyaa Braazil bara 2014 irratti tabataa cinine
2	Tabataa kubbaa miilaa kilabii biyya Afrikaa kan ture amma dargagoota kilabii QPRtiif tabachuuf carraa kan argate
3	Tabataa kubbaa miilaa biyya Amerikaa kibbaa, dorgommii waancaa kubbaa miilaa addunyaa Braazil bara 2014 irratti lammaffaa baateef tabatu fi dorgommii sanirratti kubbaa warqii badhaafame
4	Atileetii “babay face destroyer” jechaan beekamtu fi oboleettin ishee dorgommii metira 10000 dubartii olompiikii Ateens bara 2004 irratti lammaffaa baate.
5	Atileetii biyya Afrikaa, abbaa rikordii figicha metira 10000 addunyaa fi olompiikii kan taate, leenjisaan ishee atileetii durii fi dhirsa ishee kan ta’e
6	Tabataa kubbaa miilaa biyya dorgommii waancaa kubbaa miilaa addunyaa Biraazil bara 2014 irrattii waancaa mooye fi kilabii biyya isaa keessatti tabachaa kan jiru
7	Lammii biyya Afrikaa kan Olompiikii Addunyaarrattii worqii mooye fi Maaraatonii Iskotlaand 7ffaa irratti rikordii bakkee kan cabse

5.3.2 Test Result

All queries described in Table 5.1 describe about an entity or collection of entities. These queries describe about an item that the system is expected to search for. Performance of Ontology aided Afaan Oromo text retrieval system is compared against retrieval performance when the user query is used as it is. In the ontology based retrieval, the queries are converted to SPARQL query that will query the AOSO knowledge base. In the direct use of the queries, all descriptions and other information delivered are directly employed in the retrieval process.

During evaluation of the system, the retrieval performance of the system for each query articulated above is evaluated by two advanced users of the language and the average precision for top documents retrieved by the system is summarized in Table 5.2. Number of documents retrieved by ontology-aided retrieval system is used as the number of top documents for comparison purpose. If ontology-aided retrieval retrieves 7 documents, comparison is made on top 7 documents for both retrieval systems.

Table 5.2 Evaluation result of Afaan Oromo text retrieval with and without the use of AOSO

S.N.	Number of Total Docs retrieved	Ontology based searching		Keyword based searching	
		Relevant doc. retrieved	Precision (%)	Relevant doc retrieved	Precision (%)
1	7	5	71.4	2	28.5
2	6	4	66.6	2	33.3
3	5	5	100	2	40
4	6	5	83.3	2	33.3
5	5	5	100	2	40
6	4	3	75	1	25
7	4	3	75	1	25
Average			81.61		32.16

5.3.3 Discussion

Query-1: “Tabataa kubbaa miilaa kilabii biyya Awuroppaa shaampiyons liigii maroo shan mooyeef tabatu, dorgommii waancaa kubbaa miilaa addunyaa Braazil bara 2014 irratti tabataa cinine”.

This query is querying about football player who is playing in a European professional football club who won European champions league five times and bite player of another country on 2014 world football cup competition hosted by Brazil. Querying the knowledge base will retrieve the specific entity by which the document will be searched. However, if we use the query as it is to search for Afaan Oromo document corpus, a lot of unrelated documents that result in a poor retrieval performance is retrieved due to the nature of the query. 71.4% against 28.5% precision is obtained when AOSO is used and not used respectively.

The query given above is composed of a lot of information. In the case of ontology-aided system, it is possible to split this information into concepts, relationships and individuals. A form based interface that displays all information captured in the ontology can aid users to decide the category of a given entity and create query. The query is composed of concepts like “Tabataa_Kubbaa_Miilaa” who is playing for a football club based in Europe that won European champion league five times. We can also find relations like “tabataa Kilabii, “dorgommii Kubbaa Miilaa Addunyaa Bara 2014 Irratti Tabataa Cinine”, “maroo Shan mooye”. This query has also information that can be categorized as concepts like “Shaampiyoons liigii Awurooppaa”. This all information can be selected from the concepts, relationships and individuals delivered through form based interface and organized into SPARQL query as shown in Figure 5.22. During selection of those items from the list, a simple query that enables retrieval of AOSO is created. When the query is executed, specific item or set of items that fulfil the query are retrieved. Having this result from the output of query execution, it can further be employed to search for more information from AO document corpus. However, in the direct use of the query as it is, many unrelated documents are retrieved. It has general information which resulted in the retrieval of unrelated documents from the document corpus. Concepts like “Tabataa kubbaa miilaa” resulted in the

retrieval of documents that have this keyword. It can be about any player as far as the word "Tabataa Kubbaa Miilaa" is mentioned in the document. Similarly, information like "tabataa kilabii", "maroo shan mooye", "shaampiyoons liigii Awurooppaa", "dorgommii waancaa kubbaa miilaa addunyaa bara 2014 irrattii tabataa cinine" etc., lead to the retrieval of junk of documents. In the retrieval system based on word matching techniques, all the words given herewith lead to the retrieval of documents whether the documents contain all or part of words mentioned in the query irrespective of whether the documents contain information about the item mentioned in the query.

As compared to the ontology based retrieval, which retrieves document corpus based on the output of the execution of the SPARQL query, retrieval performance that retrieves documents based on the description of an item has very poor performance. In ontology based system, execution of SPARQL query is performed against AOSO prior to the retrieval of AO document corpus which enables to deliver items that fulfil all the conditions specified in the query and hence, AO text retrieval system performs retrieval of documents that have information about these specific items that leads to good precision of retrieved documents.

Query-2: "Tabataa kubbaa miilaa kilabii biyya Afrikaa kan ture amma dargagoota kilabii QPRtiif tabachuf carraa kan argate".

A football player formerly playing in a football club found in African country and now get a chance to join QPR youth academy.

When AOSO knowledge base is queried before the actual document search is commenced, a list of semantic entities tagged as a former player of football club found in African country and now get a chance to join QPR football club youth academy is displayed. The result of this query will in turn be used as search word to query for the AO document corpus. In this way, the quality or accuracy of the retrieval system is increased as compared to the junk of list of documents returned when search is made through the description of the query as it is. 66.6% precision with the knowledge base retrieval against 33.3% when the knowledge base is not used.

Similar to query-1, we can also identify concepts like "Tabataa Kubbaa Miilaa" and "Biyya_Afrikaa" in this query. The query is about an entity or set of entities which are type

of “Tabataa_Kubbaa_Miilaa” related to concept “Biyya_Afrikaa”. This query consists of a list of relationships like “tabataa_Durii”, “tabataa_Kubbaa_Miilaa_Dargagootaa_Kilabii” and “kilabii_Biyya” relating individuals represented as variables and QPR. The SPARQL query representation of query-2 is given in Figure 5.23. In ontology aided system, an entity or set of entities fulfilling query-2 conditions are retrieved and used in the document retrieval purpose which leads to good retrieval performance. In the case of direct use of query-2 for document retrieval purpose, documents in which all or some of words mentioned above are retrieved irrespective to whether the retrieved documents are about an entity the query is looking for. Hence, it leads to poor retrieval performance.

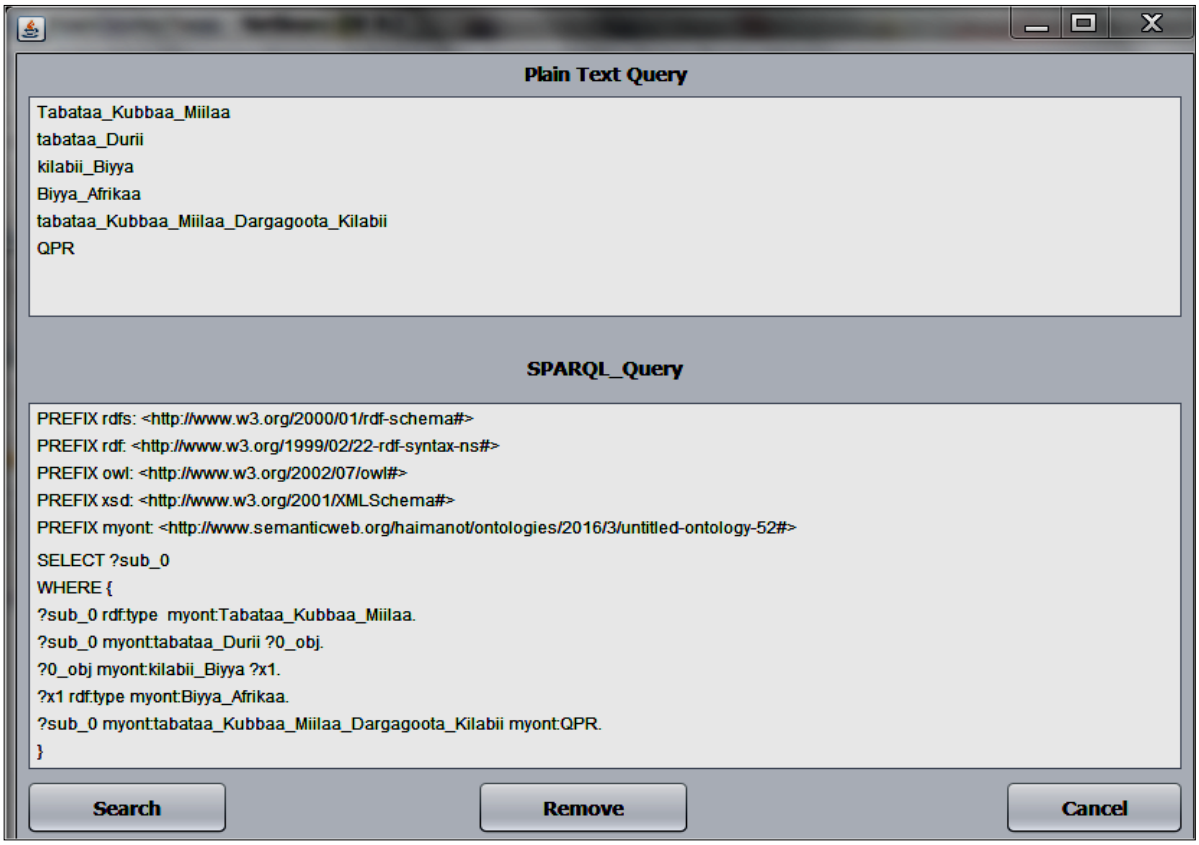


Figure 5.23 SPARQL query representation of Query-2

Query-3: “Tabataa kubbaa miilaa biyya Amerikaa kibbaa dorgommii waancaa kubbaa miilaa addunyaa Biraazil bara 2014 irratti lammaffaa baateefi tabatu, dorgommii sanirratti kubbaa warqii badhaafame”.

This query is querying about a player who played for a South American country that was a runner up on the 2014 world football cup hosted by Brazil and won golden boots on that competition. If this concept is captured in the knowledge base, a player that fulfils the specified conditions is retrieved from AOSO which in turn be used in the AO document retrieval process to improve the retrieval effectiveness. The direct use of the query for the AO document retrieval process has got a precision of 40% against 100% in the case of Ontology aided retrieval.

Query-3 comprises of concepts like "Tabataa_Kubbaa_Miilaa", and "Biyya_Amerikaa_Kibbaa" and "tabataa_Biyya", "lamaffaa_Baate", "badhaafame" are object properties relating individuals specified in the query. "Dorgommii_Waancaa_Kubbaa_Miilaa_Addunyaa_Braazil_Bara_2014" and "Kubbaa_Worqii_Dorgommii_Waancaa_Kubbaa_Miilaa_Addunyaa_Bara_2014" are individuals or objects of the concepts of the ontology. The SPARQL query representation created from the query is given in Figure 5.24.

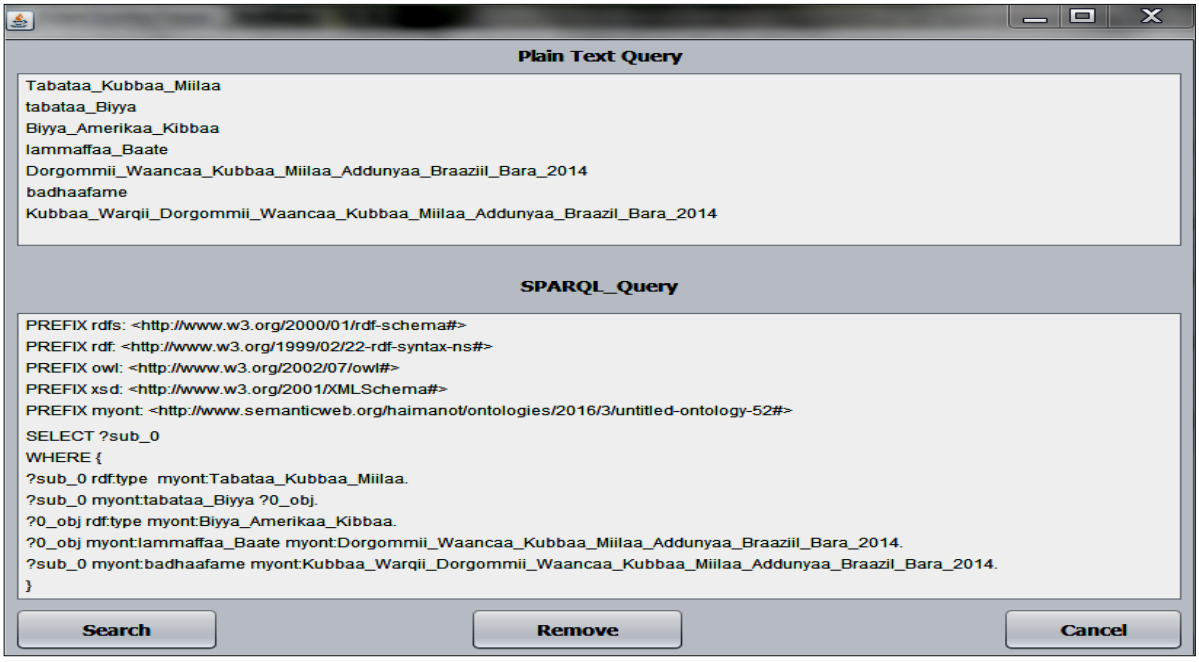


Figure 5.24 SPARQL query representation of Query-3

Query-4: "Atileetii biyya Afrikaa" baby face destroyer" jechaan beekamtu fi oboleettin ishee dorgommii metira 10000 dubartii olompiikii Ateens bara 2004 irratti lammaffaa baate."

It is a query to retrieve an athlete of African country who has a nick name of “baby face destroyer” and has a sister who had finished second on 2004 Athens Olympic women 10000 metre competition. This is also a description about an entity that is going to be queried. Ontology aided search has good precision of 83.3% as compared to 33.3% for direct use of the query.

During ontology retrieval, the system searches for an entity or entities of type “Atileetii” representing a nation of African country represented by “Biyya_Afrikaa”. Both “Atileetii” and “Biyya_Afrikaa” are concepts of AOSO. ”atileetii_Biyya”, “maqaa_Ittin_Beekamtu”, “obbolleettin_Ishee”, “lammaffaa_Taate” are object properties. “Dorgommii_Metraa_10000_Dubartii_Olompiikii_Ateens_Bara_2004” and the variables “?sub_0”, “?obj_0” and “?obj_1” represent Individuals of the concepts.

During the SPARQL query execution, the system searches from AOSO an entity or set of entities that fulfil the specified conditions. The output of the SPARQL query execution is used for the retrieval of information from AO document corpus. Direct use of the information given in the query-4 leads to the retrieval of junks of information due to broad concept specified in the query. The SPARQL query representation of the query-4 is given in Figure 5.25.

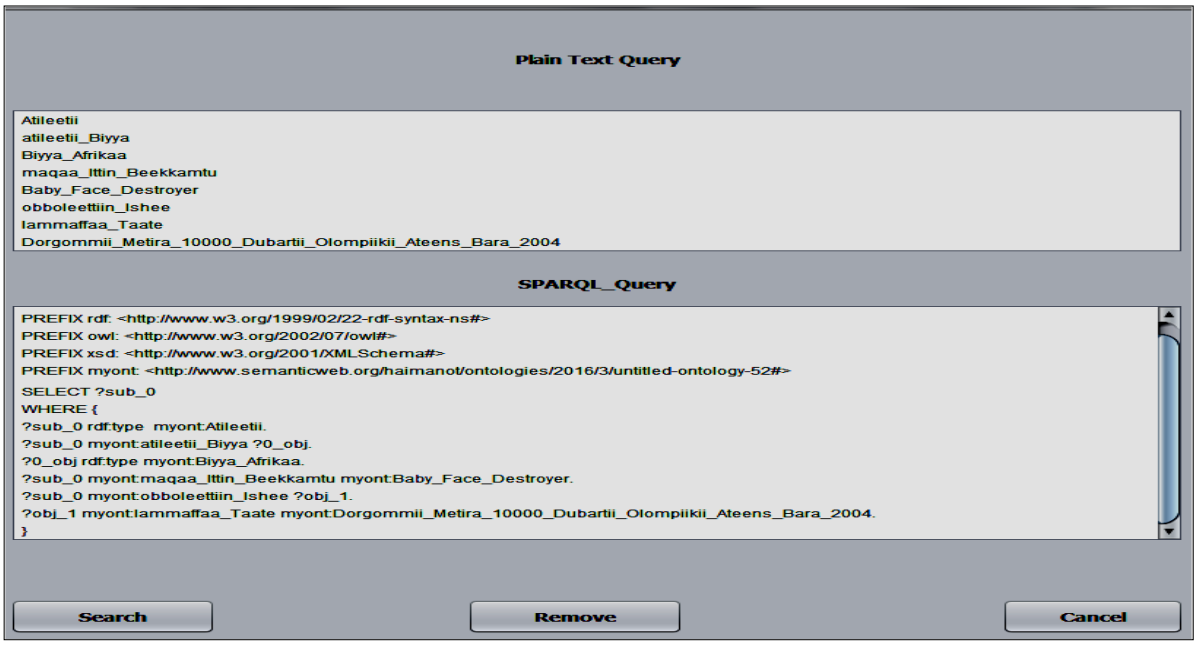


Figure 5.25 SPARQL query representation of Query-4

Query-5: “Atileetii biyya Afrikaa, abbaa rikordii figicha metira 10000 dubartii addunyaa fi olompiikii kan taate, leenjisaan ishee atileetii duriifii dhirsa ishee kan ta’e,”

This query describes about an athlete who competes for an African nation and has world and Olympic women 10000 metre record and her manager is a veteran athlete and her husband.

This is also a description of entities or an entity that is going to be searched. In the ontology aided system, description of entities or an entity is delivered to the knowledge base through the SPARQL query. Ontology aided retrieval of AO corpus delivers 100% precision as compared to 40% precision when the query is directly used for retrieval purpose.

The SPARQL query which is the representation of query-5 is composed of concepts like “Atileetii” and “Biyya_Afrikaa”, object properties like “rikordii_Isheen_Qabdu”, “leenjisaan_Ishee” and “niitii” are relating individuals of concepts specified in the query. “Figicha_Metira_10000_Dubartii_Addunya” and “Figicha_Olompiikii_Metira_10000_Dubartii” are individuals of concepts “Dorgommii_Atleetiksii” and “Dorgommii_Olompiikii” respectively. The rest individuals which are types of concepts “Atileetii” and “Biyya_Afrikaa” and are not known before the execution of the SPARQL query are represented by variables like “?sub_0” and “?obj_0”, “?obj_1”.

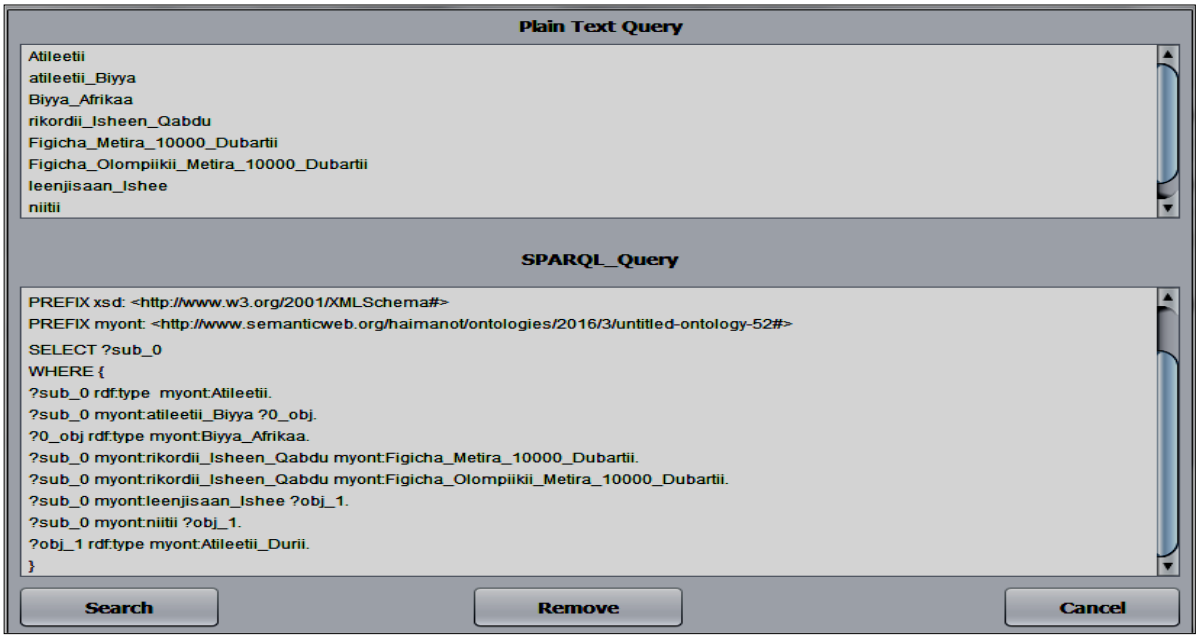


Figure 5.26 SPARQL query representation of Query-5

Execution of the SPARQL query given in Figure 5.26 retrieves set of semantic entities that fulfil the specified conditions in the where clause. The output of the SPARQL query is used in the retrieval of information from AO text corpus. In the case of keyword based retrieval, each word specified in the query serves as a search keyword which will result in the retrieval of unrelated documents.

Query-6: "Tabataa kubbaa miilaa biyya dorgommii waancaa kubbaa miilaa addunyaa Braazil bara 2014 irrattii waancaa mooye fi kilabii biyya isaa keessatti tabachaa kan jiru".

This query describes about players who represented their country that won 2014 world cup hosted by Brazil and playing for football club based in that country. The knowledge about a country that won 2014 world football cup, football clubs found in that country and players playing for those clubs and represented their nation on 2014 word cup should first be retrieved from AO sport ontology.

Similar to the above queries 1 to 5, direct application of this query for information retrieval purpose will result in poor performance. This query comprises of concepts like "tabataa", "kubbaa miilaa", "biyya", "dorgommii", "waancaa", "addunyaa", "kilabii", "taba". Documents having these keywords could be retrieved by keyword retrieval system. The outcome of the retrieval system based on matching of keywords is a set of documents that contain these words in their document index which might not have any relation with the required information. In the application of ontology, however, the ontology is retrieved for the entity about which the query describes. The result of information retrieval is a list of documents containing information about entities that fulfil all the conditions stated in the where clause of the SPARQL query. The SPARQL query consists of concepts like "Tabataa_Kubbaa_Miilaa", object properties like "tabataa_Kilabii", "tabataa_Biyyaa", "waancaa_Moote" and individuals of concepts like "Dorgommii_Waancaa_Kuubaa_Miilaa_Addunyaa_Braazil_Bara_2014" as shown in figure 5.27. Application of ontology has got better precision of 75% against 25% when it is not used.

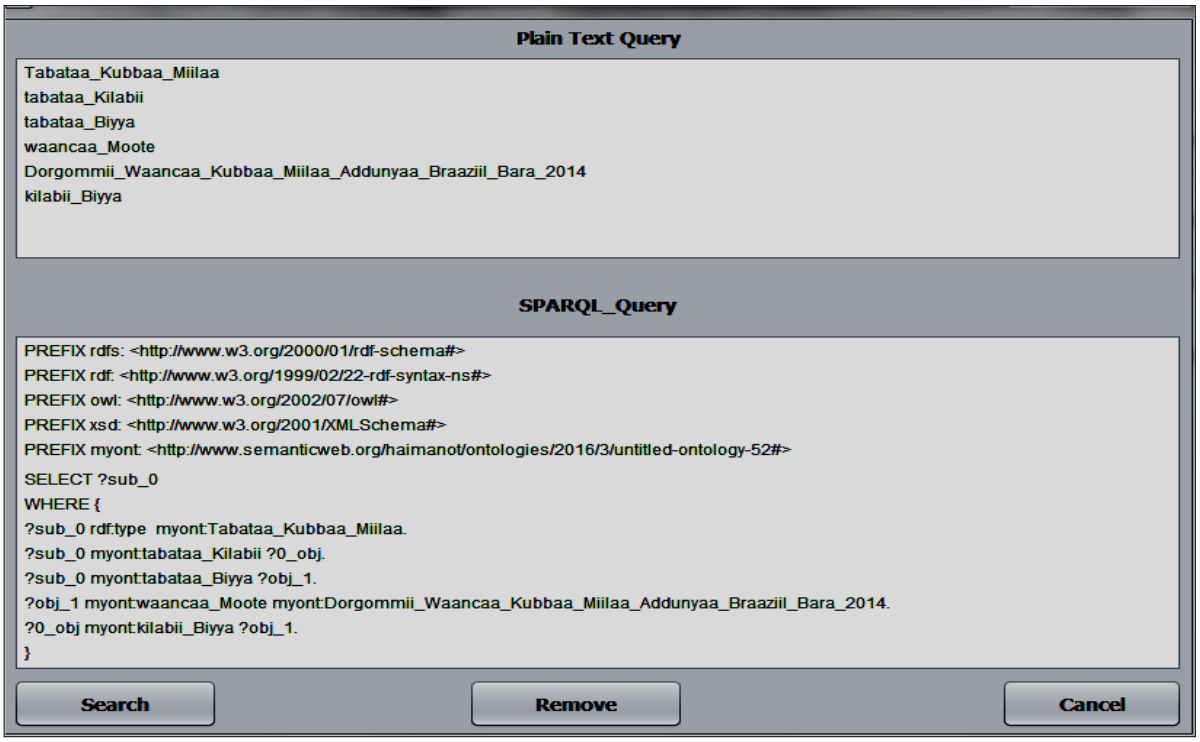


Figure 5.27 SPARQL query representation of Query-6

Query-7: “Lammii biyya Afrikaa kan Olompiikii Addunyaarrattii worqii mooye fi Maaraatonii Iskotlaand 7ffaa irratti rikordii bakkee kan cabse”.

This query requests the system to retrieve information about an Athlete who is citizen of African country and won Olympic gold medal and has broken Marathon record of a place on the 7th edition of Scotland Marathon competition. Ontology-aided retrieval has got better performance than directly using the query to retrieve the information about the concept described in the query. Ontology aided retrieval has precision of 75% against 25% precision when the query is used as it is for document searching.

The query consists of concepts like “Atileetii” and “Dorgommii_Olompiikii” as shown in figure 5.28. The entity required from the ontology retrieval is a type of “Atileetii” who won athletics competition that is type of “Dorgommii_Olompiikii”. “lammii_Biyya”, “medaaliya_Warqii_Mooye” and “rikordii_Cabse” are relationships relating objects of concepts. “Maaratonii_Iskotlaandii_7ffaa” is an object or instance of specific competition. In the direct keyword retrieval, documents indexed with “atileetii”, “lammii”, “biyya”, medaaliya”, “warqii” “rikordii” “cabse”, “iskotlaandii”, “maaraatonii” keywords are

retrieved and results in a retrieval of unrelated documents. However, in the ontology-aided retrieval system specific set of entities that fulfil the specified conditions in the where clause are retrieved from the AOSO and information retrieval system looks for documents describing about these items.

Figure 5.28 SPARQL query representation of Query-7

Chapter Six: Conclusion and Future works

6.1 Conclusion

In the system that is based on the word matching principle, performance of the retrieval system is affected by the presence of similar words in the target document. In this system, documents having words that are similar to the query words are retrieved irrespective of the content of the documents retrieved and the query. Retrieval is made when query and document have similar words. In this case, performance of the system is affected by the way the query is presented since the focus of the retrieval system is to conduct word comparison. In some cases the user may describe the entity about which the information is required instead of directly delivering the required information to the retrieval system. In such scenario, retrieving information through word matching has no satisfactory performance.

In the developed system, ontology has played key role between the delivered query and the document retrieved. The ontology is retrieved before the actual retrieval process. Even though the user delivers the query in the form of the description of the required items, these descriptions are converted into SPARQL query to query for ontology from where the semantic representation in the form of relationship, concept and individuals are retrieved which will further be employed to search for the documents. In this way, the draw back of the retrieval system that is based on word comparison principle can be mitigated.

Performance of the developed system is evaluated using precision of retrieved documents with and without the use of ontology. According to the evaluation result, precision of the retrieval aided by the ontology has by far better performance than the retrieval when the queries are used directly for document searching purpose. Ontology-based Afaan Oromo text retrieval has an average precision of 81.61% as compared to average precision of 32.16% when the queries are used as they are. Quality of retrieval performance in the case of direct use of the queries for the retrieval purpose is highly affected by the nature of the queries which describes the entities required instead of the direct use of entities for searching purpose. In the case of ontology aided retrieval, the entities about which the query describe are first retrieved from the ontology and then the search for the document comes next using the output of the ontology retrieval.

In the query processing stage, different items that are labelled by the same names (the case of polysemy) have no significant impact as the description in the query pattern excludes the unintended entity. Only an entity that fulfils all the conditions of query pattern is retrieved. Even though different items could be known by the same name, they are filtered by the query patterns given in the where clauses.

Since the use of ontology in this work is restricted to the retrieval of semantic entities on query stage, the document retrieval part which is performed by keyword matching mechanism is prone to similar naming. Query operation which retrieves "Luwis Suwaarez" from AOSO retrieves documents describing about this entity. In addition to this, the document retrieval part could retrieve documents containing entities with similar names.

Despite this fact, this thesis work proved the role ontology could play in overcoming the problem on the way queries are described by AO text retrieval process.

6.2 Future works

Developing ontology based applications like text retrieval system is very broad and is not a simple task to fully entertain in works like this. There is lack of basic prerequisites to even think about it for works bounded in a time frame. In general, we would like to forward the following recommendations for the future works:

1. There is lack of Afaan Oromo corpus that could be used as a source of concepts extraction during ontology development and used in the performance evaluation purpose. It would be better if there is standardized Afaan Oromo text corpus for the future works.
2. This work is done on sport domain, but for works like this, sport domain is very broad and it is difficult to capture all the concepts and related issues found in the domain with the available time and resources. Better work can be done by making the scope narrower so that it would be manageable in capturing most of the key concepts and instances of the concept.
3. To make Ontology-based text retrieval system complete and have a system that is less vulnerable to most of the factors that affect quality of the retrieval system, it is recommended to support the system with document annotations. Documents can be

annotated with the concepts from the knowledge base to improve the information retrieval effectiveness.

4. In this work, precision of text retrieval system is improved through application of ontology in the information retrieval system. This is true if and only if the ontology retrieval process has an output that could be used for text retrieval process. To make the system more complete for queries that have no output from the ontology, we would recommend further works to integrate ontology based application with keyword retrieval system so that the keyword retrieval could take over the retrieval process when the queries are about concepts that are not captured in the knowledge base.
5. To develop different applications that exploit ontology, we need to have ontologies on different domains. Therefore, the development of ontology on every domain should be encouraged.

References

[1] Cambridge University Press, *Boolean Retrieval*, Online edition Cambridge University Press, 2009

[2] W. Bruce Croft, D. Mtzler and T. Strohman, *Search Engines Information Retrieval in Practice*, Pearson Education , Inc, 2015

[3] C. Beeks, and T. Seidl., "Efficient-Content based Information Retrieval: A New Similarity Measure for Multimedia Data", Aachen University, Germany, in *Symposium for Future Directions in Information Access* (FDIA 2009), Germany

[4] A. Singhal., "Modern Information Retrieval: A Brief Overview", in *Bulletin of IEEE Computer Socienty Technical Committee on Data Engineering*, 2001

[5] G. Kowalski, *Information Retrieval System Theory and Implementation*, Kluwer Academic Publisher, USA, 1999

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley ACM Press, Addison-Wesley Longman Limited, Edinburgh gate, Harlow, Essex CM20 2JE, England, 1999

[7] H. Beck, and H. S. Pinto, "Overview of Approach, Methodologies, Standards and Tools for ontologies", The Agriculture ontology service (UNFAO), 2002.

[8] Gezahegn Gutema, "Afaan Oromo Text Retrieval System", Unpublished Masters Thesis, Department of Information Science, Addis Ababa University, 2012.

[9] Tesfaye Guta Debela, "Afaan Oromo Search Engine", Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, 2010

[10] K. Supekar., C. G. Chute, H. Solbrig, "Representing lexical components of medical terminologies in OWL", in *Symposium Proceedings*, pp. 719-723, Stanford University, Stanford, 2005.

[11] M. F. Sanchez, "Semantically enhanced Information Retrieval: an ontology-based approach", *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 9, No. 4, 2011, pp. 434-452

[12] H. B. Styltsvig, "Ontology-based Information Retrieval", Unpublished PHD Thesis, Computer Science Section, Roskilde University, Denmark, 2006

[13] S. Kara, "An Ontology-Based Retrieval System Using Semantic Indexing", *Information systems*, Vol. 37, Issue 4, 2012, pp. 294 – 305

[14] S. T. Dumais, "Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval", Bellcore, 1992.

[15] M. W. Berry, Z. Drmac, and E. R. Jessup, "Matrices, Vector Spaces, and Information Retrieval", 1999.

[16] Tewodros Hailemeskel, "Amharic Text Retrieval: An Experiment using Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD)", Unpublished MSc Thesis, Addis Ababa University, 2003.

[17] Girma Debele Dinegde and Martha Yifru Tachbelie, "Afan Oromo news text summarizer", *International Journal of Computer Applications, Vol. 103, No. 4, 2014, pp. 0975* - 8887

[18] Debela Tesfaye and Ermias Abebe, "Designing a Rule Based Stemmer for Afaan Oromo Text", *International Journal of Computational Linguistics (IJCL),* Vol. 1, Issue 2

[19] D. Hiemstra, *Using Language Model for Information Retrieval for Information*, Netherlands, 2001

[20] C. D. Manning, P. Raghavan, and H. Schutze., *An Introduction to Information Retrieval*, Online edition @2009 Cambridge UP, Cambridge University Press, Cambridge, England, 2009

[21] D. Hiemstra, *Information Retrieval Models*, John Wiley, Online, New York: John Wiley & Sons Ltd, 2009

[22] W. R. Hersch, D. L. Elliot, D. H. Hicham, S. L. Wolf, A. Molnar , C. Lechtensteim, "Towards new measures of Information Retrieval evaluation", in *Proceedings of the 18th Annual International ACM SIGIR conference on Research and Development in information retrieval*, pp, 164 – 170, 1995

[23] E. Ofer, M. Shaul and G. Evgenly, "Concept-based Information retrieval using explicit Semantic Analysis", *ACM Transaction on Information System*, Vol.29, No.2, Article 8, 2011

[24] E. Greengrass, *Information Retrieval: A survey*, 2000

[25] R. Cummins, "The Evolution and Analysis of Term-weighting Schemes in Information Retrieval", National University of Ireland, Galway, PHD Thesis, 2008

[26] T. Heigl., Information Retrieval in the Legal Domain, Institute for Software technology and Interactive system, 2008

[27] Y.Y. Yao, "Information Retrieval Support Systems", Boca Raton: Taylor & Francis Group, pp. 1 – 778, 2012

[28] N. Fuhr, "Probabilistic Models in Information Retrieval", SpringerLink, Vol.11, no.3, pp. 251-265, 2008

[29] V. N. Gudiva, V. V. Raghavan, W. I. Grosky, and R. Kasanagottu, *Information on the World Wide Web*, IEEE Internet Computing, http://www.cse.iitb.ac.in/, 1997

[30] N. J. Belkin and W. B. Croft, "Retrieval Techniques", in *Annual Review of Information Science and Technology (ARIST)*, Volume 22, 1987

[31] H. P. Luhn, "A statistical approaches to mechanized encoding and searching of litary information", *IBM Journal of Research and Development*, 1957

[32] A. K. Singhal, "Term Weighting revisited", Un Published PHD Thesis, Cornell University, 1997

[33] F. Crestani, M. Lalmas, C.J. Van Rijsbergen and I. Campbell, "Is this document Relevant? …Probably. A Survey of Probabilistic Models in Information Retrieval", University of Glasgow, Computing Science department, Glasgow, *ACM Computing Surveys*, Vol.30, No.4. 1998

[34] L. Khan, D. Mcleod and E. Hovy, "Retrieval effectiveness of an ontology-based model for information selection", *The VLDB Journal*, 13, 2004, pp. 71 - 85

[35] G. Salton and C. Buckley., "Term Weighting Approach in Automatic Text Retrieval", *Information Processing and Management* Vol. 24, No. 5, pp. 513 – 523, 1988

[36] Abhishek Jain, Aman Jain, N. Chauhun., V. Singh, and N. Thakur., Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model, Bharati Vidyapeeth's College of Engineering, International Journal of Computer Applications (0975 – 8887), vol. 194 – No. 6, 2017

[37] C. Carstens, "Ontology Based Query Expansion Retrieval support for the domain of education research", University of Hildesheim, 2011

[38] Van Rijsbergen C.J., *Information Retrieval (2nd edition)*, Butterworths, London, 1979

[39] T. Saracevic , "Evaluation in Information Retrieval", in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in the information retrieval*, pp, 137 – 146, 1995

[40] Solomon Nega, "Analysis of Semantic Technologies for Ethiopic Church Manuscript, Art and Music", MSC Thesis, Department of Computer Science, Addis Ababa University, 2007

[41] G. Kalem, "Semantic Web Application: Ontology-Driven Recipe Querying", A Master Thesis in Computer Engineering, Atilim University, 2005

[42] N. Gantayat, "Automated Construction of Domain Ontologies from Lecture Notes", Master of Technology, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, 2011

[43] J. Hebler, M. Fisher, R. Blace, and A. Perez-Lopez, *Semantic Web Programming*, Wiley Publishing, USA, 2009

[44] van Nguyen, "Ontologies and Information System: A Literature Survey", Command, Control, Communication and Intelligence Division Defence Science and Technology Organization, Australia, 2011

[45] N. Guarino and P. Giaretta, "Ontologies and Knowledge base: Towards a Terminological classification", In Mars, N. (Ed.), Toward very large knowledge-base knowledge Building and Knowledge sharing, IOS press, 1995

[46] T. R. Gruber, "Toward Principles for the Design of Ontologies used for Knowledge Sharing", Stanford Knowledge Systems Laboratory, 1993

[47] W. N. Borst, "Construction of Engineering Ontologies for Knowledge sharing and Reuse", PHD thesis, University of Twente, Enschede, 1997

[48] R. Studer, V. R. Benjamins and D. Fensel, "Knowledge Engineering: Principles and Methods", *Data and Knowledge Engineering*, Vol.25, pp. 161 – 197, 1998

[49] N. F. Noy and D. L. Mc Guinness, *Ontology Development 101, A Guide to create your first ontology*, Stanford University, 2001

[50] S. K. Malik, N. Prakash., and S.A.M. Rizivi, "Developing a University Ontology in Education domain using Protégé for Semantic Web", *International Journal of Engineering Science and Technology*, Vol.2 (9), pp. 4673 – 4681, 2010

[51] M. Uschold and M. Gruninger, "Ontologies: Principles, Methods, and Applications", *Knowledge Engineering Review*, Vol.2, No.11, 1996

[52] K. Weller, *Knowledge representation in the Social Semantic Web*, De Gruyter Saur, Germany, 2010

[53] M. Hadzic, P. Wongthongtham, T. Dillon and E. Chang, *Ontology-Based Multi-Agent Systems*. *Studies in Computational Intelligence*, Springer-Verlag Berlin Heindelberg, ISBM: 978-3-642-01903-6, 2009

[54] N. Guarino, "Formal Ontology and Information System", in *Proceedings of the first international Conference on Formal Ontologies in Information System* (FOIS 1998), pp. 3 – 15, Trento, Italy, 1998

[55] M. C. Daconta, L. J. Obrst, and K. T. Smith, "The Semantic Web: A guide to the future to XML. Web Services and Knowledge Management", New York, USA, Wiley, 2003

[56] A. R. Aronson, T. C. Rindflesch, A. C. Browne, Exploiting a large Thesaurus for Information Retrieval, National Library of Medicine, Rockville Pike

[57] A. Gomez-Perez., M. Fernandez., and A. J. de Vicente, "Towards a Method to Conceptualize Domain Ontologies", Madrid, Spain

[58] K. Vanitha, K. Yasudha, K. N.Soujanya, S. Venkatesh, K. Ravindara, and S. Venkata Lakshimi, "The Development Process of the Semantic Web and Web Ontology", *(IJACSA) International Journal of Advanced Computer Science and Applications,*Vol. 2, No. 7, 2011, pp. 122 -125

[59] M. Gruninger, and M. S. Fox, "Methodology for the design and evaluation of ontologies", in *Proceeding of the workshop on Basic Ontological Issues in Knowledge sharing*, IJCAI, Toronto, 1995

[60] J. Bermajo, *A Simplified Guide to Create an Ontology*, The Autonomous system laboratory, Madrid, 2007

[61] M. Fernandez Lopez, "Overview of methodologies for building ontologies", in *Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5),* Stockholm, Sweden, August 2, 1999

[62] D. Jones, T. Bench-Capon, and P. Visser, "Methodologies for ontology development", Department Computer Science, University of Liverpool, Liverpool, U.K., 1998

[63] R. Iqbal, M. A. Azmil Murah, A. Mustapha, and N. M. Sharef, "An Analysis of Ontology Engineering Methodologies: A Literature Review", *Research Journal of Applied Science, Engineering and Technology*, 6(16): pp. 2993 – 3000, 2013

[64] G. Brusa, Ma. Laura Caliusco, O. Chiotti, "A Process for Building a Domain Ontology: an Experience in the Development of a Government Budgetary Ontology", in *Australasian Ontology Workshop (AOW 2006), Hobert, Australia*

[65] M. Uschold, and M. King, Towards a Methodology for Building Ontologies, in *Proceeding of the Workshop on Basic Ontological Issues in the knowledge sharing,* 1995

[66] A. Gomez-Perez, N. Juristo, and J. Pazos, "Evaluation and Assessment of knowledge sharing technology in NJ Mars (ed.) Towards Very Large Knowledge bases": *Knowledge building and Knowledge Sharing*, KBKS 95 IOS Press 289 – 296, 1995

[67] M. S. Fox, M. Barbuceanu, M. Gruninger, "An Organization Ontology for Enterprise Modeling: Preliminary Concepts for linking structure and behavior", *Computer in Industry* 29 (1996), pp. 123 – 134

[68] M. Fernandez Lopez, A. Gomez Perez, A. Pazos Sierra, J. Pazos, "Building a chemical ontology using Methontology and ontology design environment", IEEE intelligent system and their applications 4(1) 1999

[69] D. B. Lenat and R.V. Guha, *Building Large Knowledge-Based System: Representation and inference in the CYC Project*, Addison-Wiseley publishing company, Inc,. reading, Massachusetts, 1993

[70] C. Sandeep, and S. Rao, "Integrated Approach to Ontology Development Methodologies with Case Study", *International Journal of Database Management Systems (IJDMS)* Vol. 2, No. 3, August 2010

[71] M. M. Taye, "Web-Based Ontology Languages and its Based Description Logics, Faculty of Information Technology", Philadelphia University, *the Research Bulletin of Jordan* ACM, ISSN: 2078-7952 vol. II (II)

[72] F. Baader, D. L. McGuinness, D. Nardii, and P. F. Patel-Schneider, *The Description Logic Handbook: Theory, implementation and Applications*, Cambridge University Press, 2008

[73] F. Baader, and U. Sattler, "An overview of tableau algorithms for description logics", LuFG Theoretical Computer Science, RWTH Aachen, Germany, 2000

[74] D. Calvanese, G. De Giacomo, M. Lenzerini, and D. Nardi, *Reasoning in expressive description logics*, Handbook of Automated Reasoning chapter 23, pages 1581-1634, Elsevier Science Publishers, Amsterdam, 2001

[75] I. Horrocks, P. F. Patel-Schneider, D. L. McGuinness, C. Welty, OWL: "A Description Logic Based Ontology Language for the Semantic Web", 2003

[76] O. Korocho., A. Gomez-Perez, "A Roadmap to Ontology Specification language", in *12th International Conference on Knowledge Engineering and Knowledge Management, Madrid, Spain*

[77] T. Bray, J. Paoli , C.M. Sperberg, and E. Maler (ED), *Extensible Markup Language (XML) 1.0*, Second Edition, W3C Recommendation, http://www.W3C.org/TR/REC-xml, 2000

[78] D. Kalibatiene, and O. Vasilecas., "Survey on Ontology Languages", *LNBIP 90, pp. 124 – 141, 2011*

[79] D. Brickley and R.V. Guha, "Resource Description Framework (RDF) Schema Specification 1.0", retrieved from http://www.w3.org/TR/2000 CR-rdf-Schema-20000327/, last accessed on 25/07/2016

[80] X. Su, and L. Ilebrekke, "A Comparative Study of Ontology Languages and Tools", Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[81] D. Fensel, I. Horrocks, F.Van Harmelen, S. Decker, M. Erdmann, and M. Klein, "OIL in a Nutshell"

[82] M. Horridge, *Practical Guide to Building OWL Ontologies Using Protégé 5 and CO-ODE Tools*, edition 1.1, University of Manchester, 2007

[83] E. Alatrish, "Comparison of some of Ontology Editors*", Management Information System* vol. 8, No. 2, 2013, pp. 88 – 22

[84] S. C. Buraga, L. Cojokura, and O. C.Nichifor, "Survey of Web Ontology Editing Tools", Faculty of Computer Science, University of Iasi, Berthelot Street, 16 Iasi, Romania, 2006

[85] Parveen, D. K. Sahni, D. Khurana, and R. Nandal, "Ontology Development Tools and Languages: A Review", M. Tech. (CSE), UIE, MDU, Rohtak, Haryana, *International Journal of Enhanced Research in Management and Computer Applications*, ISSN: 2319 – 7471, Vol. 5, Issue 6, 2016, pp. 92 - 96

[86] K. Dentler, R. Cornet., A. ten Teije., and N. de Keizer.," Comparison of Reasoners for Large Ontologies in the OWL 2 EL Profile", *Semantic Web Journal, 2011, pp. 1-5*.

[87] S. Abburu, "A Survey on Ontology Reasoners and comparison", *International Journal of Computer Application* (0975 – 8887), Vol. 57, No. 17, 2012, pp. 33 - 39

[88] D. Wu and H. Wang, "Role of Ontology in Information Retrieval", *Journal of Electronic Science and Technology of China*, vol. 4, no. 2, 2006, pp. 148 - 154

[89] J. Saias, and P. Quaresma, "A Methodology to Create Ontology-based Information Retrieval System", Departamento de Informatica, Universidide de Evora, Portugal

[90] A. Jamgade, and S. Karale, "Ontology Based Approach for Semantic Information Retrieval System", *INTERNATIONAL JOURNAL FOR TRENDS IN ENGINEERING AND TECHNLOGY* vol. 4, Issue 1, ISSN: 2349 – 9303, pp. 67 – 72, 2015

[91] P. Yadew and R.P. Singh, "An Ontology-based Augmented method for document retrieval", *International Journal of Computer Application* (00975- 8887), Vol. 53, No. 17, 2012

[92] Kula Kekeba, V. Varma, and P. Pingali, "Evaluation of Oromo-English Information Retrieval, in *International Joint Conference on Artificial Intelligence (IJCAI)-2007, January 12, 2007, Hyderabad, India*

[93] Daniel Bekel, R. Babu, and Dereje Teferi, "A Cross-Lingual Information Retrieval (CLIR) System for Afaan Oromo-English using a Corpus-based Approach", *International Journal of Engineering Research and Technology (IJERT*), ISSN: 2278 – 0181 Vol. 4 Issue 05, 2015

[94] Eyob Nigussie, "Afaan Oromo-Amharic Cross Lingual Information Retrieval: A Corpus based Approach", Unpublished MSc Thesis, Department of Information Science, Addis Ababa University, 2013

[95] Meron Sahilemariam, "Concept-based Automatic document categorization", Unpublished MSc Thesis, Department of Computer Science, Addis Ababa University, 2009

[96] J. Paralic, and I. Kostial, "Ontology-based Information Retrieval", Department of Cybernetics and AI, Technical university of Koisice, Letna 9, 040 11 Koisice, Slovakia, 2003

[97] L. R. Khan, "Ontology-based Information Selection", PHD Thesis, Computer Science, University of Southern California, 2000

[98] N. N. Aung, and T. T. Naing, "Sports Information Retrieval with Semantic Relationships of Ontology", University of Computer Studies, in *3rd International Conference on Information and Financial Engineering*, IPEDR Vol-12, pp: 86-92, 2011, Singapore

[99] R. Lakshmi Tulasi, M. Srinivas Rao, and G. Rayana Gouda, "Ontology based Information Retrieval: A Case Study for Sports and Eminent Personalities Domain". *International Journal of computer Applications (*0975 – 8887), Vol. 39, No. 11, 2012

[100] Gaihua Fu, C. B. Jonhes and A. I. Abdlmoty, "Ontology-based spatial Query Expansion in Information Retrieval", School of Computer Science, Cardiff University, Cardiff, UK

[101] S. Rajasurya, T. Muralidharan., S. Devi, and Dr. S. Swamynathan, "Semantic Information Retrieval Using Ontology in University Domain", Department of Information and Technology, College of Engineering, Guindy, Anna Univiersity, Chennai-25

[102] M. McCandless, E. Hatcher and O. Gospodnetic, *Lucene in Action*, 2nd Edition, manning Publishing co., USA, 2010

[103] Debela Tesfaye, "Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach", Unpublished MSc Thesis, Department of Information Scinec, Addis Ababa University, 2010

[104] Class Similarity, URI:https://lucene.apache.org/core/3-6-2/api/org/apache/lucene/search/ Similarities.html, last accessed, March-25-2016

[105] E. Sirin., B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A Practical OWL-DL Reasoner", University of Maryland, MIND Lab, 8400 Baltimore Ave, College Park MD 20742, USA, 2003

[106] Farquhar, A., Fikes, and R. Rice, "The Ontolingua Server: A Tool for Collaborative Construction", in *Proceedings of KA W 96*, Banff, Canada, 1996.

[107] H. Rajagopal., "JENA: A Java API for Ontology Management", *Colorado Software summit*, IBM Corporation, 2005

[108] M. Hert., "Semantic Web Engineering", Department of Informatics, University of Zurich, @ Gerald Reif HS 2010

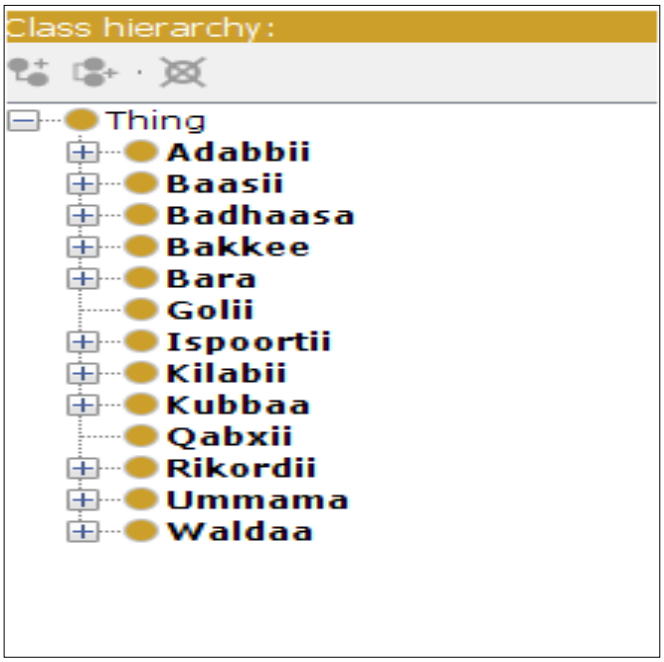
Annex A: Afaan Oromo Stop words

The following are stop words that do not add value for Afaan Oromo Text Retrieval system and need to be removed prior to indexing.

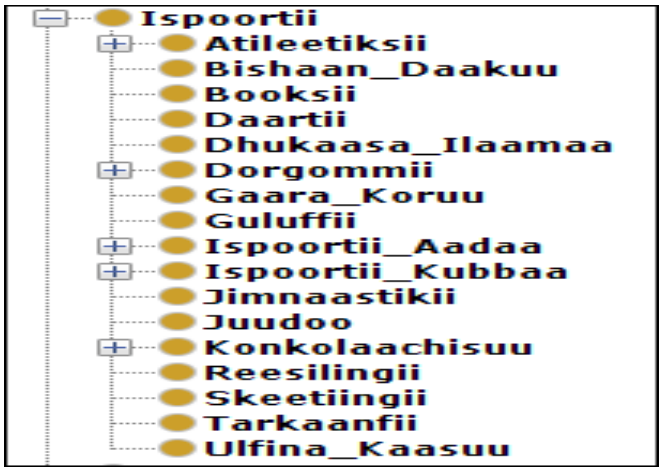
aanee	eenaa	irra	isiniif	keessatti	oo	tawullee
agarsiisoo	erga	irraa	isiniin	kiyya	osoo	teenya
akka	ergii	irraan	isinirraa	koo	otoo	teessan
akkam	f	isa	isinitti	kun	otumalle	tiyya
akkasumas	faallaa	isaa	ittaanee	lafa	otuu	too
akkum	fagaatee	isaaf	itti	lama	otuullee	tti
akkuma	fi	isaan	ittumallee	malee	saaniif	utuu
ala	fkn	isaani	itu	manna	sadi	waa'ee
alatti	fullee	isaanii	ituullee	maqaa	sana	waan
alla	fuullee	isaanitiin	jala	moo	saniif	waggaa
amma	gajjallaa	isaanirra	jara	na	si	wajjiin
ammo	gama	isaanitti	jechaan	naa	sii	warra
an	gararraa	isaatiin	jechoota	naaf	siif	woo
ana	garas	isarra	jechuu	naan	siin	yammu
ani	garuu	isatti	jechuun	naannoo	silaa	yeroo
ati	gidduu	isee	kan	narra	simmoo	yommii
bira	gubbaa	iseen	kana	nati	sinitti	yommu
booda	ha	ishee	kanaa	nuu	siqe	yoo
boodde	hamma	isheen	kanaaf	nu'i	sirraa	yookiin
dabalatees	hanga	ishii	kanaafi	nurra	sitti	yoom
dhaan	henna	ishiif	kanaafuu	nuti	sun	ufuu
dudduuba	hoggaa	ishiin	kanaan	nutti	tahullee	
dugda	hogguu	ishiirra	kannatti	nuu	tana	
dura	hoo	ishiitti	karaa	nuuf	tanaaf	
duuba	illee	isii	kee	nuun	tanaafi	
eega	immoo	isiin	keenna	nuy	tanaafuu	
eegana	ini	isin	keenyaa	odoo	ta'ulle	

eegasii	innaa	isini	keessa	ofi	ta'uyyu
yoolinimoo	inni	isinii	keessan	ogga	ta'uyyuu

Annex B: Concepts in AOSO



These are top level concepts indentified during Afaan Oromo Sport Ontology development process. "Ispoortii" is one of the concepts that contain most of sport related activities.



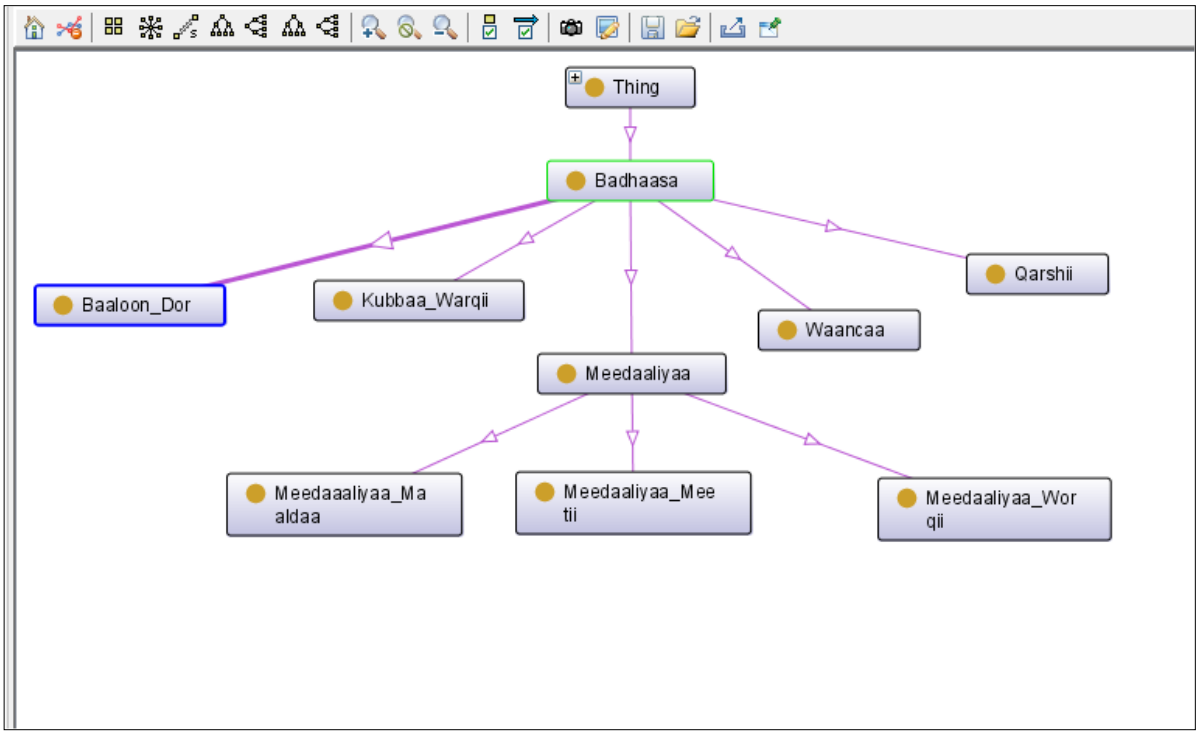
This concept consists of numerous sport related concepts that are categorized under it. Subclasses of class "Ispoortii" includes "Atileetiksii", "Bishaan_Daakuu", "Booksii",

“Daartii”, “Dhukaasa_Ilaamaa”, “Dorgommii”, “Gaara_Koruu”, “Guluffii”, “Ispoortii_Aadaa”, “Ispoortii_Kubbaa”, “Jimnaastikii”, “Juudoo”, “Konkolaachisuu”, “Reesilingii”, “Tarkaanfii”, and “Ulfina_Kaasuu”

“Dorgommii” sub class is one of the major sub concepts of “Ispoortii” concept. “Dorgommii” means competition and major international sporting events like world cup, Olympics, international athletics events, continental cup of nations, club competitions etc. “Ispoortii” concept lies at the core of AOSO and most of the concepts of the developed ontologies are categorized under this concept.

“Badhaasa” class is related to different prizes given for winning competitions or showing exceptional performance on any sporting event.

Hierarchical representation of class “Badhaasa”

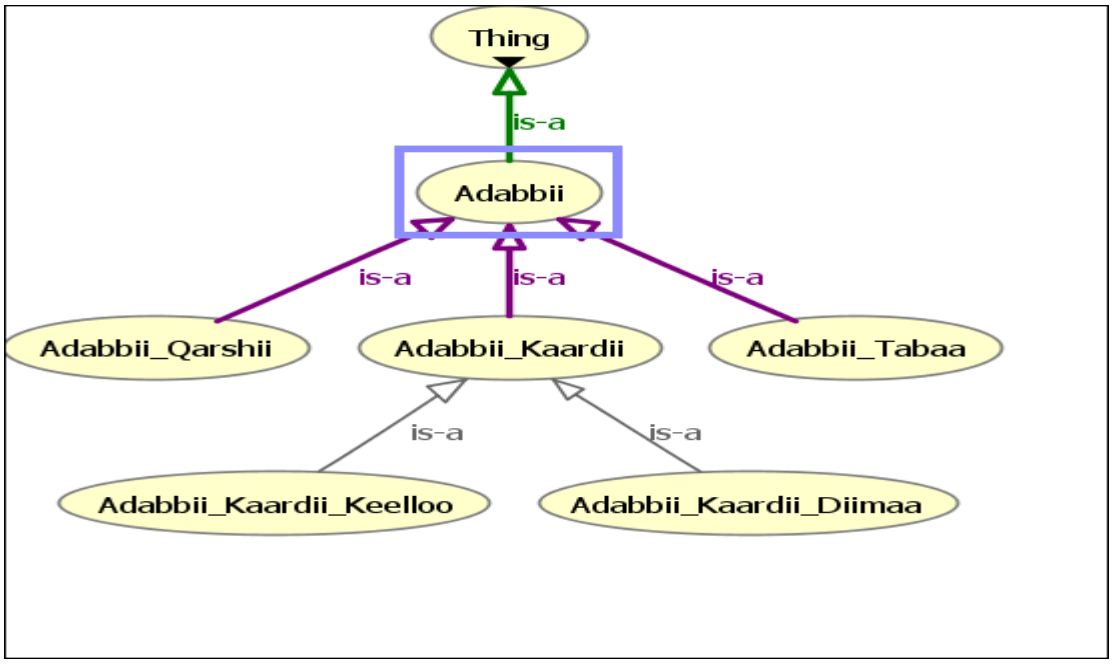


Graphical representation of class "Badhaasa" showing is-a relationship between class and subclass.

"Bakkee" is a concept related to a place like continent, country or sporting area like stadium etc.,

Hierarchical representation of class "Bakkee"

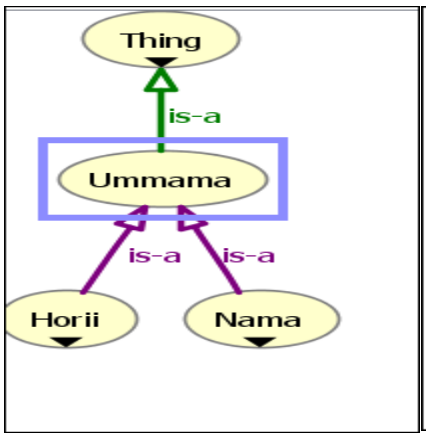
Class “Adabbii” is related to punishment for any infringement in sport domain



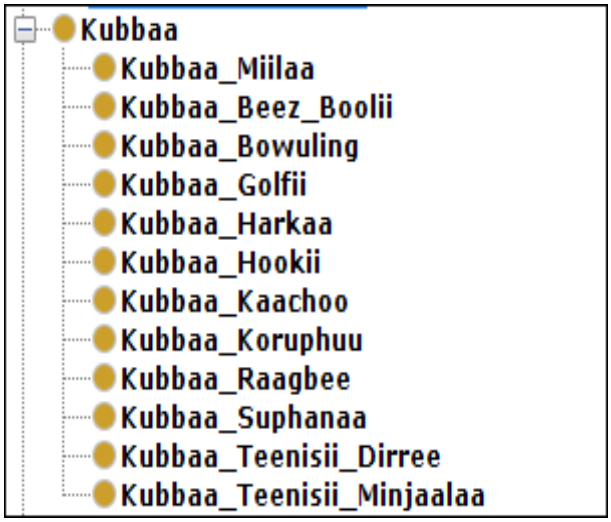
Graphical representation of class “Adabbii” and its subclasses.

Class “Ummama” is about actors or participants of any sport activities. It has subclass of “Nama” and “Hori”. Class “Nama” is to mean Person and “Hori” is animals like horses involving in sport.

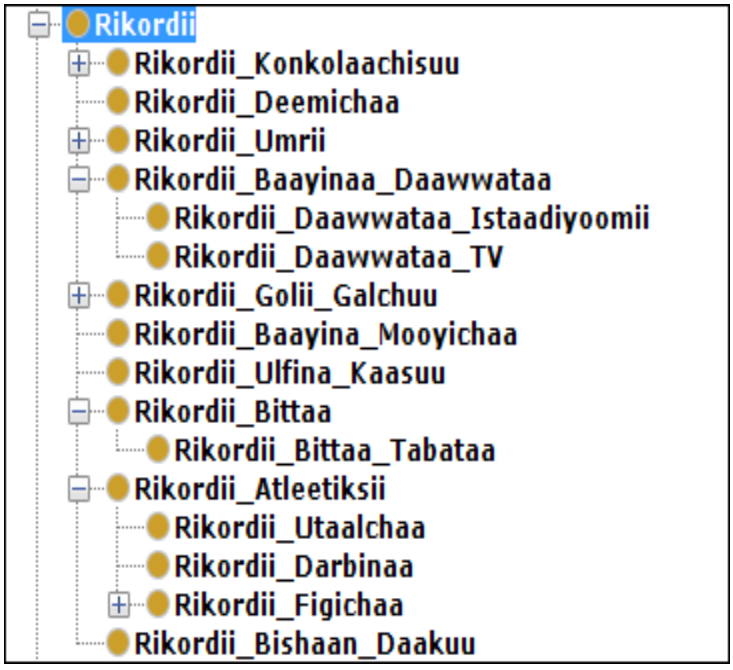
Class “Nama” comprises of subclasses like “Atileetii”, “Bakka_Bu’aa”, “Tabataa”, “Bulchinsa_Ispoortii”, “Deeggartoota” (supporters or fans), “Leenjisaa”, “Gaazexeessaa” etc.,



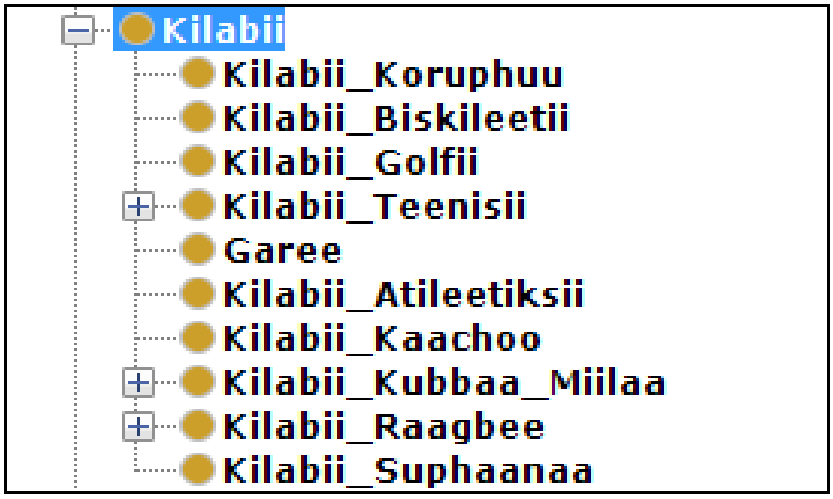
Class “Kubbaa” is about different balls used for sport competition or executing a game. Football, basketball, volleyball, tennis ball etc. are subclasses of this concept.



Class “Rikordii” deals with different records broken during sport competitions. Among the subclasses of “Rikordii” concept, records of Athletics “Rikordii_Atleetiksii” are the most common types of records known to most of sport family.



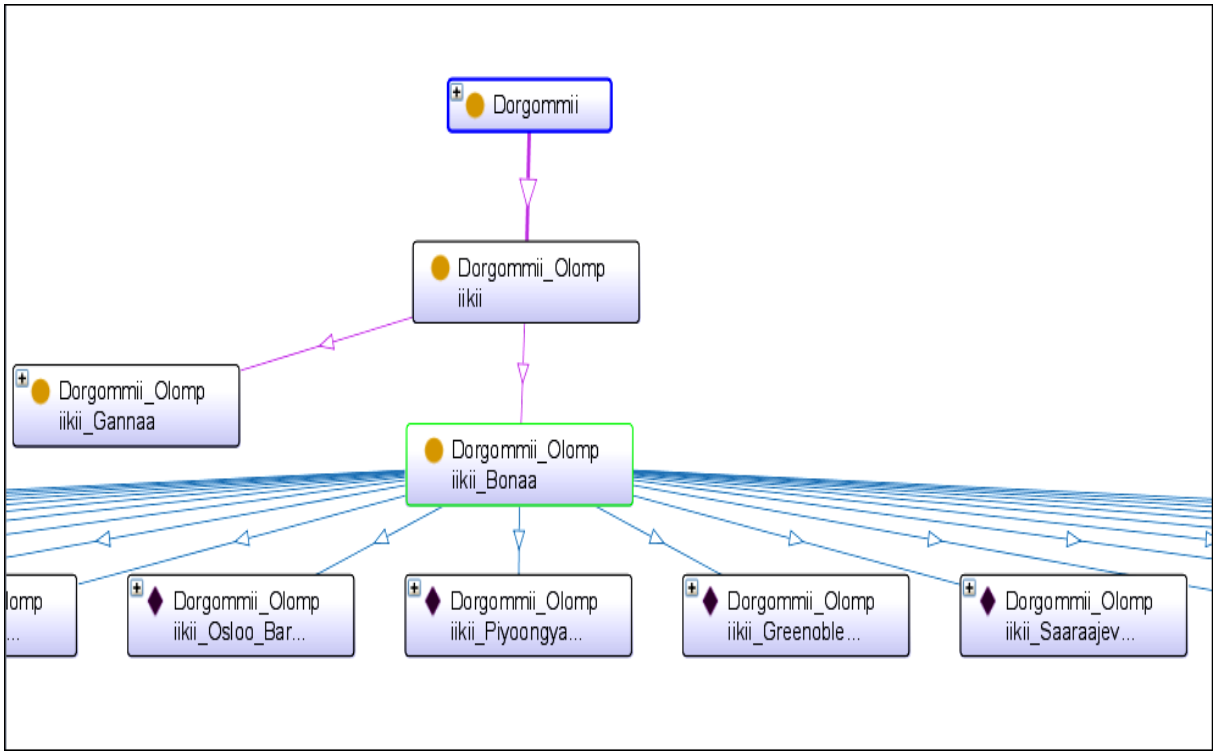
“Kilabii” concept deals with the organization of group of people for the purpose of undertaking sport.

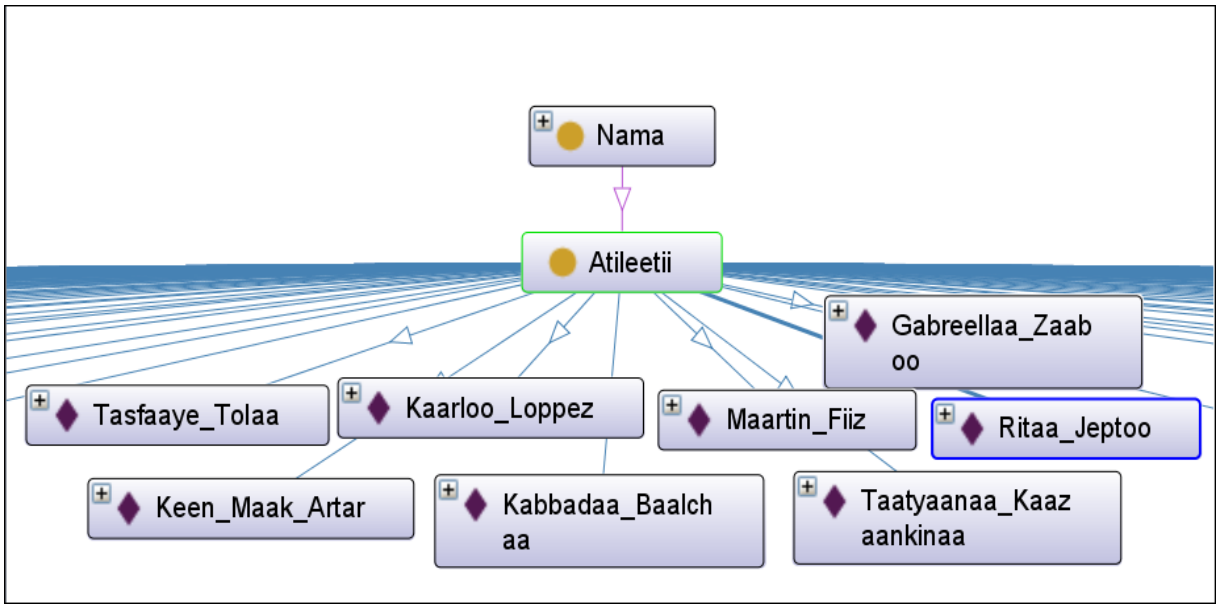


“Waldaa” concept deals with the organizations like federations who is responsible to lead, organize the overall sport related operations.

Annex C: Objects in AOSO

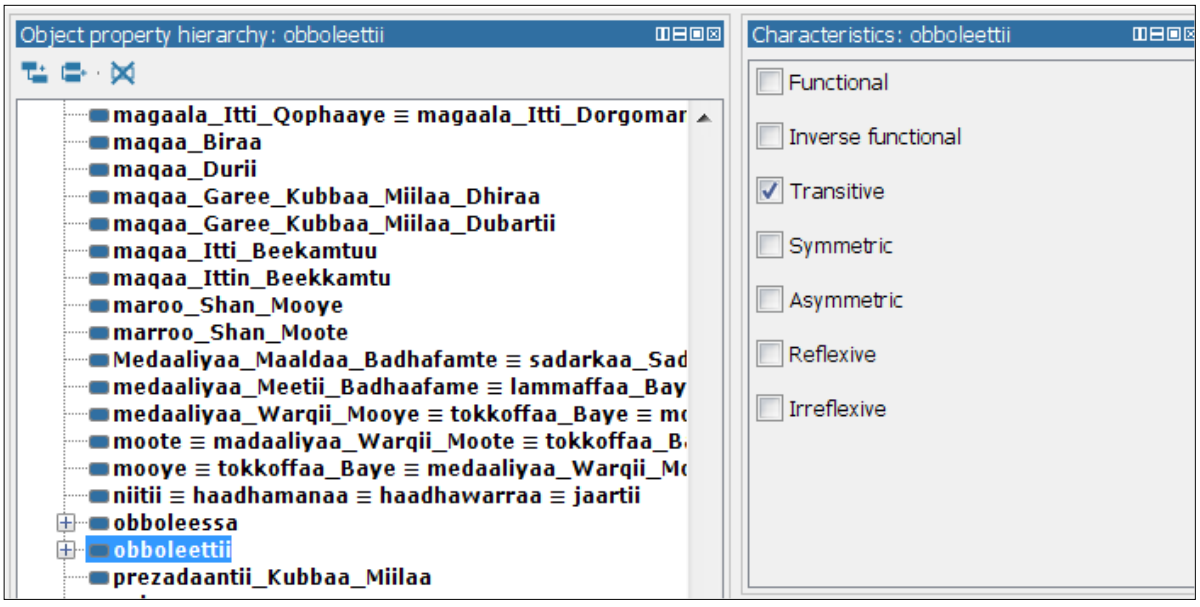
“Dorgommii_Olompiikii_Bonaa” or “Winter_Olympic_Competition” is one of the sub classes of “Dorgommii” concept. 23 individuals were categorized in AOSO as a type of “Dorgommii_Olompiikii_Bonaa”. Similarly, class “Atileetii” is a concept under which 244 individuals are populated are examples of concepts and their instances.



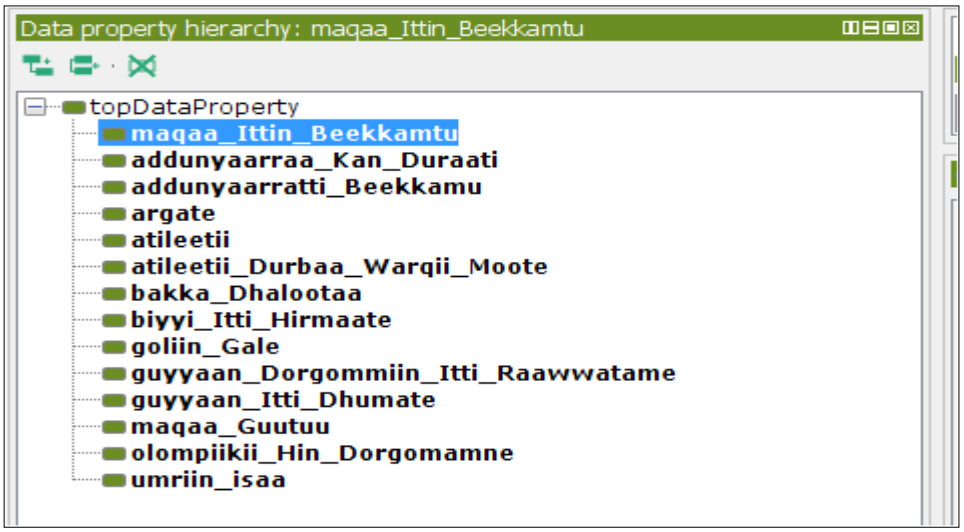


Annex D: Relationships in AOSO

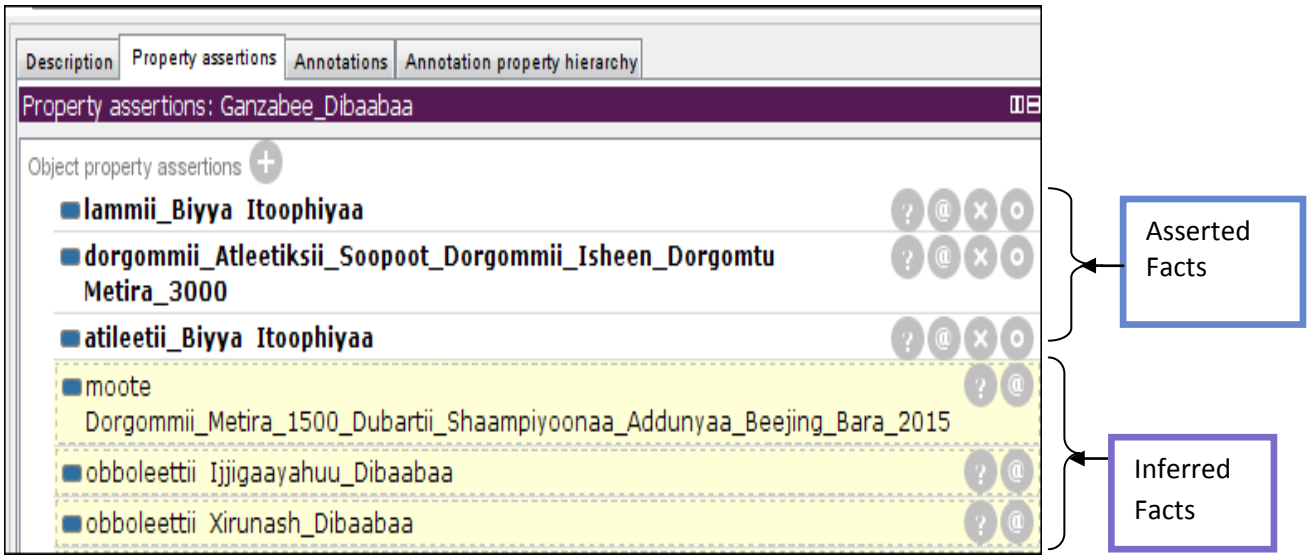
There are two types of relationships in AOSO. Object properties and data type properties. Object properties related two individuals. In Afaan Oromo sentence “Xirunash oboleettii Ganzabee” . “oboleettii is a property relating individual “Xirunash” with “Ganzabee”

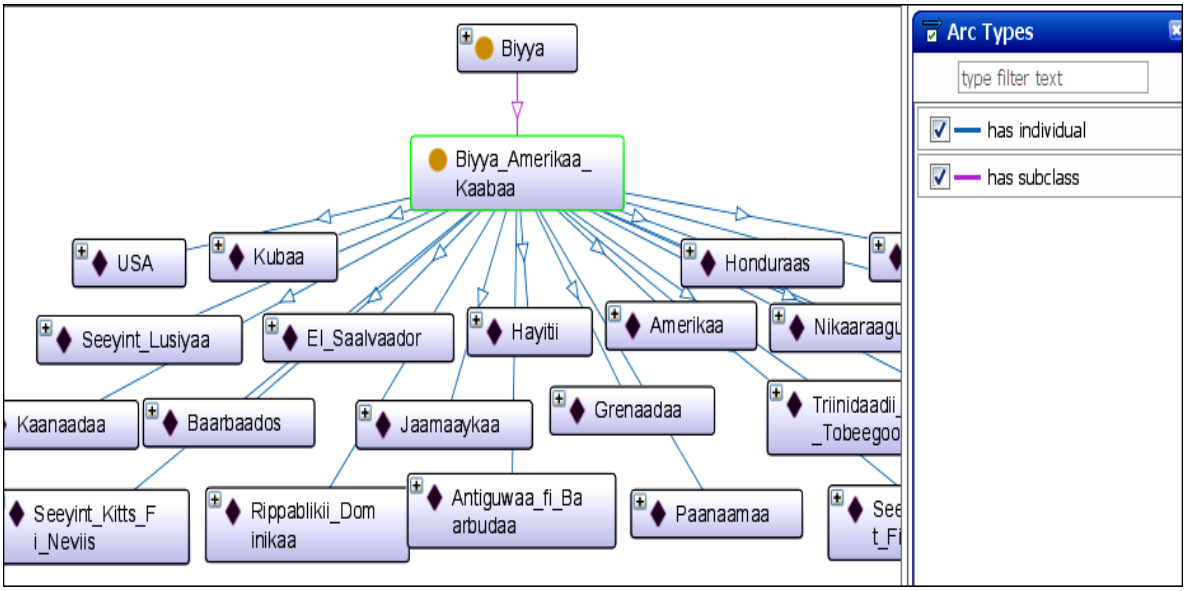


Data type properties relate individuals with the data value. In "Xirunash maqaa_Ittin_Beekkamtu baby face destroyer". "maqaa_Ittin_Beekkamtu" is data type property relating object "Xirunash" and data value "baby face destroyer".



Two types of property assertions are identified in AOSO knowledge base called asserted and inferred relationships. Asserted relationships are relationships between individuals or concepts asserted by the ontology developer where as inferred relationships are relationships inferred by reasoner.





This figure shows the graphical representation of relationships between concept and individuals.

Declaration

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: Haimanot Kebede

Signature: _____

Date: _____

Confirmed by advisor:

Name: Yaregal Assabie (PHD)

Signature: _____

Date: _____