



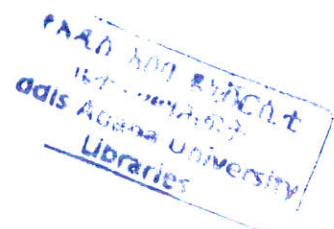
Addis Ababa University
School of Graduates Studies
College of Humanities, Language Studies, Journalism
and Communication Department of Linguistics

Amharic DBpedia Extraction

A THESIS SUBMITTED TO
THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE IN COMPUTATIONAL LINGUISTICS

Biniyam Getahun

March, 2015

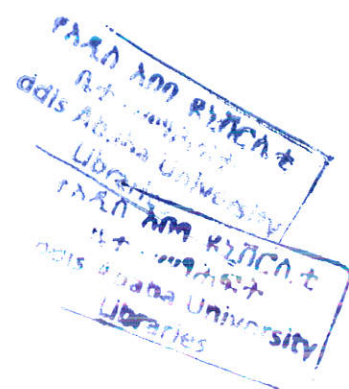


ACKNOWLEDGEMENT

First and foremost, I am ever grateful to the omnipresent God who made things possible for me to complete my study. My deepest heartfelt gratitude goes to my advisors Prof. Baye Yimam and Ato Demeke Asres for their critical comments and helpful advice.

My special thanks go to Ato Melkamu Beyene, for sharing his precious time and energy in helping me understand concepts in this study. I also would like to thank my friends and colleagues who have supported me by providing good working environment and moral support.

Last but not least, I would like to thank the DBpedia Community members for their continuous support, without them this work would have been impossible.



Contents

ACKNOWLEDGEMENT.....	I
LIST OF TABLES.....	VI
LIST OF FIGURES.....	VII
LIST OF APPENDICES.....	IX
ABBREVIATIONS.....	X
ABSTRACT.....	XI
CHAPTER ONE.....	1
INTRODUCTION.....	1
1.1. Background.....	1
1.2. Statement of the Problem.....	3
1.3. Objective of the Study.....	4
1.3.1. General Objectives.....	4
1.3.2. Specific Objectives.....	4
1.4. Scope and Limitation of the Study.....	4
1.5. Methodology.....	5
1.5.1. Data Source.....	5
1.5.2. Development Tools.....	5
1.6. Evaluation.....	7
1.7. Significance of the Study.....	7
1.8. Organization of the Thesis.....	8
CHAPTER TWO.....	9
LITERATUE REVIEW.....	9
2.1. Semantic Web.....	9

2.2. Semantic Web Technologies.....	10
2.2.1. Extensible Markup Language (XML)	10
2.2.2. Resource Description Framework (RDF).....	10
2.2.3. RDF Versus XML	11
2.2.4. The Web Ontology Language (OWL).....	11
2.3. Architecture of Semantic Web.....	11
2.4. Linked Data.....	13
2.5. Wikipedia.....	15
2.5.1. Structure of Wikipedia.....	16
2.5.2. Wikipedia's Community	18
2.5.3. Language Editions of Wikipedia	19
2.5.4. Accuracy of Contents in Wikipedia.....	19
2.5.5. Software Operations and Support of Wikipedia.....	20
2.5.6. Scientific Use of Wikipedia.....	20
2.6. DBpedia	21
2.7. DBpedia Extraction.....	21
2.7.1. Dump Based Extraction.....	21
2.7.2. Live Extraction	21
2.7.3. DBpedia Information Extraction Framework (DIEF)	22
2.7.4. I18 DBpedia Information Extraction Framework (I18n DIEF).....	22
2.8. Amharic Writing System	23
CHAPTER THREE	25
RELATED WORKS.....	25
3.1. Internationalization of DBpedia.....	25
3.1.1. Greek DBpedia (Dimitris Kontokostas, 2012)	25

3.1.2. Korean DBpedia (Kim, 2010).....	26
3.1.3. Germany DBpedia (Sebastian Hellmann, 2012).....	27
3.1.4. Arabic DBpedia (Al-Feel, 2013).....	27
CHAPTER FOUR.....	29
RESEARCH DESIGN.....	29
4.1 Design of Amharic DBpedia Extraction	29
4.1.1. Wikipedia Dump	30
4.1.2. Local Amharic Wikipedia.....	30
4.1.3. Amharic Wikipedia Page Structure	31
4.1.4. Amharic Wikipedia Infobox	32
4.1.5. Page Filter	33
4.1.6. Parsers	34
4.1.7. Extractors	34
4.1.8. Amharic DBpedia RDF Serialization	35
4.1.9. Triple Dump.....	36
4.2 Amharic Wikipedia Infobox Application Issues.....	36
4.3 Amharic Infobox Improvement	40
CHAPTER FIVE	41
EXPERIMENTATION AND EVALUATION	41
5.1 Architecture of DBpedia Extraction Framework.....	41
5.2 Installation and Extraction of DBpedia Extraction Framework	43
5.3 Extracted RDF Triple Representation.....	46
5.4 Amharic RDF Serialization.....	46
5.5 Extraction Discussion	48
5.5.1. Amharic Infobox Extraction and RDF Representation	48

5.5.2. Amharic Inter-language Link Extraction	51
5.5.3. Amharic DBpedia RDF Graph.....	52
5.6 Extraction Results	53
5.7 Amharic and Other Language DBpedia Statistics	55
5.8 Amharic Infobox Quality Improvement	55
5.9 Evaluation	56
5.9.1. Evaluation of Extracted RDF Triples against Source on Amharic Wikipedia	56
5.9.2. Manual Evaluation.....	57
5.10 Evaluation Results and Discussion	57
5.10.1. Evaluation Result of Extract RDFs Triples against Source on Amharic Wikipedia.	57
5.10.2. Manual Evaluation Result	58
5.11 Result Discussion.....	59
CHAPTER SIX	61
CONCLUSION AND RECOMMENDATION	61
6.1 Conclusion	61
6.2 Recommendation	62
REFERENCES.....	63
ANNEXES	66

LIST OF TABLES

Table 2.1 XML Tag Example	10
Table 2.2 Top ranked global language edition in Wikipedia.....	19
Table 2.3 Top ranked Ethiopian language edition in Wikipedia	19
Table 5.1 Extracted triples	54
Table 5.2 Amharic RDF extraction test result	58
Table 5.3 Manual evaluation result.....	59

LIST OF FIGURES

Figure 2.1 Layered architecture of semantic web	12
Figure 2.2 Excerpt of sample open linked data cloud.....	14
Figure 2.3 Wikipedia structure	16
Figure 4.1 Amharic DBpedia extraction design	30
Figure 4.2 Amharic Wikipedia URL page structure	31
Figure 4.3 Existing Amharic infobox template.....	32
Figure 4.4 Example of Amharic Wikipedia source	33
Figure 4.5 Extractor flow diagram.....	35
Figure 4.6 Excerpt of English person_politician infobox template	38
Figure 4.7 Amharic person_politician infobox.....	38
Figure 4.8 Excerpt of English country infobox template.....	39
Figure 4.9 Amharic country infobox template.....	39
Figure 4.10 Flow diagram of Amharic infobox improvement.....	40
Figure 5.1 Overview of Amharic DBpedia Extraction Framework.....	42
Figure 5.2 Sample Amharic DBpedia RDF with namespace	43
Figure 5.3 Local Amharic Wikipedia	45
Figure 5.4 Amharic DBpedia RDF Example.....	46
Figure 5.5 Sample Amharic N-Triples format.....	47

Figure 5.6 Example of Turtle Serialization.....	47
Figure 5.7 An excerpt of Amharic Wikipedia Infobox (Ethiopia).....	50
Figure 5.8 An excerpt of Amharic Wikipedia and DBpedia infobox alignment (Ethiopia).....	50
Figure 5.9 Sample Inter-language link pictorial representation.....	51
Figure 5.10 Sample Amharic inter-language RDF triples	52
Figure 5.11 Sample DBpedia RDF Graph	53

LIST OF APPENDICES

Annex A: Amharic Article Category RDF Triples..... 66

Annex B: Amharic Article Template RDF Triples 67

Annex C: Amharic Category Label RDF Triples 68

Annex D: Amharic External Link RDF Triples 69

Annex E: Amharic Geo-coordinate RDF Triples 70

Annex F: Amharic Infobox Property RDF Triples..... 71

Annex G: Amharic Infobox Property Definition RDF Triples..... 72

Annex H: Amharic Inter-language Link RDF Triples... .. 73

Annexes I - Amharic Infobox Templates... .. 74

Annexes P – List of Updated Infobox Property Values... .. 75

Annexes K – Pom.xml File Customization 76

ABBREVIATIONS

AST	Abstract Syntax Tree
DIEF	DBpedia Information Extraction Framework
DTD	Document Type Definition
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
ILL	Inter-lingual Link
NLP	Natural Language Processing
OKFN	Open Knowledge Foundation Network
OWL	Web Ontology Language
RDF	Resource Description Framework
SW	Semantic Web
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language
SKOS	Simple Knowledge Organization System
SPARQL	Sparql Protocol and RDF Query Language
SQL	Standard Query Language

ABSTRACT

Knowledge base is a technology used to store complex structured and unstructured data used by computer. Today, most knowledge bases cover just particular domain that is created by a small group of knowledge engineers because building general domain base knowledge is costly and time taking to cover all domains. Wikipedia has developed into one of the focal knowledge sources for everyone and is kept up by a large number of contributors but its structure has some issue to use as knowledge source. The DBpedia project goes for extracting information based on semi-structured information by presenting Wikipedia articles, interlinking it with other knowledge bases, and publishing it as RDF triples openly on the Web. So far, the DBpedia project has succeeded in creating one of the largest knowledge bases on the web data, which is used in many applications and research prototypes.

DBpedia extraction is extracts structured data (RDF) from Wikipedia. This study describes the effort to extract Amharic DBpedia. During the extraction process, the extraction design present by considering Amharic language. The tool used to extract Amharic DBpedia is I18n extraction framework. The result shows more than quarter million Amharic RDF triples extracted. In addition to this achievement, the improvement of Amharic Wikipedia infoboxes could increase the quality of extracting RDF triples. The result also shows extracting Amharic DBpedia is applicable and the language can be a part of the internationalized DBpedia chapter.

Even if the study shows encouraging results, there are some remaining work needs to be done to get full Amharic DBpedia chapter. Abstract and homepage extractions must include having a full version of Amharic DBpedia chapter. Live base DBpedia extraction can be a considerable in the future work because it can get dynamic knowledge from Wikipedia and has a capability to deliver instant RDF triples.

Building Amharic knowledge bases, including Amharic DBpedia RDF store helps in order to facilitate access and querying structured data. Furthermore, the Amharic triple store can be a knowledge source for NLP tasks and web applications.

Keywords: DBpedia, I18n Extraction framework, RDF, Semantic Web, Wikipedia.

CHAPTER ONE

INTRODUCTION

1.1. Background

Semantic Web represents the next major evolution in connecting information. It allows information to be connected from one source to another source and to be seen by computers, so they can perform progressively complex tasks on our behalf (Fred Fonseca, 2011). The main principle of the Semantic Web is to advance the processes of knowledge representation and information retrieval on the Web. Semantic Web proposes changes on the markup languages utilized as a part of the development of pages to improve the interconnection between such pages and system. With the use of semantics, it is possible to increase the associations of documents to their meanings through descriptive metadata. The regular web design is for peoples make things simple, at the same time, not for computers. So the key thought of Semantic Web is to make computers read, see the relationship of all data, and make a choice like humans do. Semantic Web brings structure to the meaningful content of web pages, creating an environment where software agents' crawls storm to the page (Berners-Lee T. , 2001) can readily carry out sophisticated tasks for users. Unfortunately, the full functionality of the Semantic Web is still not completed (Al-Feel, 2013).

First of all to semantic web computers must have access to structured collections of information and sets rules that to apply automated reasoning. To get structured information Wikipedia is a good choice because it has well-structured and high quality data (Al-Feel, 2013) and it is the largest encyclopedia in online format and written collaboratively by people all around the world. Everybody can access this knowledge as well as add and edit contents. Right now Wikipedia has over 30 million articles in 286 languages and it became one of the largest encyclopedia providers of multilingual content in the world. English Wikipedia is one of the most resourceful with more than 3 million pages; due to this it is preferable to apply Semantic Web with it.

Knowledge bases are playing an undeniably vital role in improving the intelligence of web and in supporting information searching and integration. Today, most knowledge bases cover

particular areas, created by moderately little gatherings of information designers, and are extremely cost intensive to keep up with the latest change. In the meantime, Wikipedia has developed into one of the main knowledge sources for everyone, maintained by thousands of contributors and available freely. There are different kinds of project applied by semantic web to make semantic web technologies to structurally and semantically enrich the web by using Wikipedia. One of the projects to such end is DBpedia.

DBpedia is a community effort to extract structured information from Wikipedia and make it available on the Web. The DBpedia project was started in 2006 and it pulls massive knowledge by extracting structured information from Wikipedia. Recent studies (Georgi Kobilarov, 2009) show DBpedia has more than 274 million pieces of information in 15 languages. DBpedia gives a wide variety of benefits such, as it works as a central hub for Linked Data on the web, help to ask sophisticated queries against datasets derived from Wikipedia, accessing multilingual information and it can use for different kinds of Natural Language Process research and applications. Currently DBpedia is the largest hub on the Web of Linked Data (Dimitris Kontokostas, 2012). Meanwhile DBpedia attracted significant interest in research and practice.

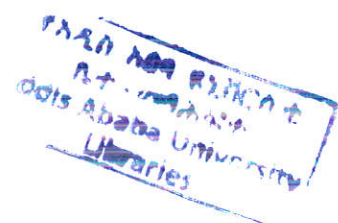
Even if Amharic is the second most spoken Semitic language next to Arabic (Lars Asker, 2008) a localized chapter of DBpedia is not available for this language. A step towards the internationalization of DBpedia, extracting Amharic DBpedia chapter can contribute a fundamental part in the deployment of the global DBpedia collections. As a new chapter it would be added to the available multilingual DBpedia chapters which can support the maintenance, extraction and enriching of DBpedia especially for being Amharic language users. Transforming Wikipedia into knowledge base is the main application of DBpedia. Similar to other language DBpedia chapters, the Amharic DBpedia can be a central interlinking hub where all Amharic linked datasets will be connected. It can also be guidelines on how Amharic linked data could be published and how Amharic characters can be handled in linked open data. Evidence presented by (Charalampos Bratsas, 2011) shows the majority of users are attracted by DBpedia projects and web sites if the information is available in their native language. Furthermore, extracting the Amharic DBpedia and providing links it to many other Wikipedia language resources in the linked data. This central interlinking knowledge base could also be used as background

knowledge in many natural language processing techniques especially in a multilingual environment.

1.2. Statement of the Problem

To extract Amharic DBpedia it needs to handle the problem of Amharic Wikipedia. It faces some problems such as Wikipedia contents not follow the Wikipedia structure and Wikipedia infoboxes has inconsistency problems. Because of this the DBpedia knowledge representation will have limitations. In addition, when to extract DBpedia it needs to use a proper serialization method to represent RDF triples.

- Does the current DBpedia extraction framework can be used to extract Amharic DBpedia Chapter?
- What characteristics of the Amharic Wikipedia hinder the effective extraction of the Amharic DBpedia?
- How can we improve Amharic Wikipedia infobox issues?
- Which RDF serialization method is applicable for Amharic characters?



- What adaptations are needed to the DBpedia extraction framework so as to extract quality RDF triples from Amharic Wikipedia?

1.3.Objective of the Study

1.3.1. General Objectives

The general objective of the study is extraction of the Amharic DBpedia from Wikipedia.

1.3.2. Specific Objectives

To achieve the general objective of the study, the following specific objectives have to be considered.

- Review the literature on information extraction specifically on DBpedia extractor and related works.
- Identify the Amharic Wikipedia infobox problems, which are basic to DBpedia extraction, and improving their quality.
- To investigate existing DBpedia extraction methods and techniques.
- Implementation of II8n DBpedia extraction framework to extract Amharic DBpedia.
- Evaluate the extraction result.
- To reach conclusions taking into account experimental result and suggest further research works.

1.4. Scope and Limitation of the Study

The scope of the study is extracting Amharic DBpedia from Wikipedia. In the process, Wikipedia infobox enrichment and Amharic DBpedia dataset extraction have been done. In addition to these, the extraction shows the Amharic DBpedia extraction methods, challenges to extract Amharic DBpedia and achievement of this research.

Out of all extracted Amharic DBpedia datasets, abstract and home page of Wikipedia extraction is not done in this study because DBpedia extraction framework didn't work for these datasets. To accomplish this it need further research to upgrade the extraction framework.

1.5.Methodology

In order to achieve the objectives, the following resources and methods will be employed:

1.5.1. Data Source

Amharic Wikipedia is a data source for this research. It can be downloaded¹ from Wikimedia and dumped to a personal computer for extraction purpose.

1.5.2. Development Tools

Extraction procedures of the research comprised a series of processing stages. In order to meet the objective of the research, dumping Amharic Wikipedia, parsing and RDF extraction accomplished. The tools we used for Amharic DBpedia extraction is: I18n DBpedia Extraction Framework it can allow extracting structured data from Wikipedia and turns them into RDF. MySQL (Structured Query Language) is used for this research to store Amharic Wikipedia in local computer and python Programming Language also used for experimentation. In detail the development tool we use for this study are:

- **I18n DBpedia Extraction Framework**

I18n extraction framework able to process input data from Wikipedia. Amharic Wikipedia, like other Wikipedia languages, has different types of features (Shown 2.5.1). After a proper configuration made, I18n DBpedia extraction framework extracts these different kinds Amharic Wikipedia features and turns them into RDF triple.

- **MediaWiki**

MediaWiki is free software, and was originally developed for Wikipedia. The software allows the creation of pages in which information is written in a markup language known as Wiki text. Every article on Wikipedia is technically realized with a MediaWiki page and it can allow to access local Wikipedia pages. For DBpedia extraction purpose, we use MediaWiki 1.24.1 to have local Amharic Wikipedia.

¹ <http://dumps.wikimedia.org/amwiki/>

- **MySQL**

MySQL is an open source relational database management system (RDBMS) based on Structured Query Language (SQL). MySQL can associate with web-based applications and online publishing and is an important component of an open source enterprise stack called XAMP. XAMP is a Web development platform that can use Linux as the operating system. All Amharic Wikipedia content is imported from an Amharic MediaWiki dump in MySQL table to use for extraction purpose.

- **PHP**

PHP (an acronym for *PHP: Hypertext Preprocessor*) is a widely-used open source general-purpose server side scripting language that is especially suited for web development and can be embedded into HTML. Apache is the most popular Web server software. To have MediaWiki install on the local server, PHP platform is needed and we configure PHP in local computer for the purpose of MediaWiki configuration.

- **Scala and Maven**

Scala is an acronym for “Scalable Language”. Its types get out of the way because the compiler can infer them. Scala is the preferred workhorse language for many mission critical server systems. Maven is built using a plugin-based architecture that allows it to make use of any application controllable through standard input. Maven can be used to build and manage projects written in Scala and other programming languages. A DBpedia extraction framework is developed by Java programming language so that to create an appropriate run time environment, the framework needs a proper configuration of scala and maven.

- **Python Programming Language**

Python is a programming language that is freely available and that makes solving a computer problem almost as easy as writing out one's thoughts about the solution. Python is an excellent tool for scanning and manipulating textual data. We use python for Amharic infobox enrichment and experimentation purpose of extracting Amharic RDF triples.

1.6. Evaluation

Evaluation is a vital step to test the performance of the extraction. The outcome of the study was evaluated with appropriate evaluation methods to check Amharic DBpedia extraction. For evaluation purposes, we follow two methods to check the extracted result. The first method is check the number of extracted triples features with Wikipedia source and the second method check the extracted RDF triples whether they are really exist in Amharic Wikipedia.

1.7. Significance of the Study

Like any other DBpedia language chapters, Amharic DBpedia extraction has a wide variety of significances. These are:

- Using language-specific resource naming approach, an inter-language link (ILL) can be utilized to connect resources across different DBpedia language editions and create a multilingual semantic space.
- Extracted Amharic DBpedia datasets can be a knowledge source NLP research and applications. For example:
 - *Named Entity Recognition (NER)*
 - *Multilingual entity recognition task of real world entities.*
 - *Voice Reorganization*
 - *Gender Identification*
- Amharic DBpedia can be a central interlinking hub with other languages for Web of Data.
- DBpedia allow user to search information by sophisticated query.
- Contribute to web technology to be multilingual helps content creators, language tools developers and it can address the new opportunities and challenges of the Web.
- Moreover, the result of the study will play a role in academics for future study in the area of Amharic DBpedia extraction, NLP and Semantic Web.

Hopefully this study will make Amharic Wikipedia to be used in some new useful ways. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself.

1.8. Organization of the Thesis

This study is organized in six chapters. Chapter One introduces Semantic web and DBpedia. The chapter also presents a statement of the problem, objective of the research, the scope of the study and general statement about the methodology of the research. Chapter Two reviews Semantic Web in general and DBpedia in particular. Detailed discussions on DBpedia works on foreign languages are discussed in Chapter Three. Chapter Four puts the methodology to adopt for solving the research theme. Here, explanation of research design provided for Amharic DBpedia extraction as well as flow of process and constraints is presented. Details issues surrounding on the implementation of the proposed model, the experiment conducted, and the results observed are reported in Chapter Five. Finally, the conclusions and recommendations made based on the findings of the study are presented in Chapter Six.

CHAPTER TWO

LITERATURE REVIEW

This chapter gives a brief description of the literature review by presenting what a typical Semantics Web is and explaining how its components function in the system also describes the concepts in relate to semantic web and DBpedia.

2.1. Semantic Web

The term “Semantics” comes from the Greek word Semantiké which is the art of signification. Most literatures define semantics as the study of meaning, or as the science that studies the meaning of words (Fred Fonseca, 2011).

The Semantic Web (SW) is not a separate web but an extension of the current one (Berners-Lee T. , 2001), in which information is given with well-defined meanings, to enable computers and people to work in cooperation. Semantic web is a collaborative movement led by the international standards of the World Wide Web Consortium (W3C). It is the main international standards organization for the World Wide Web. The standard upholds developing protocols and guidelines that ensure long term growth of the World Wide Web.

The semantic web intends to improve the existing web with a layer of machine-interpretable metadata so that a computer program can understand what a web page is about (Majdi Beseiso, 2010). Semantic web is a network of knowledge compared to what we have today that can be defined as a network of information. Information networks are created when a number of information sources are properly interconnected (Mazzocchi, 2001).

The advantages of the semantic web include the ability to share more information than individuals ever have been before. An alternate point of interest would be the capacity to search and extract information much simpler and quicker. More innovative advances can spring from the semantic web also. Everything is easier and much more accessible for everyone.

2.2. Semantic Web Technologies

It is a W3C technology that forms the basis of the semantic web including Extensible Markup Languages, Resource Description Framework and Web Ontology Language.

2.2.1. Extensible Markup Language (XML)

XML is a general purpose data format that is intended to facilitate the distribution of information between wide varieties of computer in a computer network such as an Internet. XML is a language for documents identifying structured data in a quite simple way. Structured data includes content (text, pictures, etc.). XML works through tags. A tag surrounded by left and right chevron and the inside of the start and end tags the data placed (E.g. <title> የኢ.ፔ.ዲ.ዎ ሕክምና </title>). Start tag marks the start of some data and end tag marks end of some data. The table shows an example of XML tags.

Starts marks beginning	The data	End tag marks end
<Title>	የኢ.ፔ.ዲ.ዎ ሕክምና	</Title>

Table 2.1 XML Tag Example

2.2.2. Resource Description Framework (RDF)

Semantic information used for computers to understand what we put on the web. RDF helps to add semantic information on the web pages. RDF is a language for representing resources on the World Wide Web using RDF triple statements and is the basis for the semantic web. Previously RDF was developed to provide a formal means of describing any resources, both online and offline. Computer applications make use of structured information in the semantic web and spread in a distributed and decentralized way throughout the web. RDF is an abstract model, a way to break down knowledge into discrete pieces, while it is most popularly known for its RDF/XML syntax.

W3C give standard for web technologies based on considering the aspect of accessibility, privacy, security, and internationalization². RDF is the W3C standard for encoding knowledge

² <http://www.w3.org/standards/about.html>

because it describes any concept, relationship or things. It is a general method of decomposing any type of knowledge into small pieces, with some rules about the semantics, or meaning, of those pieces. In addition to these, RDF is easy to write and flexible to use. It makes it easy to modeling knowledge³ and to holding as a net distributed knowledge sources.

2.2.3. RDF Versus XML

XML is syntax and data model is a tree. RDF data model is based on a graph that uses URIs and has much diverse syntax. In any case, both XML and RDF can be utilized to represent structured data on the web and move information around between applications, but RDF is more powerful than XML because it contains meaning and make computer to understand the content.

2.2.4. The Web Ontology Language (OWL)

OWL is a language for representing ontology of web resources in the semantic web. Knowledge Representation OWL language proposed by the W3C as a standard to represent ontology in a prospective semantic web.

There are three types of entities in OWL ontology.

- **Individuals:** the objects of the knowledge domain.
- **Properties:** that can be used to link individuals in binary relations and
- **Class:** is a set of individuals with common characteristics.

2.3. Architecture of Semantic Web

The architecture of semantic web is shown in Figure 2.1, the first layer, a URI is a standardized form that allows to uniquely identifying resources and Unicode is a standard of encoding international character sets which allows all languages to be used on the web using one standardized form. XML layer with XML namespace and XML schema definitions makes sure that there is a common syntax used in the semantic web. An XML document contains elements that can be nested and may have attributes and content. XML namespaces allow specifying different markup vocabularies in one XML document. XML schema serves for expressing a schema of a particular set of XML documents (Obitko, 2007).

³ <https://help.github.com/>

A core data representation format for semantic web is called RDF. Previously RDF was used for representing metadata about WWW resources, such as the title, author, and modification date of a web page, but it can also be used for storing any other data. RDF is based on triples subject-predicate-object that forms a graph of data. All data in the semantic web use RDF as the primary representation language (Michael Kifer, Jos de Bruijn, Harold Boley, Dieter Fensel, 2005).

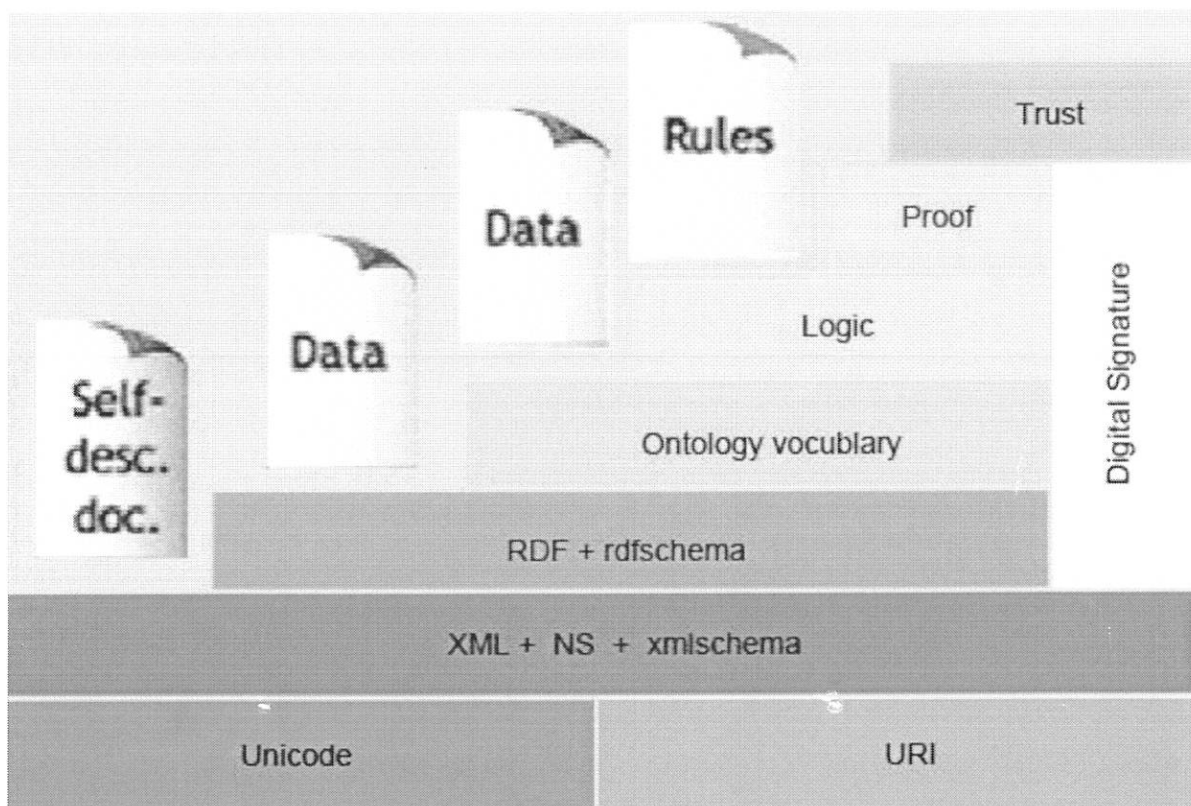


Figure 2.1 Layered architecture of semantic web

As shown in Figure 2.1, Unicode is an international encoding standard for different languages and scripts where each letter, digit, or symbol is assigned a unique numeric value that applies across different platforms and programs (Gulzar, 2007) and it can as an identification of resources across the web established by a Uniform Resource Identifier (Ian Jacobs, Norman Walsh, 2006).

All the above technologies are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain and any of these technologies can be used by any application, but only applications which support RDF are truly semantic web application (Krupansky, 2009). The main goal of the semantic web is to improve the processes of knowledge representation (KR) and information retrieval (IR) on the web.

2.4. Linked Data

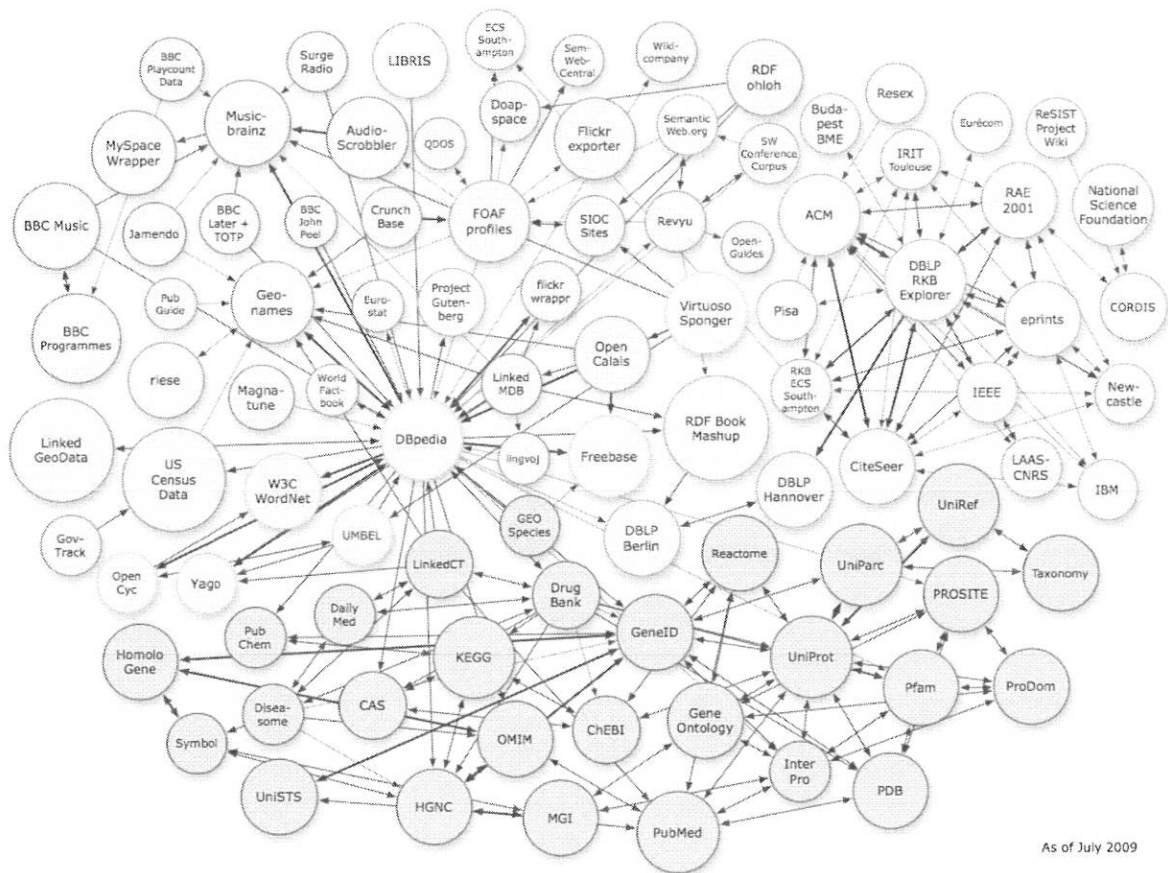
A controversial dream of many is to enable the web itself to be a global united database. Linked data is a best practice for publishing and connecting structured data on the web. Linked data uses a web to create links of data from different resources and it gives an approach to proficiently share and reuse these resources. Today, more than 30 organizations publish their libraries of data into semantic web formats and make them query able from the web itself (Pollock, 2009). Linked data is considers as a heard of semantic web⁴.

Open linked data, unlike a regular database, the semantic web data and data models can be directly and precisely linked together over the web itself. Even if linked data has a precise like between its RDFs, There are some problems for the application. One of the common problems is packaging. Packaging is a problem of representing data in common format like HTML and XML. The other problem is how to find out how to link the data in the web. To solve these problems the representation of data should follow property-value mechanism (Sporny, 2012).

Linked data is a set of principles and W3C standard technologies that can be used for the distributed access to data over web protocols. The basic principle⁵ of linked data is depending on web architecture (terminology of resource and resource identifier) and RDF data model (RDF). It works by making typed links between pieces of RDF data. The RDF format makes it machine readable. (Berners-Lee T. , 2009). By publishing and interlinking various data sources on the web, the dinking data community has created a crystallization point for the Web of Data and a challenging test bed for Linked Data technologies. Figure 2.2 shows sample open linked data cloud.

⁴ <http://www.w3.org/standards/semanticweb/data>

⁵ <http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/#principles>



6

Figure 2.2 Excerpt of Sample Open Linked Data cloud

Publishing data on the web according to the linked data principles, data providers most cases government, add their data to a global data space, which allows data to be discovered and used by various applications. Open Data information not only get from governmental institutions, but also includes those from other relevant stakeholder groups such as business/industry, NGOs, and science /education institutes. Some of the best-known institutions are currently undertaking Open Data activities by providing their organizational and published information are World Bank, United Nations, The New York Times, The Guardian and the Open Knowledge Foundation

⁶ http://www4.wiwiw.fu-berlin.de/bizer/pub/lod-datasets_2009-07-14_colored.png

Network (OKFN) (Florian Bauer, Martin Kaltenböck, 2009). These organizations are contributing a high volume of information for linked data.

2.5. Wikipedia

Wikipedia is online encyclopedias that provide open source information to everyone. Wikipedia is built by through collaborations and publishing different kind information from various contributors worldwide. Now a day, Wikipedia has grown and become one of the largest online sources of encyclopedia with millions of articles. Wikipedia covers more than 280 languages (Rada Mihalcea, Andras Csomai, 2007).

Wikipedia represents its data in a simple format called wiki text which is simple to edit and modify. It can allow editing Wikipedia pages except sensitive pages that are protected. Although Wikipedia's principle is that anyone can edit the content and it therefore aims to have as many of its pages as possible open for public editing except some pages that have a restriction in protection policy⁷ of Wikipedia. Protections applied by only Wikipedia's administrators and it can be unlimited or can expire after a specific period of time. According to Wikipedia's protection policy, there are different type of Protection polices Wikipedia have. These are: Full protection, Semi protections, Creation protections, Move protections, Upload protections and Pending-changes protections. All of these are available for Wikipedia administrator. All Wikipedia users can locate and read articles, access photos and other information, and follow hyperlinks to resources related to the topic by using keyword searches

According to alexa⁸, the well-known website data analyzer, Wikipedia is the 6th most visited site in 2013. Wikipedia powered by Google search engine and this make the site appears 50% of search engine traffic in Google search. Wikipedia content used for different kind of purpose such as academic references, scientific research, court case and governmental publication reference⁹ and it can also use as journalism resources.

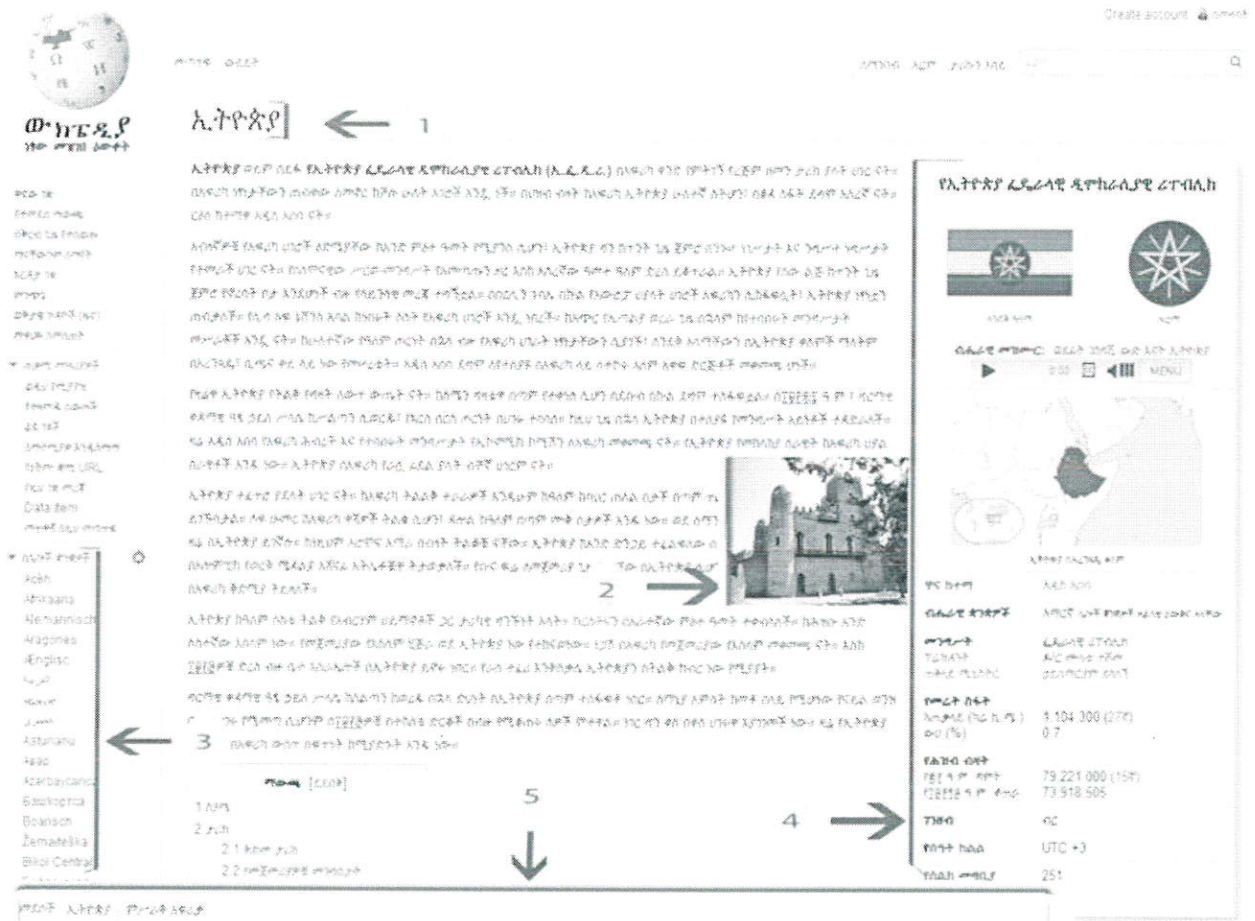
⁷ http://www.wikipedia.org/wiki/Wikipedia:Protection_policy

⁸ <http://www.alexa.com>

⁹ <http://www.parl.gc.ca>

2.5.1. Structure of Wikipedia

MediaWiki provides essential markup for the creation of Wikipedia contents. The design of Wikipedia articles must also follow conventions in order to present information in a coherent way. Figure 2.3 shows an excerpt of standard layout of Wikipedia pages.



Description: 1.Title
2.Image
3.Interlanguage links
4.Infobox
5.Categories

Figure 2.3 Sample Amharic Wikipedia structure

Wikipedia has different kinds of structure (features). Each feature has its own kind of information representation. The most common features of Wikipedia in all language editions are:

➤ **Infobox**

An infobox is a fixed-format table mostly to be placed at the top right-hand corner of articles to present a summary of articles share and sometimes to improve navigation to other interrelated articles. Many infoboxes also produce structured metadata which is used by DBpedia and others. The generalized infobox feature grew out of the original infoboxes (taxonomy infoboxes) that editors developed to visually express the scientific classification of organisms. Wikipedia infoboxes contain very specific information about things and are thus a very valuable source of structured information that can be used to ask expressive queries against Wikipedia.

➤ **Inter-lingual Links**

Wikipedia articles on the same topic in different languages are connected via inter-lingua links. These links serve as an excellent resource for obtaining lexical translations, or building multilingual dictionaries and semantic networks.

➤ **Categories**

Categories is one of the navigational links to all Wikipedia pages in a hierarchy of categories which help readers to browse and quickly find sets of pages on topics that are defined by those characteristics.

➤ **Abstract**

The abstract of a Wikipedia article is the first few paragraphs of that article.

➤ **Geo-coordinate**

The geo-coordinate has geographical information of latitude/longitude. Places mentioned in the infobox mostly have geo-coordinate information.

➤ **Wikipedia Page Links**

Wikipedia articles together are essential because of this Wikipedia page links allow users to go to another page related to the article they are reading and this links help readers to get more information.

➤ **External link**

External links are helping the reader to go to another page to get related information. Unlike Wikipedia page links, these links in the Wikipedia take the reader to other page that is not part of Wikipedia.

➤ **Revision ID**

Each historical version of a page has revision ID¹⁰, which help users to find from the history of the page. The page holds history of which page were made to any editable Wikipedia page.

➤ **Home Page**

A home page that is the first page that visitor navigating to a website of Wikipedia page. This page is the landing page of Wikipedia content based on their titles.

2.5.2. Wikipedia's Community

The Wikipedia community is the community of contributors and these individual contributors are known as "Wikipedians". The majority of Wikipedians are volunteers. With the increased maturity and visibility of Wikipedia other categories of Wikipedians emerged.

The number of contributors in November 2011, there were approximately 31.7 million registered user accounts across all language editions, of which around 270,000 accounts were active on a monthly basis (Alike, 2014). As of October 2013, the community of volunteers declined and is continuing to drop. Some 31,000 editors are active in the project, but the decreasing contributor number didn't affect the increasing of Wikipedia's articles (Suh, 2011).

¹⁰ http://en.wikipedia.org/wiki/Wikipedia:Revision_id

2.5.3. Language Editions of Wikipedia

Based on ISO 639¹¹, the URL of the Wikipedia for a given language is xx.wikipedia.org, where xx is the 2-letter language code. Currently there are 287 language editions (language versions) of Wikipedia. Out of these numbers, nine of them have over one million articles each, six languages, have over 700,000 articles, 36 languages have over 100,000 articles, and 74 languages have over 10,000 articles¹². The following tables (2.2 and 2.3) show the rank of global and local languages in Wikipedia respectively.

Rank	Language	Article
1	English	4,462,471
2	Dutch	1,763,752
3	German	1,692,696
4	Swedish	1,612,310
5	French	1,481,635

Table 2.2 Top ranked global language edition in Wikipedia

Each language edition functions more or less independently, some efforts are made to supervise them all. Meta-Wiki provides important statistics on all language editions of Wikipedia.

Rank	Language	Article
1	Amharic	15,741
2	Oromo	285
3	Tigrinya	270
4	Afar	6

Table 2.3 Top ranked Ethiopian language edition in Wikipedia

2.5.4. Accuracy of Contents in Wikipedia

Wikipedia has different comments on reliability. Some critics agree that Wikipedia may be reliable, but not clear that the reliability of the article (Boyd, 2005). The research done by Nature

¹¹ http://www.iso.org/iso/language_codes

¹² http://en.wikipedia.org/wiki/Wikipedia#Language_editions (*checked on April, 20014*)

science journal (Giles, 2006) shows that the difference in accuracy between Encyclopedia Britannica and Wikipedia was not particularly great. The research takes to review forty-two articles from Wikipedia and Encyclopedia Britannica; it concluded the average science entry in Wikipedia contained around four inaccuracies; Britannica about three.

Open nature and lack of proper resources, Wikipedia has critics argue in the unreliability of its information (Schiff, 2006). As a consequence of the open structure, Wikipedia not give any guarantee of validity of its resources because there is no one is responsible for the information appear on the website (Giles, 2006).

2.5.5. Software Operations and Support of Wikipedia

Article write in Wikipedia use free and open source wiki software called MediaWiki. On Wikipedia, MediaWiki has more than 1000 automated and semi-automated bots and other tools have been developed to assist in editing. The platform of Wikipedia written in PHP and it built upon the MySQL database system (Bergman, 2008).

2.5.6. Scientific Use of Wikipedia

Wikipedia has seen widespread use as a corpus for linguistic research, in the area of computational linguistics, information extraction and natural language processing.

Wikipedia can use as a resource for automatic keyword extraction and word sense disambiguation, and the research result (Rada Mihalcea, Andras Csomai, 2007) shows how this online encyclopedia can be used to achieve state-of-the-art results on both these tasks.

Wikipedia in higher education use in courses ranging from humanities to science to business and with Wikipedia, students can take part in a collaborative process of creating and revising content in a global context, moving the opportunities for learning beyond the walls of the classroom or the university (ELI, 2007).

2.6. DBpedia

DBpedia is a result of community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows users to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data (Christopher, 2014)

The DBpedia project targets are extracted information based on semi-structured data present in Wikipedia articles, interlinking it with other knowledge bases, and publishing this information as RDF freely on the Web (Stadler, 2010). Over the past two years the DBpedia knowledge base has turned into a crystallization point for the emerging Web of Data (Mohamed Morsey, 2012).

DBpedia project has three major parts. These are the website, the datasets and the Extraction Framework. The DBpedia datasets are getting from processes of Wikipedia. The extraction work is performed by the DBpedia Extraction Framework, it can be freely downloaded (Charalampos Bratsas, 2011) from DBpedia webpage.

2.7. DBpedia Extraction

The extraction process in DBpedia can be made via two ways, dump based extraction and live extraction (Mohamed Morsey, 2012).

2.7.1. Dump Based Extraction

The Dump Based Extraction depends on the Wikipedia dumps published in SQL for all editions and updated monthly and collected in DBpedia via the page collection of Wikipedia database.

2.7.2. Live Extraction

DBpedia extraction extracts a large volume of knowledge from Wikipedia. This large information extraction procedure has been a disadvantage. It needs the involvement of manual effort to extract knowledge on the fly. To handle this issue DBpedia live extraction is a good option. It depends on the live update of DBpedia according to any changes or modifications made in Wikipedia due to the privilege given to the DBpedia project to access the Wikipedia

Open Archives Initiative Protocol for Metadata Harvesting (OAI – PMH) live feed that instantly reports all Wikipedia changes (Jens Lehmann, 2009).

2.7.3. DBpedia Information Extraction Framework (DIEF)

DBpedia information extraction framework (DIEF) is the first version used for extracting different kinds of Wikipedia features from English version of Wikipedia. The framework is designed to extract only English language (Dimitris Kontokostas, 2012).

2.7.4. I18n DBpedia Information Extraction Framework (I18n DIEF)

I18n DIEF able to extract RDF data from different structure of Wikipedia data. The framework has different kind's extractors to extract articles from Wikipedia (Mohamed Morsey, 2012). The framework extracts different features of Wikipedia by using different kinds of extractors. These I18n extraction framework extractors are:

- **Abstracts.** The main body of the Wikipedia part, which is abstract, is extracted from Wikipedia.
- **Labels.** All Wikipedia articles have a title, which is used as an `rdfs:label` for the corresponding DBpedia resource.
- **External links.** Articles contain references to external Web resources and it can represent by using the DBpedia property `dbpedia:reference`.
- **Redirects.** In order to identify synonymous terms, Wikipedia articles redirect to other articles. Extracting redirects help to resolve references between DBpedia resources.
- **Page Links.** By using the `dbpedia:wikilink` property, Page Links extract all links between Wikipedia articles and a page represent them.

- **Homepages.** This extractor obtains links to the homepages of entities by looking for the terms homepage or website within article links. The representation is Friend of Friend homepage (foaf:homepage).
- **Infobox.** The extractor extracts infobox property and values from Wikipedia content.
- **Revision ID.** This extractor extracts the history page edited by users.
- **Categories.** Each Wikipedia articles are arranged in categories. Categories become skos:concepts; category relations are represented by using skos:broader.
- **Geo-coordinates.** The geo-extractor expresses coordinates use the BasicGeo and the GeoRSS. Besides coordinates extracted from infoboxes and the DBpedia data set contains additional geo-coordinates for geographic locations. Geo-coordinates are using (Sahnwaldt, 2013) the W3C Basic Geo Vocabulary.

2.8. Amharic Writing System

Amharic is a Semitic language used as the working language of the Federal Government of Ethiopia and is spoken and written as a first or second language in many parts of the country¹³. Amharic has more than 25 million speakers and it is the second most spoken Semitic language in the World Arabic (Lars Asker, 2008).

Writing systems can be divided into two main types: those that represent consonants and vowels (alphabets), and those which represent syllables (syllabaries), though some do both. The Amharic writing system is syllables. Amharic is written with a version of the Ge'ez script known as Fidel (ፊደል) which grew out of the Geez. Geez, which is still the classical and church language of Ethiopia, took and modified Sabeian scripts to become a developed written language and Amharic took all the 26 characters of its predecessor (Yimam, 1991).

¹³ <http://www.omniglot.com/writing/amharic.htm>

Modern Amharic has inherited its system of writing by way of the language of Axum, Geez, which is still the classical and church language of Ethiopia. The roots of Amharic orthography; like those of the language itself, are Semitic¹⁴. The Sabaean alphabet is said to have had twenty-nine symbols. All twenty-nine had been in use in Ethiopia. As (Afework, 2011) quote, when Ge'ez became the spoken and written language, it took over only twenty-four of the twenty-nine symbols. In contrast to the other Semitic languages like Hebrew and Arabic, both Geez and Amharic languages are written and read from left to right.

¹⁴ <http://www.ethiocloud.com>

CHAPTER THREE

RELATED WORKS

The emergence of the World Wide Web has provided new opportunities to access information easily from different sources and enables researchers to make new research to get better technologies. Previous year, most of web information was in English language. Through time, the number of non-English web resources increased and now non-English web pages are 60% and 40% English web pages are available on the Internet (Welde, 2011). This research shows that people are being an urgent to use their own language by developing tools and Internet resources that can allow them to better access of information on the Internet.

In the past few years, many researchers have tried to address DBpedia extraction for different languages. To show state of the art and effort of DBpedia extraction, relate works have been reviewed in this chapter.

3.1. Internationalization of DBpedia

An early version of DBpedia was found only in English language, but now DBpedia covers 15 localized DBpedia chapters that include Catalan, Spanish, German, Croatian, Greek, French, Irish, Hungarian, Italian, Dutch, Polish, Portuguese, Russian, Slovene, Turkish languages. The recent figure shows the DBpedia dataset describes over 3.64 million things across 97 languages, providing 1 billion pieces of information in the form of RDF triples with an open license in various domains (Al-Feel, 2013). After the use of I18 framework, the Greek DBpedia is the first successful application towards Linked Data. Including Greek DBpedia (Dimitris Kontokostas, 2012), other DBpedia extractions, such as: Arabic (Al-Feel, 2013), Korea (Eun-kyung Kim, 2010) and Germany DBpedia extraction papers also reviewed in this chapter.

3.1.1. Greek DBpedia (Dimitris Kontokostas, 2012)

The Greek DBpedia is the first step towards Linked Data internationalization. The research uses I18n extraction framework to extract Greek DBpedia chapter and introduce two extractors that can be used as a plugin for the I18n extraction framework. The research is a first successful

attempt of using I18n extraction framework and they can significantly contribute the IRI adoptions. Language specific extractors such as home page and abstract extractors work properly for this language. Greek language use non-Latin characters and the research show a way to handle by IRIs. The research outcome shows the main DBpedia edition can also significantly contribute towards the IRI adoption of Greek language. In all the research describes the deployment of the Greek DBpedia and the contribution to the DBpedia information extraction framework with regard to internationalization (I18n) and multilingual support. Even if the research shows many achievements, the final result of template extractor parameter not present and the study try to add 'Template Parameter Extractor' but it is not fully functional and they leave it as a future work.

Amharic DBpedia extraction uses the same framework (I18n extraction framework) for DBpedia extraction. Like Greek language, Amharic also use non-Latin characters and the research show the way to how Amharic DBpedia extraction application in I18n extraction framework to extract Amharic DBpedia from Wikipedia. The characters used for Amharic language are present correctly by using Turtle serialization.

3.1.2. Korean DBpedia (Kim, 2010)

Korean DBpedia and the English version of DBpedia use the first version of DBpedia extraction framework (DIEF) to extract Korean DBpedia. The first DBpedia extractor it limited by English language. Because of this, they use their own plug-in extractors for language specific features. In the result they can extract all Korean Wikipedia features with effective IRI adaptation. In addition to this they use. However, the previous framework does not have sufficient support for non-English languages, especially for the non-Latin alphabet based languages. Korean DBpedia extraction gets some improvement in the extraction of non-Latin characters and extended the framework with pluggable internationalization components in order to facilitate the extraction of localized information. By the improvement of extraction framework, they can increase the amount of extracting triples. The previous Korean infobox was manually created by Wikipedia authors that create or edit an article with low quality. As a result, many articles have no infoboxes and other articles contain infoboxes which are not complete with each other. To tackle this problem they use python programming language to improve the Korean infobox extraction

module and they complete extracted Korean DBpedia by translating English DBpedia. The limitation of the research is the dictionary they use Korean – English dictionary to enrich Korean infobox but their attempt is limited by the number of word that find in the dictionary.

Non-Latin alphabet base languages have a challenge to use the extraction framework to extract DBpedia. Even if we are not use DIEF like Korean language, we apply an IRI configuration to handle Unicode character issue. Before they extract Korean DBpedia, they apply some improvement on Wikipedia to get high quality and quantity of RDF triples. For future work of improving Amharic DBpedia RDF triples, the Korean way of advancing Wikipedia is a good option; especially for extract Amharic infoboxes. To apply this, first it can uses a customized template for Amharic infoboxes for Amharic DBpedia extraction and it will be a considerable option to use English-Amharic dictionary to align all English infobox to Amharic infoboxes to enrich Amharic infobox.

3.1.3. Germany DBpedia (Sebastian Hellmann, 2012)

The German Wikipedia is currently the second largest Wikipedia with about 1.3 million articles. By the use of I18n DBpedia extraction framework, Germany DBpedia is extracted and the language is the third largest localized DBpedia with a total of 73 million RDF triples. Because of the vast about of DBpedia resource, the main focus of this research paper is the application of German DBpedia. The main application they show is NLP Interchange Format. This format can be used to represent this linking transparently for applications and using DBpedia ontologies to represent linguistic annotations because the extracted triples have excellent quality of RDF triple representation as well as the number quite large. In their research they also try to use geographical datasets of DBpedia to create a knowledge base called LinkedGeoData but it is not fully complete. In the future research Amharic DBpedia application, it can use Amharic DBpedia for NLP recourse (like Named Entity Recognition and Gender Identification).

3.1.4. Arabic DBpedia (Al-Feel, 2013)

A step towards the internationalization of DBpedia, this research try to describe its practices and the efforts performed as part of the work in the deployment of the Arabic DBpedia as a new chapter. Arabic DBpedia extraction uses a tool (I18n DIEF) to extract Arabic Wikipedia. Unlike

the above reviewed papers, this research is a prototype of Arabic DBpedia. The main challenge they face is unavailability of Arabic Wikipedia infoboxes for many Arabic Wikipedia articles and to solve this problem they prepare Arabic infobox template from scratch. To extract information written in the Arabic language, they use different extractor like label, abstract, Geo and wiki pages and they can extract all Wikipedia attributes inside infoboxes. As a limitation, some Arabic Wikipedia features are not extracted and they suggest in their future work to include the remaining un-extracted features.

The main objective of reviewing these works is to gain background knowledge about DBpedia extraction framework. Amharic DBpedia extraction has some commonality with Arabic DBpedia in extraction and development. The challenge to extract Amharic DBpedia is Wikipedia contents are not well organized and it affects the number of extracted triples and the knowledge representation of RDFs. To tackle this problem, we apply infobox enrichment (improvement) to improve the extracted triples.

As it mentions in chapter one, this study shows a way to extract Amharic DBpedia chapter and this will be the first step of Amharic DBpedia to be a part of internationalize DBpedia. This will have a contribution to existing linked data internationalization effort and all of the reviewed works have some leading way for this research.

CHAPTER FOUR

RESEARCH DESIGN

This chapter presents the research design of a system that allows DBpedia extraction. The proposed design focuses on the base level of the system that gives an understanding of what is to be built and how it is expected to build.

Wikipedia articles comprise for the most part of free content and various types of features such as infobox, category, inter-lingual link and geo-coordinates. The proposed design considers each of the Wikipedia article features. The main focus of the design is on how to preprocess the Amharic Wikipedia, extractor configuration, and represent the extracted RDF triples.

4.1 Design of Amharic DBpedia Extraction

Design model of the system illustrates major components and schemes of communication and collaboration working on a given theoretical solution. It also shows responsibility of the components involved and the constraints under which the model remains functional¹⁵.

In this thesis, the proposed system can be divided into three main parts; Wikipedia input, extraction process and RDF output. The input part has a Wikipedia collection into the system and from the collection, Amharic Wikipedia is imported to local server. The second part is the most critical part of the system to process Amharic Wikipedia content by applying filters, parsers, and different kinds of extractors on it. The final part is the output. In this stage extracted Amharic RDFs are dumped into local computer. The main processes of the Amharic DBpedia extraction are shown in Figure 5.1.

¹⁵ <http://www.systems-thinking.org/modsim/modsim.htm>

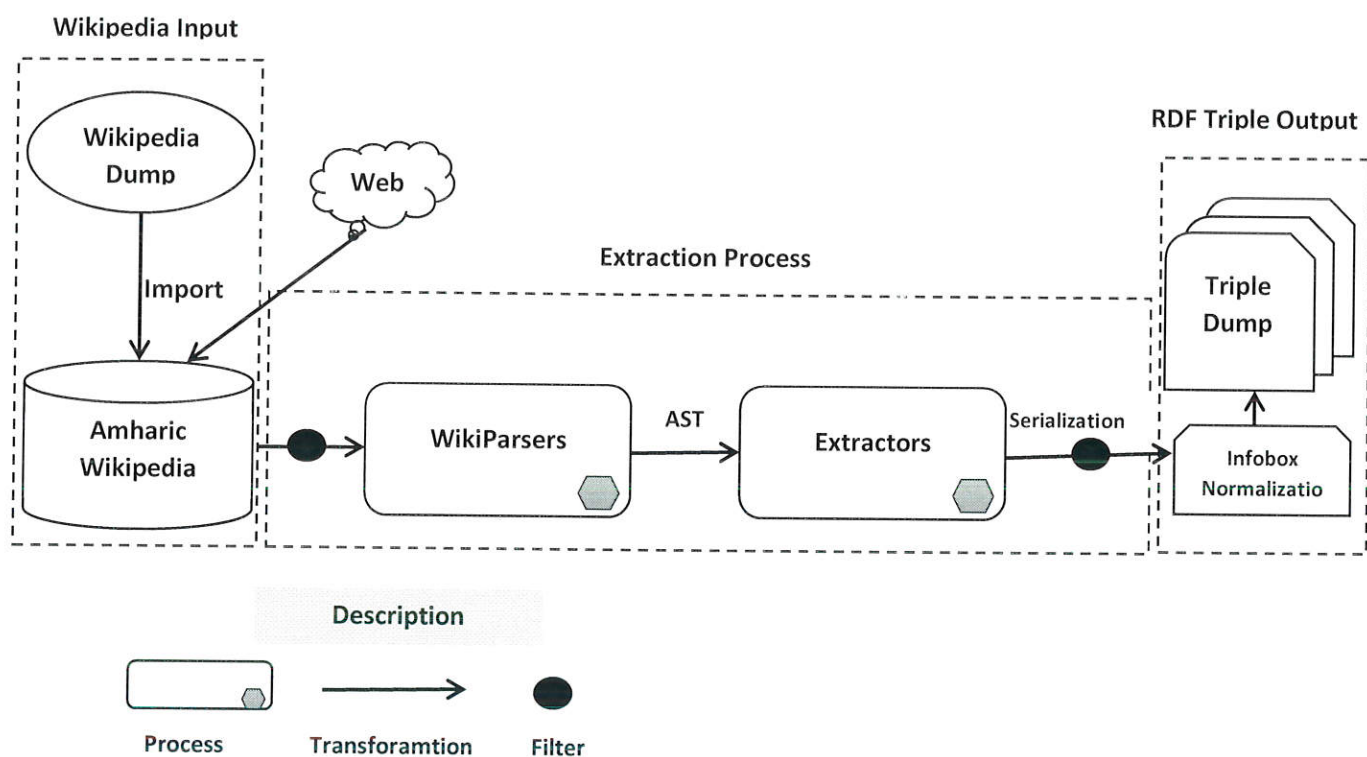


Figure 4.1 Amharic DBpedia extraction design

4.1.1. Wikipedia Dump

Wikipedia offers free copies of accessible material to interested users. The materials in Wikipedia are available in reading and downloading format. For this research, we use a downloaded format from Amharic Wikipedia information collection website¹⁶. Before extraction process start, it needs to be dumped Amharic Wikipedia collation into local computer.

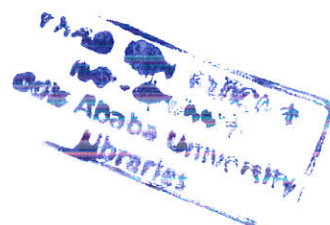
4.1.2. Local Amharic Wikipedia

Amharic Wikipedia contents are imported from Wikipedia collection and stored in local server. In addition, the latest version of Amharic Wikipedia¹⁷ and Wikidata¹⁸ dump downloaded in the local computer (database). The database used for mirroring and offline use of Amharic Wikipedia. These Wikipedia texts and metadata of all pages places as XML files and database table files include:

¹⁶ <http://dumps.wikimedia.org/>

¹⁷ <http://dumps.wikimedia.org/amwiki/>

¹⁸ http://www.wikidata.org/wiki/Wikidata:Database_download



- Lists of pages with links outside of the project (external links, language links tables)
- Page-to-page link lists (page links, category links, image links, template links tables)
- Media metadata (image tables)
- Information about each page (page, page_props, page_restrictions tables)
- Titles of all pages in the main namespace, i.e. all articles.
- List of all pages that are redirects and their targets (redirect table)
- Log data, including blocks, protection, deletion, uploads (logging table)
- Miscellaneous (interwiki, site_stats, user_groups tables)

4.1.3. Amharic Wikipedia Page Structure

Namespace is a collection of pages which have content for similar purpose¹⁹. Each of the page names reserved and recognized by MediaWiki software. Wikipedia URL page title is composed of a namespace and a page name, which are usually separated by the first colon. MediaWiki defines a set of default namespaces and each name must be declared as a namespace. (E.g. Amharic Wikipedia²⁰ has a word ‘በር’; the meaning is similar to the English Wikipedia²¹ version of ‘Portal’. Therefore a page named በር:ሳይንስ will be in the ‘በር’ namespace, whereas a page named ሳይንስ will be located in the main namespace.) In general the MediaWiki software allows every page to have subpages, where the subpage name is separated from its parent page by a slash Figure 5.2.

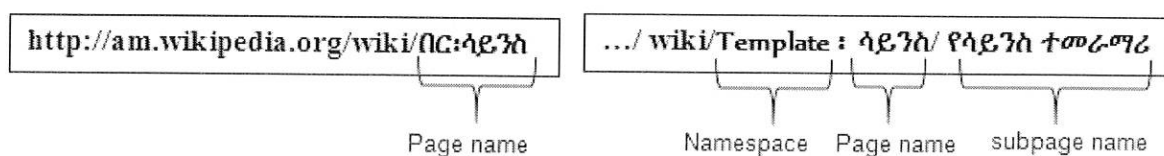


Figure 4.2 Amharic Wikipedia URL page structure

¹⁹ <http://www.mediawiki.org/wiki/Manual:Namespace>

²⁰ <http://am.wikipedia.org/wiki/በር>

²¹ <http://en.wikipedia.org/wiki/Portal:Science>

4.1.4. Amharic Wikipedia Infobox

In general almost all infoboxes including Amharic are a table place at the top right-hand of Wikipedia page (Arabic and Hebrew infoboxes are placed on the left). Infobox templates contain important facts and statistics of a type which are common to related articles. They quickly summarize important points in an easy-to-read format. Infobox improves the appearance of an article on Wikipedia.

Amharic infoboxes contain facts about the described subject and displayed as attribute-value pairs and each of them created based on Wikipedia templates. The main purpose of infobox is to summarize key facts that appear in the article. The more information it contains, the more effectively it serves that purpose, allowing readers to identify key facts when they look at it. Figure 5.3 shows an infobox template of the existing Amharic Wikipedia version.

የኢትዮጵያ ስልጣን

የዕለድ ክፍል ስልጣን የዕለድ ክፍል ስልጣን	ስልጣን
---	-------------

1. የኢትዮጵያ ስልጣን: የኢትዮጵያ ስልጣን (New York: Palgrave, 2000), p. 60

2. የኢትዮጵያ ስልጣን: የኢትዮጵያ ስልጣን (Oxford: Clarendon Press, 1972), p. 72

የዕለድ ክፍል

የዕለድ ክፍል

የዕለድ ክፍል

የዕለድ ክፍል

Figure 4.3 Existing Amharic Infobox template

4.1.5. Page Filter

The main purpose of the filters is supporting the extraction process to get proper and language specific RDF triples. Each filter has its own specific task and applied by extraction framework automatically.

- **Preprocess Filter:** Wikipedia is developed by open source software called MediaWiki. To represent Wikipedia content, this software use different kinds of markups (Shown in Figure 4.4), but these markups are not useful for DBpedia extraction. On preprocessing stage, it needs to remove these MediaWiki²² markups that are not useful for Amharic RDF representation.

[illegible]

Figure 4.4 Example of Amharic Wikipedia source

- **Namespace Filter:** Wikipedia collection has more than 280 language and all languages have its own unique namespace. Namespace filter enables to select language specific namespace for Amharic language. (E.g. 'am' is a namespace for Amharic and URL is <http://am.dbpedia.org>).

²² http://en.wikipedia.org/wiki/Help:Wiki_markup

- **IRI Filter:** there are two types of resource identification in the web, URI and IRI. For non-Latin character languages like Amharic, it is needed to use IRI²³ to get human readable resource identifier. IRI filter validate and apply IRI to resources and make the final output of the resource is readable.

4.1.6. Parsers

Each of Amharic Wikipedia pages parse by wiki parser. There are different parsers available in DBpedia extraction framework. For Amharic DBpedia extraction, we use:

- **Geo Parser:** allow to select and identify geographical coordinates from Wikipedia infoboxes.
- **DateTime Parser:** enable to identifying all date and time format.
- **Number Parser:** help to select numbers only from other Wikipedia content.
- **String Parser:** enable to identify strings with other Wikipedia content.
- **Unit Value Parser:** determining measurement data types, converting values between different units.

4.1.7. Extractors

The extractor is a core of the system to manage triple extraction by processing Amharic Wikipedia contents and delivers their output to the destination. There are different kinds Wikipedia feature extractors (Shown in Section 2.7.4) normally named after their proposed output, e.g. Category Extractor extract category from Wikipedia. By using these extractors we turn a specific feature of Amharic Wikipedia into triples and each feature has its own specific information. All of available extractors are useful for DBpedia extraction because each extractor extract unique feature of Wikipedia. These extractors given to extractor the framework extract automatically without further intervention.

As shown in Figure 5.8, wiki parser transforms the source code of a Wikipedia page into an Abstract Syntax Tree (AST). Each RDF graph contents identified by the AST and the extractors extract based on its own feature.

²³ <http://www.w3.org/International/O-URL-and-ident.html>

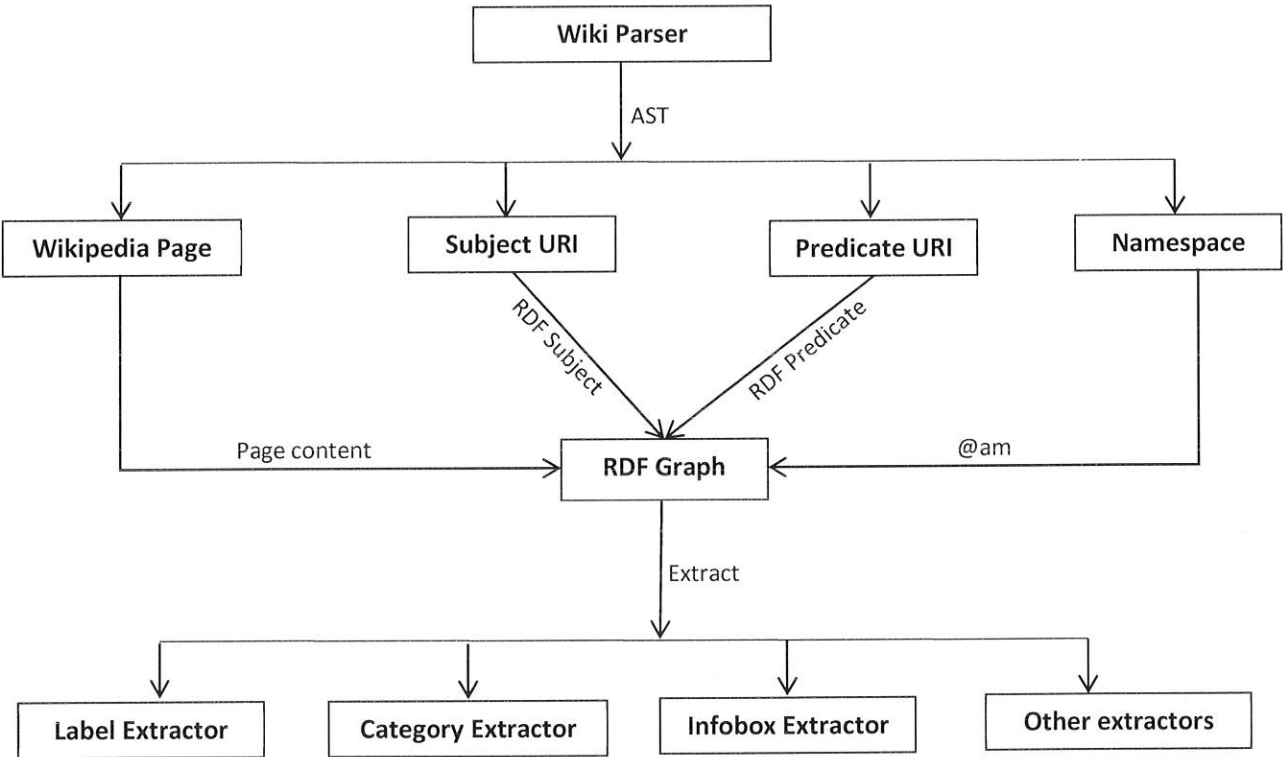


Figure 4.5 Extractor flow diagram

4.1.8. Amharic DBpedia RDF Serialization

DBpedia uses different formats of RDF serialization. The most common representations are N-triple, turtle and RDF/JSON. N-Triple²⁴ is a default serialization format of I18n framework to represent DBpedia RDF triples and it is a format for storing and transmitting data in the form of line-based plain text serialization of RDF and a subset of the Turtle. It can serialize Amharic RDF triples by N-triples. This can be applied as: subjects and predicate of N-Triple take the URI and objects can be a URI or a literal.

A Turtle²⁵ document is a textual representation of an RDF graph. It can allow writing down Amharic RDFs in a compact textual form. All RDF graphs are made up of triples consisting of a subject, predicate and object and each statement are a sequence of (subject, predicate, object)

²⁴ <http://www.w3.org/TR/n-triples>

²⁵ <http://www.w3.org/TR/turtle>

terms, separated by whitespace and terminated by single quote and full stop (' ' .). The advantage of turtle compared with N-Triple, it supports Unicode characters of human readable format.

4.1.9. Triple Dump

This stage is the destination of extracting Amharic RDF triples and each triple are stored in text file format based on their structure. Amharic language follows the standard DBpedia format to represent RDFs. As a rule representation of DBpedia resources is in the form of DBpedia web address (<http://dbpedia.org/resource/<page name>>) which are all entities in DBpedia are assigned a unique URI of the form <http://dbpedia.org/resource/Name>, where Name is taken from the URL of the source Wikipedia article and it can be defied as <http://am.wikipedia.org/wiki/ድህረገጽ>.

4.2 Amharic Wikipedia Infobox Application Issues

Because of lack of central coordination between Wikipedia editors, Wikipedia's infobox template system doesn't have a uniform standard (Dimitris Kontokostas, 2012). Amharic infobox also has shared this problem. Those are:

- *Unrelated attribute/value presentation*: using dissimilar attribute/value.
E.g. attribute Full Name (ሙሉ ስም) has a value Occupation Writer (ፀሐፊ)
- *Empty value*: unknown reason, Wikipedia authors omits attribute values.
- *No-Infobox*: some Wikipedia pages do not have infobox at all.
- *Inconsistency*: using different property to refer the same value. In some cases different Wikipedia contributors follow different templates to refer the same thing. English Wikipedia has an inconsistency problem (Mohamed Morsey, 2012). E.g. infobox referring a word *City* by using different attribute (infobox_city_japan and infobox_swiss_town). Arabic Wikipedia (Al-Feel, 2013) also has similar problems (E.g. referring a person birth place it use *place of birth* and *birthplace*. Like English and Arabic, Amharic infobox also has the problem of inconsistency.
E.g. *Date of Birth* (የትውልድ ቀን) / *Date of Birth* (የትውልድ ዘመን) and የመሪዎች ስም / *Chairperson* (ሊቀመንበሮች).

Infoboxes that appear in a Wikipedia article are based on a template²⁶ that specifies a list of attributes that can form the infobox. A wide range of infobox templates are used in Wikipedia. Common examples are templates for infoboxes that describe Country, Persons, and Organizations. Each template will have its own subpart to give detail description. For example Person template has Person_ Politician, Person_ Artist, Person_ Athlete, etc. English ‘Person_ Politician’ and Amharic ‘Infobox_ Country’ infobox templates (shown in Figure 4.6 and Figure 4.9 respectively), has all basic information to declare their specific thing. When we come to Amharic infoboxes politician infobox template called የፖለቲካ ሰው መረጃ and Infobox_ Country which is የፀገር መረጃ show in Figure 4.7 and Figure 4.8 respectively, attribute numbers are limited and it give partial information about their specific things compared to English template.

²⁶ Annex I

```

{{{የፖለቲካ ሰው መረጃ
|ስም =
|ሰዕል =
|የሰዕል_መግለጫ =
|ቢድ =
|ቀናት =
|ፕሬዝዳንት =
|ቀዳሚ =
|ተከታይ =
|ቢድ2 =
|ቀናት2 =
|ጠቅላይ_ሚኒስትር2 =
|ቀዳሚ2 =
|ተከታይ2 =
|ሌላ_ስም =
|የተወለዱት =
|የሞቱት =
|ዜግነት =
|ፓርቲ =
|ባለቤት =
|ሴጃች =
|ሀይማኖት =
|ፈርማ =
}}}

```

Figure 4.7 Amharic
Person_Politician Infobox

```

{{{ Person_Politician
!name =
!image =
!alt =
!caption =
!order =
!office =
!vicepresident =
!term_start =
!predecessor =
!jr/sr2 =
!state2 =
!term_start2 =
!term_end2 =
!predecessor2 =
!successor2 =
!office3 =
!term_start3 =
!term_end3 =
!predecessor3 =
!successor3 =
!birth_name =
!birth_date =
!birth_place =
!nationality =
!party =
!parents =
!relations =
!spouse =
!children =
!residence =
!alma_mater =
!profession =
!religion = ....
}}

```

Figure 4.6 Excerpt of English
Person_Politician Infobox
template

```

{{የሀገር መረጃ
|ስም =
|ሙሉ_ስም =
|ባንጫ_ሥልጣን =
|ባንጫ_ስፋት =
|መዝሙር =
|ካርታ_ሥልጣን =
|ካርታ_መግለጫ_ፊላዊ =
|ቅር_ከተማ =
|ባሕር_ዳንቅ =
|መንግስት_አይነት =
|መሪዎች_ማዕረግ =
|መሪዎች_ስም =
|ታሪካዊ_ቀናት =
|ታሪካዊ_ክስተቶች =
|መሪዎች_ስፋት =
|መሪዎች_ስፋት_ከዓለም =
|ውሀ_ከመቶ =
|የሕዝብ_ብዛት_ግምት_ዓመት =
|የሕዝብ_ብዛት_ፍጠራ_ዓመት =
|የሕዝብ_ብዛት_ግምት =
|የሕዝብ_ብዛት_ፍጠራ =
|የሕዝብ_ብዛት_ከዓለም =
|የገንዘብ_ስም =
|ሰዓት_ክልል =
|የክልል_መግቢያ =
|ከፍተኛ_ጆረጃ_ከባቢ =
|ግርጌ_ማስታወሻ =
}}
```

Figure 4.9 Amharic Country Infobox template

```

{{Infobox country
|conventional_long_name =
|common_name =
|image_flag =
|image_coat =
|symbol_type =
|national_motto =
|national_anthem =
|image_map =
|map_caption =
|alt_map =
|image_map2 =
|alt_map2 =
|map_caption2 =
|map_width =
|capital =
||largest_city =
|official_languages =
|languages_type =
|languages =
|regional_languages=
|official_religion =
|demonym =
|government_type =
|leader_title1 =
|leader_name1 =
.....
```

Figure 4.8 Excerpt of English Country Infobox template

4.3 Amharic Infobox Improvement

Infobox quality issues appear in different reason. To increase knowledge representation of Wikipedia as well DBpedia, each problems should be fixed to get high quality of information (knowledge representation). The best way of completing incomplete infoboxes is performed by Wikipedia writers and editors. But problems like inconsistency can be improved without Wikipedia writers. To do this, it has two options to update the knowledge. The first option directly applies the updates on Wikipedia source. The second option is applying the correction in extracting RDF triples.

The extracted triples have inconsistency problems. We write a script to correct inconstant representation of infobox property values. To update the inconsistence infobox and make it consistence, first it need to check the existing infobox with the list of wrong infobox words. If the comparison is match with one of the list of words, the infobox is updated by the correct one.

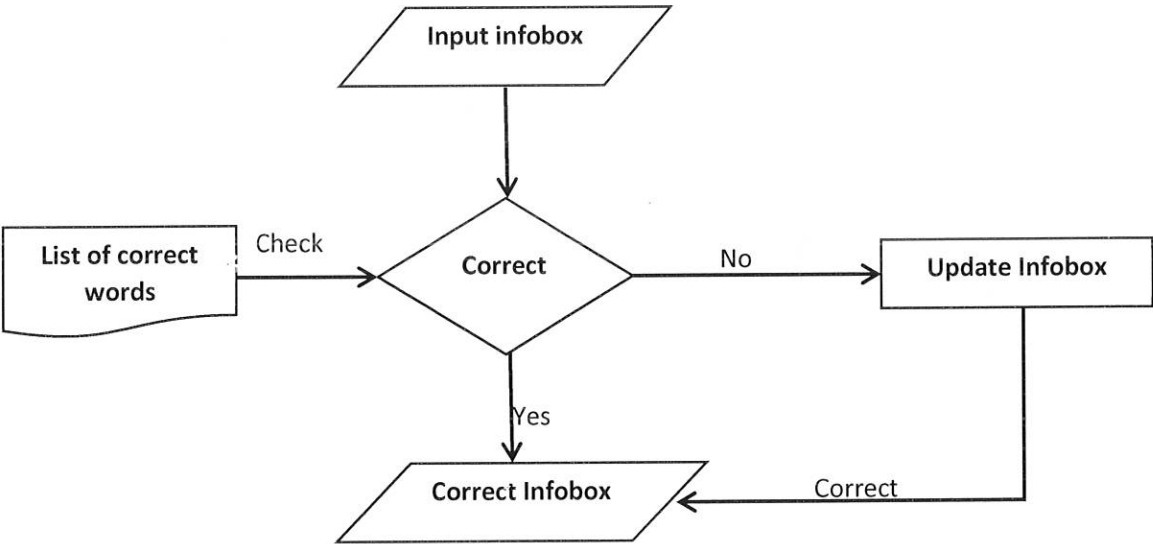


Figure 4.10 Flow diagram of Amharic infobox improvement

CHAPTER FIVE

EXPERIMENTATION AND EVALUATION

This chapter describes the implementation of Amharic DBpedia extraction. In the implementation process of Amharic DBpedia extraction, major activities are: Page Collections, Extractors, Parsers, and RDF outputs are applied to extract Amharic DBpedia datasets. In addition to this the evaluation and improvement of extraction has been discussed in this chapter.

5.1 Architecture of DBpedia Extraction Framework

The Amharic DBpedia uses the same framework (I18n extraction framework)²⁷ to extract triples from Wikipedia as other language versions. The framework needs a specific configuration for languages that are use Unicode character. To extend the framework with pluggable internationalization components in order to facilitate the extraction of localized information, it requires further adaptation to extract Amharic Wikipedia. The adaptation includes Amharic language selection and RDF triple serialization selection. After the application of a proper configuration, it can generate and maintain DBpedia editions.

Workflow of the extraction framework to extract Amharic DBpedia is; first each Amharic page is retrieved from a local Wikimedia dump to local server. Then the extraction process is configured and started. In the process requests pages from the source, passes them to the extractors and extracted RDF triples will be generated. As shown in 4.1.5, the extractor uses the parser to identify specific data and filters splitting unwanted markups. The generated data is finally sent to a destination, as Amharic RDF triples.

Figure 5.1 shows an overview of Amharic DBpedia extraction framework. The structure of DBpedia extraction starts from the input part that provide Amharic Wikipedia page from a local dump. The input Wikipedia pages parsed and the extractor extract and generate RDF triple as an output.

²⁷ <http://wiki.dbpedia.org/Documentation>

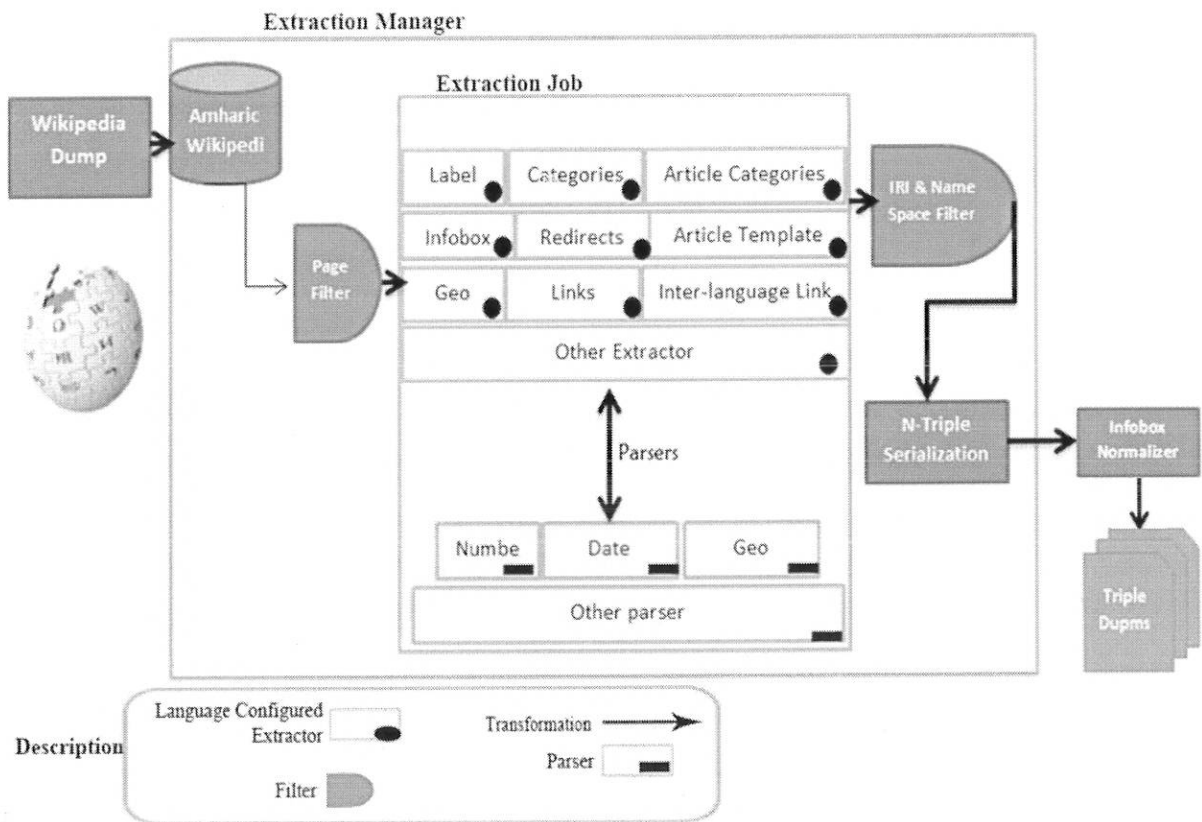


Figure 5.1 Overview of Amharic DBpedia Extraction Framework

- **IRI Filter:** The default web address representation of the framework is URI. To identify Amharic resource character on the web, it must be represented by IRI. We choose *'format.ttl.gz=turtle-triples;uri-policy.iri/'* in the framework to enables the extraction of resources in the IRI form.
- **Namespace Filter:** extracted Amharic RDF triples have only 'am' namespace. All of subject of RDF triples starts with *http://am.dbpedia.org/resource/*, where 'am' stands for representing namespace for Amharic language. To have this representation *'language='am''* include in the extraction framework. Figure 5.3 shows example of extracted RDF triple.

```

<http://am.dbpedia.org/resource/ኢትዮጵያ>

<http://am.dbpedia.org/property/መጠን>

"የኢትዮጵያ ፌዴራላዊ ዲሞክራሲያዊ ሪፐብሊክ"@am .

```

Figure 5.2 Sample Amharic DBpedia RDF with namespace

- **Extraction Job:** Main extraction work performs in here. Extraction job includes seventeen features of DBpedia datasets and each feature (Shown in 2.7.4) is a content of Amharic Wikipedia that has specific content and representation.
- **Parsing:** As shown in 4.1.7, the Wiki Parser package specifies a parser, which transforms MediaWiki page source into an Abstract Syntax Tree (AST). Through this process, all numbers (DateTime, Units, Geo-coordinates and Numbers) are identified by the parser.
- **Triple Dump:** The extracted output has a number of RDF triples. Each triple has RDF representation of subject, predicate and object and all are stored based on its own feature. Finally extracted triples are dumped into text file format.

5.2 Installation and Extraction of DBpedia Extraction Framework

Installing and using of DBpedia extraction framework will be present in this section. A step to extract Amharic DBpedia follows these steps:

1- Step One

Clone the DBpedia extraction framework in to local computer by using ‘git clone git clone <https://github.com/dbpedia/extraction-framework.git>’. This framework is available in Git²⁸ hub repository. After the extraction framework is dump in to local computer, it will be ready for dumping Wikipedia and Amharic Wikipedia.

²⁸ <https://github.com/dbpedia/extraction-framework.git>

2- Step Two

Once the deposit has been cloned, the basic configurations applied in the framework to download Wikimedia and Amharic Wikipedia. These configurations are:

- Dumping Wikimedia in to base directory by 'base-url=http://dumps.wikimedia.org/'
- Download Amharic articles by 'download=am: pages-articles.xml.bz2'.
- Downloading Wikidata by 'download=wikidata: pages-articles.xml.bz2'.

We dump Wikipedia collection form Wikipedia website and we use the extraction framework to import Amharic Wikipedia from this collection in to a local computer by using 'am' language feature of Amharic.

3- Step Three

Installation of the required software²⁹ used as a plat form for MySQL and MediaWiki to have a local Amharic Wikipedia in local computer. The compulsory steps to install Amharic Wikipedia in to local computer are:

- **MySQL:** From the local dump, Amharic Wikipedia imported in the MySQL database. To have Amharic specifically it need to configure pom.xml file³⁰.
- **Configuration of MediaWiki:** Installation of Amharic Wikipedia in local computer. A specific configuration of Amharic language of '*\$languageCode = "am"*' in the LocalSettings.php file where 'am' is a language code of Amharic language.

After the import of Amharic Wikipedia into local computer, it can get local mirror (Shown in Figure 5.2) of Amharic Wikipedia and all contents are ready for extraction process.

²⁹ PHP, Scala, Maven, Apa

³⁰ Annex K

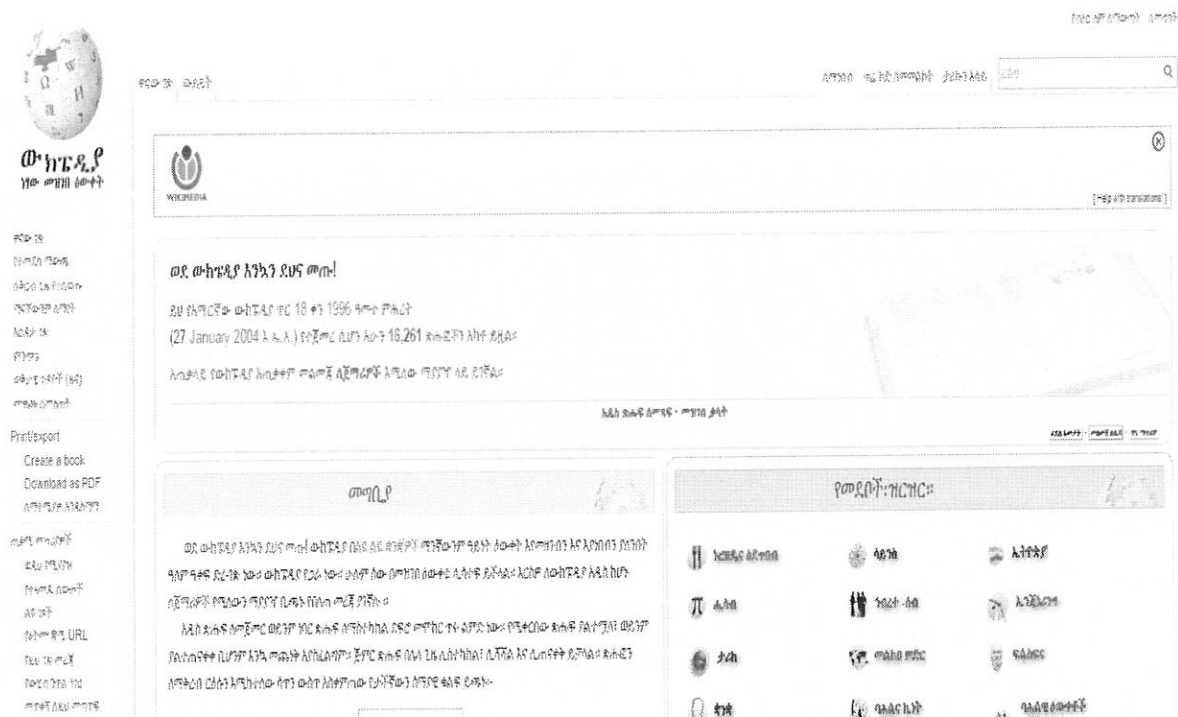


Figure 5.3 Local Amharic Wikipedia

4- Step Four

This step is to make the extraction of data from the Amharic Wikipedia dump. Before to start the extraction process the basic configurations apply in framework are:

- Selecting Amharic language by 'languages=am' Where 'am' is the language code of Amharic.
- Selecting extractors by adding components in the frame work these are:
'extractors=.ArticleCategoriesExtractor,.ArticleTemplatesExtractor,.CategoryLabelExtractor,.ExternalLinksExtractor,.GeoExtractor,.InfoboxExtractor,.InterLanguageLinksExtractor,.LabelExtractor,.PageIdExtractor,.PageLinksExtractor,.RedirectExtractor,.RevisionIdExtractor,.ProvenanceExtractor,.SkosCategoriesExtractor,.ArticlePageExtractor'
- Selecting turtle serialization method by: 'format.ttl.gz=turtle-triples;uri-policy.iri' to have IRI for Amharic character representation.

5.3 Extracted RDF Triple Representation

All extracted Amharic RDFs have a unique resource and each resource include any abstract entities. Every resource is identified by a URI which is defined to be a globally unique URL. E.g. *<http://am.dbpedia.org/resource/ፍልስፍና>*. Properties are specific resources, which can be used to describe attributes or relations of other resources.

Statements are used to describe properties of RDF resources in subject, predicate, object (Shown in Figure 5.4). According to the RDF data model all three components are triple resources. Subject and predicates (property) have a URL representation of resource and objects can also be a URL or literals consisting of a data value and optionally an associated a data type or language tag (@am).

Subject: <<http://am.dbpedia.org/resource/አክሱም>>
Predicate: <<http://am.dbpedia.org/property/አካባቢ>>
Object: "አፍሪካ"@am .

Figure 5.4 Amharic DBpedia RDF Example

5.4 Amharic RDF Serialization

Serializing RDF data is a very central issue. There are different formats for serializing RDF data. For Amharic DBpedia RDF representation, we test two types of RDF serializations. These are N-Triple and Turtle.

N-Triples are a line-based, plain text format used for encoding RDF triples. Sample of RDF statements (shown in Figure 5.5) about Ethiopia simply refers as subject identified by <http://dbpedia.org/resource/ኢትዮጵያ> has the property of <http://dbpedia.org/property/የመንግስት-አይነት> whose value is equal to "ፌዴራላዊ ፡ ሪፐብሊክ"@am. As the extracted result shows, the triples are not human readable because N-Triples are not fully support Unicode characters.

Subject: <http://am.dbpedia.org/resource/\u1240\u12F3\u121B\u12CA_\u1283\u12ED\u1208_\u1225\u120B\u1234>
Property: <http://am.dbpedia.org/property/\u1218\u12DD\u1219\u122D>
Object: "\u1293\u1205\u1291 \u1301\u1295\u12F5 \u12A0\u120B\u1205 \u1301\u1295\u12F5 \u12A0\u120D-\u12CB\u1273\u1295"@am.

Figure 5.5 Sample Amharic N-Triples format

A turtle that allows an RDF graph to be completely written in a natural text form to get human readable Amharic (Geez) characters. It provides levels of compatibility with the N-Triples format as well as the triple pattern syntax of W3C recommendation. As shown in Figure 5.6 RDF statements about Ethiopia simply refer as subject identified by `http://dbpedia.org/resource/ኢትዮጵያ` has a property `http://dbpedia.org/property/የመንግስትአደነት` whose value is equal to `"ፌዴራላዊ ሪፐብሊክ"@am.`

Subject: < http://dbpedia.org/resource/ኢትዮጵያ >
Property: < http://dbpedia.org/property/የመንግስትአደነት >
Object: "ፌዴራላዊ ሪፐብሊክ"@am.

Figure 5.6 Example of Turtle Serialization

The two examples of Amharic RDF representations shown in Figure 5.5 and Figure 5.6 to N-Triple and Turtle respectively represent the same information but their Unicode representation is different. N-Triple RDFs failed to support Unicode. Because of this we choose Turtle for Amharic RDF serialization to get readable RDFs.

5.5 Extraction Discussion

5.5.1. Amharic Infobox Extraction and RDF Representation

Infoboxes are the most important dataset for DBpedia extraction. Infoboxes are as often as used to list an article's most important facts as a table. In the extraction of infobox process; all attributes and their values in Amharic Wikipedia infoboxes are extracted and presented in the form of: the article will be the subject; the infobox attribute concatenated with the namespace (<http://dbpedia.org/property/>) will be a predicate, while the value of the infobox attribute in Wikipedia is considered an object and then it can create infobox triples in the following manner:

- **Subject**- <http://dbpedia.org/resource/PageName>, which is subsequently used as the subject. A DBpedia URI is created from the Wikipedia article URL the infobox is contained in. This URI is used as the subject for all extracted triples.
- **Predicate** URI is created by concatenating the namespace <http://dbpedia.org/property/> with the name of the infobox attribute.
- **Objects** are created from the attribute values. Sometimes the object itself can be a subject of another resource.

The extraction framework infobox extractor automatically extracts attribute and values from infobox templates that gets from Amharic Wikipedia. RDF triple infobox representation is: DBpedia URI of the Wikipedia article is used as a subject, Predicate URI is created by concatenating the namespace fragment, <http://am.dbpedia.org/property/> and the name of the infobox attribute and Objects are created from the attribute value. Property values are post-processed in order to generate suitable URI references or literal values. This includes recognizing MediaWiki links, detecting lists, and using units as data types. The advantage of the extraction is its complete coverage of all infoboxes and infobox attributes. Infobox properties and values are Amharic Wikipedia contents. The extractor extracts this content based on their structure (proper/value pairs).

Based on infobox RDF triple representation, the excerpt of extracted infobox RDF triples about Ethiopia is listed below.

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ስም> "ኢትዮጵያ"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ሙሉስም> "የኢትዮጵያ ፌዴራላዊ ፡ ዲሞክራሲያዊ ፡ ሪፐብሊክ"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ማኅተምሥዕል> "Coat of arms of Ethiopia.svg"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ባንዲራሥዕል> "Flag of Ethiopia.svg"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/መዝሙር>
<http://am.dbpedia.org/resource/ወደፊት_ገስግሺ_ውድ_እናት_ኢትዮጵያ> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/መዝሙር> "ስዕል:Wedefit Gesgeshi Widd Innat Ittyoppya.ogg"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ካርታሥዕል> "Ethiopia in its region.svg"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ካርታመግለጫፅሁፍ> "ኢትዮጵያ ፡ በአረንጓዴ ፡ ቀለም"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ዋናከተማ>
<http://am.dbpedia.org/resource/አዲስ_አበባ> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ብሔራዊቋንቋ> "አማርኛ (ሌሎች ፡ ቋንቋዎች ፡ ከልላዊ ፡ ዕውቅና ፡ አላቸው)"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የመንግስትአይነት> "ፌዴራላዊ ፡ ሪፐብሊክ"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የመሪዎችማዕረግ>
<http://am.dbpedia.org/resource/ፕሬዝዳንት> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የመሪዎችስም>
<http://am.dbpedia.org/resource/ጋዴለማርያም_ደሳለኝ> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የመሬትስፋት>
"1104300"^^<http://www.w3.org/2001/XMLSchema#integer> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የመሬትስፋትከዓለም>
"27"^^<http://www.w3.org/2001/XMLSchema#integer> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ውሀከመጽ>
"0.7"^^<http://www.w3.org/2001/XMLSchema#double> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የሕዝብብዛትግምትዓመት> "፳፻ ዓ.ም."@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የሕዝብብዛትቆጠራዓመት> "፲፱፻፺፱ ዓ.ም."@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የሕዝብብዛትግምት>
"79221000"^^<http://www.w3.org/2001/XMLSchema#integer> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የሕዝብብዛትቆጠራ>
"73918505"^^<http://www.w3.org/2001/XMLSchema#integer> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የሕዝብብዛትከዓለም>
"15"^^<http://www.w3.org/2001/XMLSchema#integer> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የገንዘብስም> "ብር"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ስዓትክልል> "+3"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የስልክመግቢያ>
"251"^^<http://www.w3.org/2001/XMLSchema#integer> .



Figure 5.7 An excerpt of Amharic Wikipedia Infobox (Ethiopia)

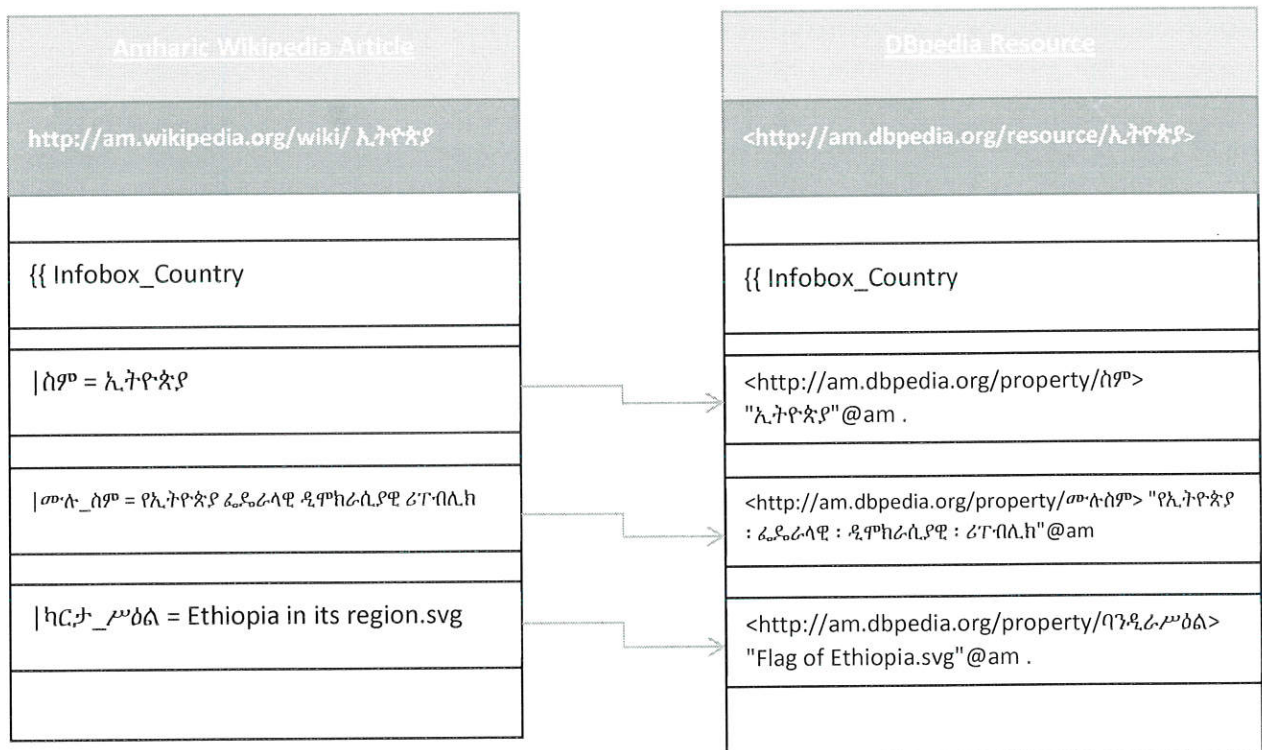
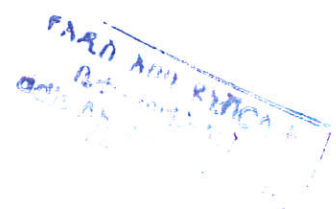


Figure 5.8 An excerpt of Amharic Wikipedia and DBpedia infobox alignment (Ethiopia)



5.5.2. Amharic Inter-language Link Extraction

Inter-language links are creating links from a page in one Wikipedia language to an equivalent page in another language. To extract inter-language link automatically generate by I18n DIF use an extractor called '*InterlanguageLinkExtractor*' (Shown in 5.2). The subject of inter-language link is <http://am.dbpedia.org/resource/> with article title and the extractor use <http://dbpedia.org/ontology/wikiPageInterLanguageLink> as a predicate for all inter-lingual RDF triples. The object of the triple is the language match of article title.

We extract Amharic RDF inter-language triples and each link (shown in Figure 5.9) has a capability to indicate their equivalent inter-language link. One of the uses of inter-language links is it can query RDFs by one language and get result in different languages.

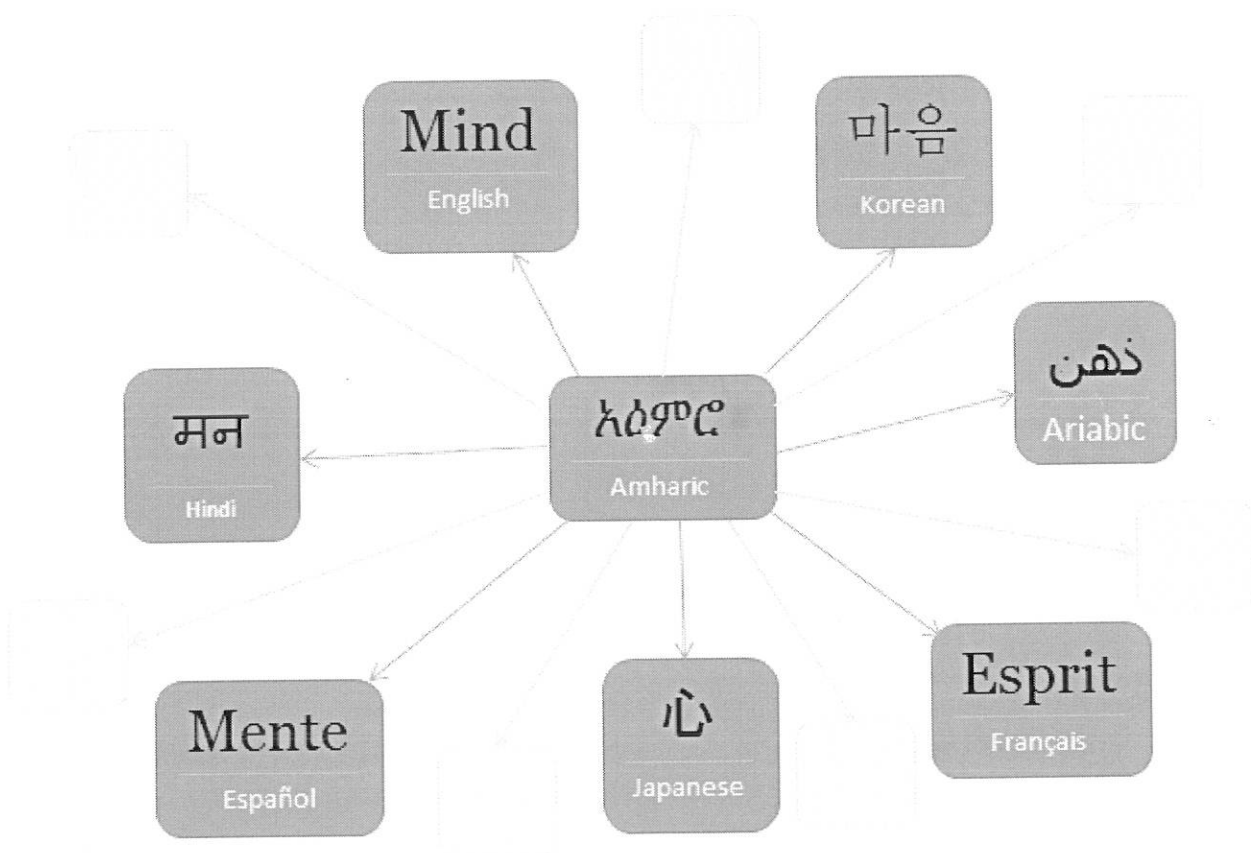


Figure 5.9 Sample Inter-language link pictorial representation

```

<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://eo.dbpedia.org/resource/Menso>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://fa.dbpedia.org/resource/آمنو>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://fr.dbpedia.org/resource/Esprit>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://gan.dbpedia.org/resource/心精>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://ko.dbpedia.org/resource/마음>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://bn.dbpedia.org/resource/মন>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://ia.dbpedia.org/resource/Mente>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://bg.dbpedia.org/resource/Дух_(фил.>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://it.dbpedia.org/resource/Mente>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://he.dbpedia.org/resource/מוח>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://la.dbpedia.org/resource/Mens>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://ml.dbpedia.org/resource/മനസ്സ്>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://fr.dbpedia.org/resource/Esprit>
<http://am.dbpedia.org/resource/አሜሮ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink> <http://ja.dbpedia.org/resource/心>

```

Figure 5.10 Sample Amharic inter-language RDF triples

The above figure (shown Figure 5.10) shows an excerpt of inter-language RDF triples talked about ‘aemero’ አሜሮ called Mind in English. The subject is <http://am.dbpedia.org/resource/አሜሮ>. The predicate is <http://dbpedia.org/ontology/wikiPageInterLanguageLink> for all triples and the object is a content talked about mental in other languages.

5.5.3. Amharic DBpedia RDF Graph

RDF graph is a set of RDF triples and each triple consists of subject, predicate and object. The subject is IRI, the predicate is IRI and the object can be and IRI or a literal (literal can be string or number). IRI representation help to read Unicode characters like Amharic writing system used.

There are two ways of RDF representation. These are statement and graphical representation. RDF as a graph expresses exactly the same information as RDF written out as triples, but the graph form makes it easier for us humans to see structure in the data. RDF graphs consist of nodes interconnected by edges.

Each arrow or edge is a part of RDF statement represents:

- The name at the start of the arrow is the statement's subject,
- The name at the end of the arrow is the statement's object,
- The name that labels the arrow is the predicate.

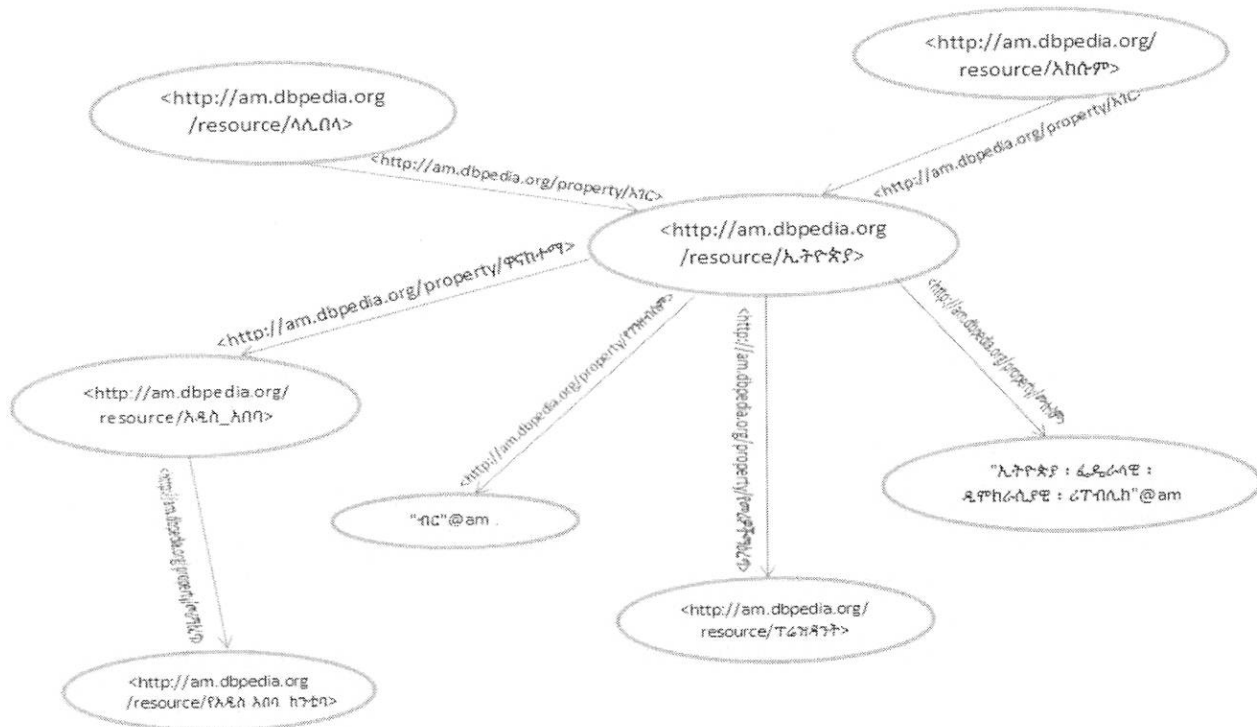


Figure 5.11 Sample DBpedia RDF Graph

Each RDF graph is a set of triples. Each RDF has two nodes (Subject and Object) and a triple connecting them (Predicate). An RDF graph can be visualized as a node and directed-arc diagram (Shown in Figure 5.11), in which each triple is represented as a node-arc-node link.

5.6 Extraction Results

Feature extractors normally named after their proposed output (Shown in 4.1.7). To extract Wikipedia features, the I18n DIF has different types of extractors (Shown in 5.2) that are specialized in extracting a single feature from Amharic Wikipedia. These extracted triples categorized based on their features. Such as: Article Category from Wikipedia Category, Geo-coordinates from Wikipedia Geographic location. We extract more than 349,000 triples from Amharic Wikipedia. Extracted RDF triples from Amharic Wikipedia are shown Table 5.1.

Types of Extractor	Description	Triples
Article Template	Standard article template	17,998
Article Categories	Extracts links from concepts into categories using the SKOS vocabulary.	15,968
Category- Labels	Extracts labels in Categories.	890
External - Links	Articles contain references to external Web resources which we represent using the DBpedia property. Extracts links to external web pages.	2,499
Geo-coordinates	The Geo-extractor expresses coordinates using the Basic Geo vocabulary and W3C Geospatial Vocabulary to extract Geo-coordinates.	1,000
Infobox Property	Extracts all properties from all infoboxes.	27,543
Infobox Property Definition	Property of infobox values.	946
Interlingua Link	These are links that connect articles about the same topic in different language editions of Wikipedia and use them for assigning labels and abstracts in different languages to DBpedia resources.	1,068
Label	All Wikipedia articles have a title and extracts labels to articles based on their title.	19,106
Page - Id	Extracts page ids of articles.	22,850
Page Links	Internal links between DBpedia instances derived from the internal page links between Wikipedia articles.	109,075
Redirects	in order to identify synonymous terms, Wikipedia articles can redirect to other articles.	5,780
Revision - Id	Extracts revision ids of articles.	22,850
Revision - URIs	Each time when a Wikipedia article is modified, it gets a new Revision ID. This extractor extracts that ID	22,850
Wikipedia Links	Links a Wikipedia article to its corresponding DBpedia resource.	76,420
SKOS categories	Extract information about which concept is a category and how categories are related using the SKOS Vocabulary.	3,041
<i>Total Triples=</i>		349,902

Table 5.1 Extracted triples

5.7 Amharic and Other Language DBpedia Statistics

The average extracted triple per page is the ration of number of extracted RDF triples (346,843) and source Amharic Wikipedia (approximately 15,800 numbers of pages). Percentage of this ratio is 21.9%. This number will increase if the remaining features (Home page and Abstract) extracted. In addition to this Amharic Wikipedia not well organized and its infoboxes are not rich (shown in 4.1.4) like English or other well developed languages.

The average percentage of extracting triple per page is 27% and 31%, respectively (Dimitris Kontokostas, 2012). These two languages have a complete DBpedia extraction but the number still needs improvement. The common reason is Wikipedia writers are not complete some page with all basic Wikipedia features. This is because some Wikipedia pages not have all kinds of features³¹.

Previous research shows only 33% of Wikipedia pages have an infobox (Lange Dustin, 2010). We also estimate the average number of extracted Amharic infobox per page. To calculate the average, we consider the total number of Amharic Wikipedia page and minimum number of infobox properties from extraction datasets. The result shows 21% Amharic Wikipedia has infobox.

5.8 Amharic Infobox Quality Improvement

It can extract Amharic infobox form Amharic version of Wikipedia without any improvement, but the extracted result may affect the quality of knowledge representation. To get a better quality of infobox information, Infobox property and value quality improvement has been done by making property values consistence (normalize). To avoid inconsistency, it can use a single property to represent one value. E.g. Selecting one of the two (የትውልድ ቀን or የትውልድ ዘመን) properties to represent 'Date of Birth'. As it mentions Section 4.2, we select property values that have inconsistency issue. The application of normalization is done by writing python script.

³¹ <http://www.socialfresh.com/wikipedia-infographic/>

5.9 Evaluation

This section describes the extracted Amharic RDF triples quality evaluation methods. Knowledge resource evaluation is difficult and there is no common approach for direct method of evaluating the knowledge base (Cuadros, 2006). We consider all extracted RDF triples in the evaluation. The evaluation of Amharic DBpedia extraction is done in two ways. These are:

5.9.1. Evaluation of Extracted RDF Triples against Source on Amharic Wikipedia

This evaluation of Amharic RDF extraction is done by comparing the number of extracted RDF triples with Amharic Wikipedia. For this evaluation, extracted Amharic RDF triple features are compared with the Amharic Wikipedia source of its own features. (E.g. Category RDF with Amharic Wikipedia Category). The objective of this evaluation is how RDF triples extracted correctly from Amharic Wikipedia source may and how many triples are missed by the extractor.

The most common way of evaluating knowledge extraction is precision and recall (Hissar, 2011). Precision and recall are the evaluation parameters of extraction. Precision (P) is the ratio of the number of correctly extracted RDFs to the total number this correct RDFs and extracted triples with errors. Recall (R) is the ratio of the number of extracted correct RDFs to the total number this correct RDFs and RDFs missed by extractor.

$P = CT / (CT + RT)$ CT denotes the number of correctly extracted RDF triple and RT denotes the number of extracted Wrong RDF triples

$R = CT / (CT + MT)$ MT is the number of RDFs triples which are missed by the extractor.

$F = 2PR / (P + R)$ F stands for an F - measure. It is a harmonic mean of precision and recall.

5.9.2. Manual Evaluation

This section describes a generalized methodology for manual assessment. The assessment methodology of manual evaluation follows these steps:

Step I - Resource selection: resources are selected in:

- *Completely random:* a random resource selection of extracting Amharic datasets.
- *Manual:* a resource selected manually from the extracted Amharic datasets.

Step II - Evaluation method: the assignment of the resources to a person³² selected in Step-I be done manually.

Step III - Resource evaluation: in the case of manual assignment of resources, the person evaluates each resource (extracted Amharic RDFs) by checking the RDF triples in Amharic Wikipedia to detect the potential data quality problems. For manual evaluation, we give a randomly selected RDF triples to individual volunteers to evaluate the extracted Amharic RDFs.

Based on the above steps the person takes RDF triples and check that the extracted RDF is really exist in the online Amharic Wikipedia website.

5.10 Evaluation Results and Discussion

5.10.1. Evaluation Result of Extract RDFs Triples against Source on Amharic Wikipedia

Table 6.4 illustrates the experimental result of extracting Amharic RDFs against Amharic Wikipedia. To count the number of Wikipedia feature in the source file, we write a python script³³. The number of the difference between source Wikipedia feature and extracted triple is count as a possible correct RDF triple but not extracted.

³² A persons involved in this test has at least first degree in Computer since or related fields. Their activity is just check the extracted RDF triples really exist in Amharic Wikipedia.

³³ Annex O

	ET	WC	DF	RT	P	R	F
Article Categories	15,968	16,302	334	215	0.98654	0.97924	0.98287
Category- Labels	890	996	106	29	0.96742	0.89038	0.9273
Page - Id	22,850	23,348	498	205	0.99103	0.97848	0.98472
Redirects	5,780	5,874	94	78	0.98651	0.98378	0.98514
Revision - Id	22,850	24,225	1,375	182	0.99204	0.94281	0.9668
Revision - URIs	22,850	24,225	1,375	208	0.9909	0.94275	0.96622

Table 5.2 Amharic RDF extraction test result

WC= Source Amharic Wikipedia content, **ET**= correctly extracted triples, **RT**= incorrect triples, **DF**= the difference between **WC** and **ET**, **P**= precision, **R**= recall, and **F**= F-measure

The average precision of Amharic RDF extraction falls into between 99% and 96. The recall also shows the results are between 89% and 97%. Accuracy of the extracted RDF triples is in the range of 92% and 98%.

5.10.2. Manual Evaluation Result

To apply manual evaluation, we give extracted RDF triples for volunteers to check the triples are present in online Amharic Wikipedia or not. As shown in Table 5.3, there are seven evaluators are participate in this evaluation. The evaluator select a single RDF triple at a time and the outcome of the result shows out of 468 tested RDF triples 84 RDF triples shows specific errors. The percentage shows out of tested RDF triples 92.6% are correct and 7.4% have a certain issues. Common errors identified by manual evaluation are:

- RDF object value is incorrectly extracted: the reason of this problem it can be:
 - *RDF values are incorrectly extracted*: This resource about Place but the object value shows Ethiopian Calendar date format which is wrong.
E.g. "፳፻፲፱.፳፻፲፱" @am.

- *Object values incompletely extracted:* This resource about date of birth of a person is extracted, but it shows only the date.
- *Object values have incorrect information:* subject and object values are not a represent similar thing.
- Some articles are not found in the Amharic Wikipedia at all.

Total no. of users	7
Total no. of resources evaluated	468
Total no. of resources without problems	384
Total no. of resources with problems	84

Table 5.3 Manual evaluation result

5.11 Result Discussion

Evaluation of knowledge extraction is not striate forward (Cuadros, 2006). We follow two ways to evaluate extracted Amharic RDF results.

The first evaluation measure accuracy of extracting Amharic RDF triples by comparing with a source of Amharic Wikipedia. Average result shows the accuracy of extracting triple against Wikipedia data source is 97%. The number shows most of extracting RDF triples found in the Amharic Wikipedia source and some of them are missing because some Amharic Wikipedia contents use a wrong template or its property values either incorrect or empty.

Even if, the first (Shown in Section 5.8.1) test results show high accuracy, this may not enough to show performance and information representation of extracting Amharic RDFs in real world knowledge representation. To complete this gap, we use the manual evaluation to measure the accuracy of extracting RDFs by checking the presence of the sample extracted RDFs from the online Amharic Wikipedia site. The result of this experiment shows some errors that are not found by the first test methods. The reasons of this, Wikipedia writers/editors add wrong information in the page. The best example to show this

(<<http://am.dbpedia.org/property/ብሔራዊቋንቋ>> አማርኛ), has correct tag and found in both extracted RDF and Amharic Wikipedia content, which make correct first and second test, respectively. But this information representation is wrong. The main advantage of manual evaluation is to spot this kind of issues.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

This chapter summarizes the research work, highlights the main contributions, recombination of some remaining work for future research and gives general conclusions over the work. It then pinpoints the future directions conducted in this research area also presented.

6.1 Conclusion

People, as an intelligent creature, have a way to exchange information across themselves. However, computers not have this sort of ability unless they train by a specific way. The Semantic Web aim is to change that by presenting Web page data in such a way that it is understood by computers, enabling machines to do the searching, accumulating and combining of the web data without a human administrator. Semantic Web technology gives a standard method for combining information from different sources, it gives a simpler approach to discover, communicate, reuse and consolidate data more easily. To do this RDF is utilized to transform essential web data into organized information that software can make utilization of it.

Working with non-English, under-researched languages like Amharic highlights practical issues. The research has its own value and benefits regarding to DBpedia extraction. This paper has described how Amharic DBpedia is extracted from Wikipedia. Most importantly, the research shows the challenges and achievement of Amharic DBpedia extraction.

This research shows to look in to techniques of Amharic DBpedia extraction. Through the study, preprocessing and fixing issues of Amharic Wikipedia, DBpedia extraction framework configuration, Amharic RDF triple representation and evaluation has been done. As shown in this research, it can fix Wikipedia issues to get high quality of knowledge source. This will increase the outcome of the RDF triple qualities. Increasing the number of Amharic Wikipedia content also will increase the potential of knowledge coverage of Amharic DBpedia. As seen in other DBpedia language chapters, extracting DBpedia require a lot of time and effort because of this the complete application is done by community effort.

6.2 Recommendation

The results found in this research showed that DBpedia extraction for Amharic can be done. However, it is also learnt that further research and development effort is needed so as to get a complete Amharic DBpedia chapter. RDF triples like Abstract and Amharic Wikipedia home page extraction are not included in this research because these extractors require more time and effort to configure extraction framework.

It has been claimed that having a complete (quality) Amharic Wikipedia is a fundamental step to get the complete knowledge source. To get a high standard knowledge source, first Wikipedia writers (editors) must follow the Wikipedia page structure. This kind of commitment is reducing the burden of researchers who use Wikipedia as a source for research

To increase knowledge representation of Amharic DBpedia, it needs further updating in the source (Wikipedia), especially on the infoboxes. Amharic infobox templates should follow the English infobox template as a standard because the English Wikipedia format has complete content and other languages follow this template. Follow this standard used for not only data quality, but also used for DBpedia infobox ontology mapping. To have a complete Amharic DBpedia chapter, ontology is also a main part in the development of the DBpedia knowledge base. The ontology characterizes the terms and relationships among each term. Each infobox is mapped to a class and each attribute is mapped to a property in the DBpedia ontology. Amharic infobox mapped to ontology based on the English infobox structure, is considered as a full and well organized infobox. To do this, first Amharic infobox need to include all information like English have. Mapping Amharic infoboxes into ontology share the knowledge base and when querying Amharic DBpedia, it can get a related result in other languages. As we can see from previous research and the application of DBpedia ontology, Amharic DBpedia ontology is done by community effort.

REFERENCES

- Afework, Y. (2011). Amharic text classification. *Masters Thesis, Addis Ababa University*, pp. 50-53.
- Al-Feel, H. (2013, October). A Step towards the Arabic DBpedia. *International Journal of Computer Applications*, pp. 27- 32.
- Alike, S. (2014). http://meta.wikimedia.org/wiki/List_of_Wikipedias. Retrieved April 29, 2014, from www.wikimedia.org.
- Bergman, M. (2008, June 27). Wikimedia Architecture. *Wikimedia Foundation*, pp. 12.
- Berners-Lee, T. (2001, May). The Semantic Web. *Scientific American*, pp. 1-3.
- Berners-Lee, T. (2009, June 18). *Linked Data*. Retrieved April 11, 2014, from <http://www.w3.org/>.
- Boyd, D. (2005, January 4). *Academia and Wikipedia*. Retrieved April 2, 2014, from <http://many.corante.com/>.
- Charalampos Bratsas, L. I. (2011). DBpedia Internationalization: A Graphical Tool for I18n Infobox-to-Ontology Mappings. 56 - 56.
- Christopher, S. (2014, May 17). Retrieved July 19, 2014, from www.dbpedia.org.
- Cuadros, M. P. (2006). An Empirical Study for Automatic Acquisition of Topic Signatures. *Global Wordnet Conference*, pp. 3.
- Dimitris Kontokostas, C. S. (2012, January). Internationalization of Linked Data. The case of the Greek DBpedia edition. *Web Science Program*, pp. pp. 1-31.
- ELI. (2007). 7 things you should know about Wikipedia. *EDUCAUSE Learning Initiative*.
- Eun-kyung Kim, M. W.-S. (2010). Towards a Korean DBpedia and an Approach for Complementing the Korean Wikipedia based on DBpedia.
- Florian Bauer, Martin Kaltenböck. (2009). *The Essentials Linked Open Data*. Vienna, Austria: mono.

- Fred Fonseca, M. B. (2011). Semantics in Semantic Web: A Critical Evaluation. *Knowledge Organization Journal*, pp. 20-23.
- Georgi Kobilarov, C. B. (2009, April 20). DBpedia A Linked Data Hub and Data Source for Web and Enterprise Applications. *Proceedings of Developers Track of 18th International World Wide Web Conference*, pp. 1-3.
- Giles, J. (2006, March 28). Internet encyclopaedias go head to head. *Encyclopædia Britannica, Inc.*, pp. 900-902.
- Gulzar, A. (2007). Introduction to Unicode. *Center for Research in Urdu Language Processing*, (pp. 3-7).
- Hissar, B. (2011, September 16). Machine Reading Between the Lines: A Simple Evaluation Framework for Extracted Knowledge Bases. *Proceedings of the Workshop on Information Extraction and Knowledge Acquisition*, pp. 1-2.
- Ian Jacobs, Norman Walsh. (2006, July). Architecture of the World Wide Web.
- Jens Lehmann, S. H. (2009). DBpedia Live Extraction. *8th International Conference on Ontologies*, pp. 2-4.
- Kim, E.-k. (2010). Towards a Korean DBpedia and an Approach for Complementing the Korean Wikipedia based on DBpedia .
- Krupansky, J. W. (2009, January). www.agtivity.com. Retrieved from http://agtivity.com/def/semantic_web.htm.
- Lange Dustin, B. C. (2010). Extracting Structured Information from Wikipedia Articles to Populate Infoboxes. pp. 3-6.
- Lars Asker, A. A. (2008). Classifying Amharic Webnews. *Springer Verlag*, pp. 19.
- Majdi Beseiso, A. A. (2010). A survey of Arabic language Support in Semantic web. pp. 1-2.
- Mazzocchi, S. (2001). Toward the semantic web. pp. 1-12.
- Michael Kifer, Jos de Bruijn, Harold Boley, Dieter Fensel. (2005). A Realistic Architecture for the Semantic Web. pp. 3-5.

- Mohamed Morsey, J. L. (2012). DBpedia and the live extraction of structured data from Wikipedia. *DBpedia and*, pp. 157-181.
- Obitko, M. (2007). <http://obitko.com/>. Retrieved March 23, 2014, from <http://obitko.com/tutorials>.
- Pollock, J. r. (2009). *Semantic Web For Dummies*. Indianapolis,USA: Wiley Publishing, Inc.
- Rada Mihalcea, Andras Csomai. (2007). Wikify! Linking Documents to Encyclopedic Knowledge. pp. 1-3.
- Sahnwaldt, C. (2013, Sep). <http://wiki.dbpedia.org/Datasets>. Retrieved 2014, from <http://wiki.dbpedia.org>.
- Schiff, S. (2006). Know It All. *The New Yorker*, pp. 4-7.
- Sebastian Hellmann, C. S. (2012). The German DBpedia: A Sense Repository for Linking Entities. In S. N. Christian Chiarcos, *Linked Data in Linguistics* (pp. 181-191).
- Sporny, M. (Director). (2012). *What is Linked Data* [Motion Picture].
- Stadler, C. (2010, May). Community-Driven Engineering of the DBpedia Infobox Ontology and DBpedia Live Extraction. pp. 20-27.
- Suh, B. (2011, July 15). The singularity is not near: slowing growth of Wikipedia. *WikiSym '09 Proceedings of the 5th International Symposium on Wikis and Open Collaboration*.
- Welde, J. (2011, september 19). <http://www.quora.com/How-many-websites-percentage-or-absolute-numbers-are-not-in-English>. Retrieved June 2, 2014, from <http://www.quora.com/>.
- Yimam, B. (1991). Retrieved June 10, 2014, from <http://www.ethiopians.com/bayeyima.html>.

ANNEXES

Annexes A – Amharic Article Category RDF Triples

<http://am.dbpedia.org/resource/ዋናው_ገጽ><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:ዋና> .

<http://am.dbpedia.org/resource/ሕክምና><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:ሕክምና> .

<http://am.dbpedia.org/resource/ሥነ_ፈለክ><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:ሥነ_ፈለክ> .

<http://am.dbpedia.org/resource/የኮምፒውተር:ጥናት><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:ኮምፒውተር> .

<http://am.dbpedia.org/resource/ኃይሌ_ገብረ_ሥላሴ><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:የኢትዮጵያ_አትሌቶች> .

<http://am.dbpedia.org/resource/ዓመት_በዓላት_እና_ታሪካዊ_ማስታወሻዎች><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:ታሪክ> .

<http://am.dbpedia.org/resource/ራስ_መኮንን><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:የኢትዮጵያ_ሰዎች> .

<http://am.dbpedia.org/resource/ራስ_መኮንን><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:ኢትዮጵያ_በ19ኛው_ክፍለ_ዘመን> .

<http://am.dbpedia.org/resource/ሚስቶች_በኖህ_መርከብ_ላይ><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:መጽሐፍ_ቅዱስ> .

<http://am.dbpedia.org/resource/ሚስቶች_በኖህ_መርከብ_ላይ><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:አፈ_ታሪክ> .

<http://am.dbpedia.org/resource/የቋንቋ_ጥናት><http://purl.org/dc/terms/subject>

<http://am.dbpedia.org/resource/መደብ:ሳይንስ> .

Annexes B – Amharic Article Template RDF Triples

<http://am.dbpedia.org/resource/ሕክምና><http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:መዋቅር-ሳይንስ> .

<http://am.dbpedia.org/resource/ሥነ_ፈለክ><http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:መዋቅር-ሳይንስ> .

<http://am.dbpedia.org/resource/ሥነ_ፈለክ><http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:Link_FA> .

<http://am.dbpedia.org/resource/ሥነ_ፈለክ><http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:Link_GA> .

<http://am.dbpedia.org/resource/የኮምፒውተር:ጥናት><http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:መዋቅር-ሳይንስ> .

<http://am.dbpedia.org/resource/መለጠፊያ:Wikipediasister> <http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:Image_label_begin> .

<http://am.dbpedia.org/resource/መለጠፊያ:Wikipediasister> <http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:Image_label> .

<http://am.dbpedia.org/resource/መለጠፊያ:Wikipediasister> <http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:Image_label_end> .

.

<http://am.dbpedia.org/resource/ኃይሌ_ገብረ_ሥላሴ><http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:ዋቢ> .

<http://am.dbpedia.org/resource/ኃይሌ_ገብረ_ሥላሴ>

<http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:መዋቅር-ስዎች> .

<http://am.dbpedia.org/resource/ኃይሌ_ገብረ_ሥላሴ>

<http://am.dbpedia.org/property/wikiPageUsesTemplate>

<http://am.dbpedia.org/resource/መለጠፊያ:Link_FA> .

Annexes C – Amharic Category Label RDF Triples

<http://am.dbpedia.org/resource/መደብ:አሪዞና_ካውንቲዎች> <http://www.w3.org/2000/01/rdf-schema#label> "አሪዞና ካውንቲዎች"@am .

<http://am.dbpedia.org/resource/መደብ:አሪዞና> <http://www.w3.org/2000/01/rdf-schema#label> "አሪዞና"@am .

<http://am.dbpedia.org/resource/መደብ:አሪዞና_ከተማዎች> <http://www.w3.org/2000/01/rdf-schema#label> "አሪዞና ከተማዎች"@am .

<http://am.dbpedia.org/resource/መደብ:ባንክ_ቋንቋዎች> <http://www.w3.org/2000/01/rdf-schema#label> "ባንክ ቋንቋዎች"@am .

<http://am.dbpedia.org/resource/መደብ:ሰው_ሠራሽ_ቋንቋዎች> <http://www.w3.org/2000/01/rdf-schema#label> "ሰው ሠራሽ ቋንቋዎች"@am .

<http://am.dbpedia.org/resource/መደብ:አውስትሮኔዚያን_ቋንቋዎች> <http://www.w3.org/2000/01/rdf-schema#label> "አውስትሮኔዚያን ቋንቋዎች"@am .

<http://am.dbpedia.org/resource/መደብ:ሮማንስ_ቋንቋዎች> <http://www.w3.org/2000/01/rdf-schema#label> "ሮማንስ ቋንቋዎች"@am .

<http://am.dbpedia.org/resource/መደብ:ፊኖ-ኡግሪክ_ቋንቋዎች> <http://www.w3.org/2000/01/rdf-schema#label> "ፊኖ-ኡግሪክ ቋንቋዎች"@am .

<http://am.dbpedia.org/resource/መደብ:ህንዳዊ-አውሮፓዊ_ቋንቋዎች> <http://www.w3.org/2000/01/rdf-schema#label> "ህንዳዊ-አውሮፓዊ ቋንቋዎች"@am .

<http://am.dbpedia.org/resource/መደብ:ኮምፒውተር> <http://www.w3.org/2000/01/rdf-schema#label> "ኮምፒውተር"@am .

<http://am.dbpedia.org/resource/መደብ:User_en-3> <http://www.w3.org/2000/01/rdf-schema#label> "User en-3"@am .

<http://am.dbpedia.org/resource/መደብ:የአሜሪካ_ተባባሪ_ክፍላገሮች> <http://www.w3.org/2000/01/rdf-schema#label> "የአሜሪካ ተባባሪ ክፍላገሮች"@am .

Annexes D – Amharic External Link RDF Triples

<http://am.dbpedia.org/resource/አሪዞና> <http://dbpedia.org/ontology/wikiPageExternalLink> <http://www.az.gov> .
<http://am.dbpedia.org/resource/ኮቻይስ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.co.cochise.az.us> .

<http://am.dbpedia.org/resource/ኮኮኒኖ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://co.coconino.az.us> .

<http://am.dbpedia.org/resource/ላ_ፓዝ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.co.la-paz.az.us> .

<http://am.dbpedia.org/resource/ሒላ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.co.gila.az.us> .

<http://am.dbpedia.org/resource/ግራህም_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.graham.az.gov> .

<http://am.dbpedia.org/resource/ግሪንሊ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.co.greenlee.az.us> .

<http://am.dbpedia.org/resource/ማሪኮፓ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.maricopa.gov> .

<http://am.dbpedia.org/resource/ናቫሖ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.co.navajo.az.us/> .

<http://am.dbpedia.org/resource/ሞሃቪ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.co.mohave.az.us> .

<http://am.dbpedia.org/resource/ፒማ_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.pima.gov> .

<http://am.dbpedia.org/resource/ፒናል_ካውንቲ፥_አሪዞና><http://dbpedia.org/ontology/wikiPageExternalLink>
<http://www.co.pinal.az.us> .

Annexes E – Amharic Geo-coordinate RDF Triples

<http://am.dbpedia.org/resource/ጋምቤላ_(ከተማ)><http://www.w3.org/2003/01/geo/wgs84_pos#lat>
"8.25"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://am.dbpedia.org/resource/አድዋ><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing> .

<http://am.dbpedia.org/resource/አድዋ><http://www.w3.org/2003/01/geo/wgs84_pos#lat>
"14.166666666666666"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://am.dbpedia.org/resource/አላማጣ><http://www.w3.org/2003/01/geo/wgs84_pos#long>
"39.55"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://am.dbpedia.org/resource/አላማጣ><http://www.georss.org/georss/point>"12.4
39.55"^^<http://www.w3.org/2001/XMLSchema#string> .

<http://am.dbpedia.org/resource/እንዳስላሴ><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing>

<http://am.dbpedia.org/resource/እንዳስላሴ><http://www.w3.org/2003/01/geo/wgs84_pos#lat>
"13.866666666666667"^^<http://www.w3.org/2001/XMLSchema#float> .

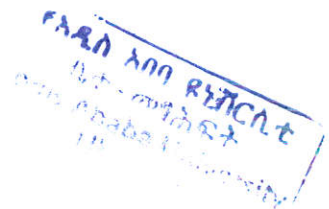
<http://am.dbpedia.org/resource/ሻሽመኔ><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing> .

<http://am.dbpedia.org/resource/ሶዶ><http://www.w3.org/2003/01/geo/wgs84_pos#long>
"35.81666666666667"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://am.dbpedia.org/resource/ሶዶ><http://www.georss.org/georss/point> "8.583333333333334
35.81666666666667"^^<http://www.w3.org/2001/XMLSchema#string> .

<http://am.dbpedia.org/resource/አርባ_ምንጭ><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing> .

<http://am.dbpedia.org/resource/አርባ_ምንጭ><http://www.w3.org/2003/01/geo/wgs84_pos#lat>
"6.033333333333333"^^<http://www.w3.org/2001/XMLSchema#float> .



Annexes F – Amharic Infobox Property RDF Triples

<http://am.dbpedia.org/resource/ቀዳማዊ_ኃይለ_ሥላሴ> <http://am.dbpedia.org/property/ተከታይ> "ዓፄ : አምሃ : ሥላሴ"@am .

<http://am.dbpedia.org/resource/ቀዳማዊ_ኃይለ_ሥላሴ><http://am.dbpedia.org/property/ባለቤት>
<http://am.dbpedia.org/resource/መነን_አስፋው> .

<http://am.dbpedia.org/resource/ሱዳን> <http://am.dbpedia.org/property/ስም> "ሱዳን"@am .

<http://am.dbpedia.org/resource/ሱዳን> <http://am.dbpedia.org/property/መጠኑም> "السودان جمهورية"@am .

<http://am.dbpedia.org/resource/ሱዳን> <http://am.dbpedia.org/property/መዝሙር> "File:Sudan.ogg"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/ብሔራዊቋንቋ> "አማርኛ (ሌሎች : ቋንቋዎች : ክልላዊ : ዕውቅና : አላቸው)"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የመንግስትአይነት> "ፌዴራላዊ : ሪፐብሊክ"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ><http://am.dbpedia.org/property/የመሪዎችማዕረግ>
<http://am.dbpedia.org/resource/ፕሬዝዳንት> .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የመሪዎችማዕረግ> "ጠቅላይ : ሚኒስትር"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ> <http://am.dbpedia.org/property/የመሪዎችስም> "ዶ/ር ሙሉቱ : ተሾመ"@am .

<http://am.dbpedia.org/resource/ኢትዮጵያ><http://am.dbpedia.org/property/የመሪዎችስም>
<http://am.dbpedia.org/resource/ኃይለማርያም_ደሳለኝ> .

http://am.dbpedia.org/resource/ሜሪላንድ> <http://am.dbpedia.org/property/ስቴትስም> "ሜሪላንድ"@am .

<http://am.dbpedia.org/resource/ሜሪላንድ> <http://am.dbpedia.org/property/ስቴትባንዲራ> "Flag of Maryland.svg"@am

<http://am.dbpedia.org/resource/ናይጄሪያ> <http://am.dbpedia.org/property/ታሪካዊከስተዮች> "የሰሜንና ደቡብ ናይጄሪያ ውህደት"@am .

Annexes G – Amharic Infobox Property Definition RDF Triples

<http://am.dbpedia.org/property/color> <http://www.w3.org/2000/01/rdf-schema#label> "color"@am .
<http://am.dbpedia.org/property/bt><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property> .

<http://am.dbpedia.org/property/bt> <http://www.w3.org/2000/01/rdf-schema#label> "bt"@am .

<http://am.dbpedia.org/property/ps><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property> .

<http://am.dbpedia.org/property/ps> <http://www.w3.org/2000/01/rdf-schema#label> "ps"@am .

<http://am.dbpedia.org/property/sel><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property> .

<http://am.dbpedia.org/property/sel> <http://www.w3.org/2000/01/rdf-schema#label> "sel"@am .

<http://am.dbpedia.org/property/ስፖ><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property> .

<http://am.dbpedia.org/property/ስፖ> <http://www.w3.org/2000/01/rdf-schema#label> "ስፖ"@am .

<http://am.dbpedia.org/property/ርዕስ><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property> .

<http://am.dbpedia.org/property/ርዕስ> <http://www.w3.org/2000/01/rdf-schema#label> "ርዕስ"@am .

<http://am.dbpedia.org/property/ስዕል><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property> .

<http://am.dbpedia.org/property/ስዕል> <http://www.w3.org/2000/01/rdf-schema#label> "ስዕል"@am .

<http://am.dbpedia.org/property/የስዕል መግለጫ><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property> .

<http://am.dbpedia.org/property/የስዕል መግለጫ> <http://www.w3.org/2000/01/rdf-schema#label> "የስዕል መግለጫ"@am .

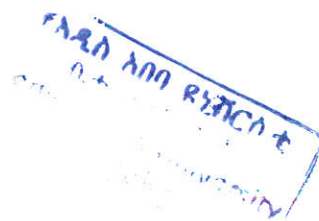
Annexes H – Amharic Inter-language Link RDF Triples

<http://am.dbpedia.org/resource/መለጠፊያ:User_en> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://ak.dbpedia.org/resource/Nhwesode:User_en-N> .
 <http://am.dbpedia.org/resource/ኮፑ_አባይ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://de.dbpedia.org/resource/Nil> .
 <http://am.dbpedia.org/resource/መለጠፊያ:መንታ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://af.dbpedia.org/resource/Sjabloon:Dubbelsinnig> .
 <http://am.dbpedia.org/resource/መለጠፊያ:መንታ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://ur.dbpedia.org/resource/سندھ:Disambig> .
 <http://am.dbpedia.org/resource/ዓለም_ንግድ_ሕንጻ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://fi.dbpedia.org/resource/World_Trade_Center> .
 <http://am.dbpedia.org/resource/ዓለም_ንግድ_ሕንጻ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://kk.dbpedia.org/resource/Дүниежүзілік_сауда_үйі> .
 <http://am.dbpedia.org/resource/ሚስኪቶ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://tr.dbpedia.org/resource/Miskitolar> .
 <http://am.dbpedia.org/resource/1926_እ.ኤ.አ.> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://nv.dbpedia.org/resource/1901_-_1950> .
 <http://am.dbpedia.org/resource/መለጠፊያ:ለማጥፋት> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://ab.dbpedia.org/resource/Ашаблон:Db> .
 <http://am.dbpedia.org/resource/መለጠፊያ:ለማጥፋት> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://eo.dbpedia.org/resourceŜablono:Forigu> .
 <http://am.dbpedia.org/resource/መለጠፊያ:ለማጥፋት> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://mzn.dbpedia.org/resource/سرید ع حذف: شاب لون> .
 <http://am.dbpedia.org/resource/መለጠፊያ:Commons> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://commons.dbpedia.org/resource/%7B%7B%7B1> .
 <http://am.dbpedia.org/resource/ኒው_ዴሊ> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://kk.dbpedia.org/resource/Нью-Дели> .
 <http://am.dbpedia.org/resource/ኑህራን> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://ml.dbpedia.org/resource/മടഹ്ററററ> .
 <http://am.dbpedia.org/resource/ቤይሩት> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 <http://pa.dbpedia.org/resource/ਬੈਰੂਤ> .
 <http://am.dbpedia.org/resource/የዓለም_የህዝብ_ብዛት> <http://dbpedia.org/ontology/wikiPageInterLanguageLink>
 http://id.dbpedia.org/resource/Penduduk

Annexes I - Amharic Infobox Templates

Person Infobox

```
{{የግንዛቤ መረጃ
| ስም =
| ርዕስ =
| ስዕል =
| የስዕል_መግለጫ =
| ግዛት =
| በዓለ_ግንዛቤ =
| ቀዳሚ =
| ተከታይ =
| ባለቤት =
| ልጆች =
| መጠ_ስም =
| ሥርወ-መንግሥት =
| አባት =
| እናት =
| የተወለዱት =
| የሞቱት =
| የተቀበሩት =
| ሀይማኖት =
}}
```



Country Infobox

{{የሀገር መረጃ
 |ስም =
 |ሙሉ_ስም =
 |ማኅተም_ሥዕል =
 |ባንዲራ_ሥዕል =
 |ባንዲራ_ስፋት =
 |መዝሙር =
 |ክርታ_ሥዕል =
 |ክርታ_መግለጫ_ፅሁፍ =
 |ዋና_ከተማ =
 |ብሔራዊ_ቋንቋ =
 |የመንግስት_አይነት =
 |የመሪዎች_ማዕረግ =
 |የመሪዎች_ስም =
 |ታሪካዊ_ቀናት =
 |ታሪካዊ_ክስተቶች =
 |የመሬት_ስፋት =
 |የመሬት_ስፋት_ከዓለም =
 |ውሀ_ከመቶ =
 |የሕዝብ_ብዛት_ግምት_ዓመት =
 |የሕዝብ_ብዛት_ቆጠራ_ዓመት =
 |የሕዝብ_ብዛት_ግምት =
 |የሕዝብ_ብዛት_ቆጠራ =
 |የገንዘብ_ስም =
 |ሰዓት_ክልል =
 |የስልክ_መግቢያ =
 |ከፍተኛ_ደረጃ_ከባቢ =
 |የግርጌ_ማስታወሻ =
 }}

Place Infobox

```
{{የቦታ መረጃ
|ስም
|ኗሪ_ስም
|ካርታ_አገር
|ክፍፍል_ዓይነት2
|ክፍፍል_ስም2
|lat_deg
|lat_min
|north_south
|lon_deg
|lon_min
|east_west
|ክፍታ
|ሕዝብ_ጠቅላላ
|ስዕል
|ስዕል_መግለጫ
}}
```

Organization Infobox

{{ድርጅት መረጃ|

ስም =

ሙሉ_ስም =

ማኅተም_ሥዕል =

ባንዲራ_ሥዕል =

ካርታ_ሥዕል =

ካርታ_መግለጫ_ፅሁፍ =

የጽ/ቤት_መቀመጫ =

የስራ_ቋንቋ =

አባል_ሀገራት =

የመሪዎች_ማዕረግ =

የመሪዎች_ስም =

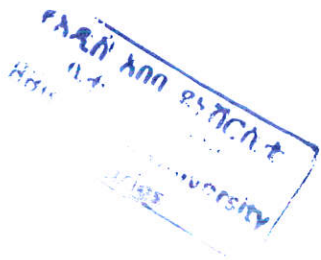
የምስረታ_ቀን =

ድህረ_ገፅ =

}}

Annexes J – List of Updated Infobox Property Values

Updated Values	Previous values
የትውልድ ዘመን	የትውልድ ቀን
ሙሉ ስም	ስም
የስዕል መግለጫ	የምስል መግለጫ
የመሬት ስፋት	የቦታ ስፋት
አሰልጣኝ	ማናጀር



Annexes K – Pom.xml File Customization

```
<args>

    <!-- base folder of downloaded dumps -->
    <arg>/home/binyam/Desktop/dbpedia/extraction-framework/dump/dumps</arg>
    <!-- location of SQL file containing MediaWiki table definitions -->
    <arg>/var/www/mediawiki/maintenance/tables.sql</arg>
    <!-- JDBC URL of MySQL server. Import creates a new database for each wiki. -->
    <arg>jdbc:mysql://localhost:3306/?characterEncoding=UTF-
8&user=root&password=bini</arg>
    <!-- require-download-complete -->
    <arg>true</arg>
    <!-- file name: pages-articles.xml{,.bz2,.gz} -->
    <arg>pages-articles.xml.bz2</arg>
    <!-- languages and article count ranges, comma-separated, e.g. "de,en" or "@mappings" etc. -->
    <arg>am</arg>
</args>
```