Semantic Web-Based Digital Business Ecosystem: The Case of Amharic Semantic Blogging

Eyoss Girma
MFI Trading and Office Solutions/Alcatel-Lucent Department, Addis Ababa, Ethiopia
eyosg@yahoo.com

Workshet Lamenew
School of Information Science, Addis Ababa University, Ethiopia
workshet@gmail.com

Abstract

While few researches are conducted elsewhere, semantic blogging specific to a local language such as Amharic was not given an attention. This research is aimed at investigating the current situation of business ecosystem in Ethiopia with the view to design and develop Amharic semantic web-based blogging system in web-based DBE.

Various methodologies are used to gather primary and secondary data from IT-based SMEs like profiles, events, and items. Classes, subclasses, and the relationships between concepts are defined and an architecture of semantic blog searching for IT-based SMEs is developed.

The proposed semantic blog searching architecture consists of four modules: preprocessing, query processor, knowledgebase, and crawler. The preprocessing module handles most of the Amharic unique features. The query processor module lets the user enter queries in Amharic and retrieve the relevant result for the given query. The knowledgebase module handles the ontological structures of the business of concepts. The crawler component traverses through the World Wide Web and fetches those web page blogs with Amharic content.

The query and retrieval of the semantic blogging system is measured in terms of precision and recall and the summary of the results are presented. The study shows that the use of Amharic semantic blog searcher results in 76 % accuracy which is a promising outcome.

Keywords: Digital Business Ecosystem; Small and Medium Enterprise; Semantics; Ontology

1. Introduction

Semantic web services are designed to create easy communication between computers and humans so that they can accomplish certain tasks and services. These services are often used in digital environments - so called Digital Business Ecosystems to implement and organize services of businesses like SME's (Small and Medium Enterprises) interaction for demand and supply.

SMEs can lead to effective knowledge creation and growth. Relevant information at the right time could provide SMEs with the appropriate tools to make more economically sound decisions. This is the main objective of the paper: "To investigate the current situation of IT-based SMEs business ecosystem in Ethiopia with the view to design and develop architecture for Amharic semantic web-based blogging system in web-based DBE".

The rest of the paper is organized as follows. Section 2 covers literature review which includes the description of SMEs, semantic web-based digital business ecosystem and its major components. Section 3 covers the review of related works. Section 4 covers the architectural design and implementation of the proposed architecture for semantic blogging system and Amharic blog searching system in digital business ecosystem. Section 5 presents the experimental result of the proposed design for semantic blogging and searching in digital business ecosystem for IT-based SMEs. Finally Section 6 presents the result and conclusion.

2. Background

The success of an SME in the Business Ecosystem depends on its ability to find the right partners in the Business Ecosystem, to operate across the Business Ecosystem. SMEs are increasingly adopting Internet technologies and are recognizing the opportunities offered by e-Commerce and use it to offer their products and services online.

The Digital Business Ecosystem is a special Internet-based environment in which businesses can interact with each other in effective and efficient ways. Being part of the Digital Business Ecosystem means that we are aware of the range of products and services available from all of the other partners and can easily match them with our business requirements.

The exposure is twofold which means our products and services are also being showcased to other SMEs so they can identify us as potential business partners. The Digital Business Ecosystem brings together the best of all of the Internet technologies, tools and applications as well as legal, business and revenue information that SMEs need if they are to have the competitive edge in the market.

Blogging's greatest benefit is in strengthening the social and business relationships in social and DBE's environment of the world society. First, ease of use makes it likely that more people will publish and publish more often, and that more information will be communicated. The structure of the information is often different from more static home pages, more like online journals.

There are around 80 languages in Ethiopia which are used in day-to-day communication. Although many languages are spoken in Ethiopia, Amharic is the dominant one since it is spoken by the substantial part of the Ethiopian population. Based on Tessema Mindaye [7], Amharic language is the working language of the Federal government and it is a Semitic Language of the Afro-Asiatic language group that is related to Hebrew, Arabic, and Syrian.

The unique feature of Amharic language needs to be handled properly to be used in the DBEs. Handling these features is mandatory so that the input texts are processed properly before they are manipulated in different steps of natural language processing and stored in the index or retrieved from a database.

3. Related Work

This section covers the various researches and projects done in the areas of digital business ecosystem [3], web based semantics and semantic blogging systems [2, 5], Amharic language search engine [7], and document categorization in Amharic language [6].

Batra and Salzburg [2] made some development in the area of ontology systems, specifically in the field of medicine which refers to drugs, practice or diagnosis. Chang and Boley [3] made a comparison of different forms of DBE with most advanced communication platforms or environments such as client-server, P2P and web services. They also described how digital ecosystems can benefit from semantic web ontologies. When we come to the language of Amharic, Surafel Lemma [4] worked on semantics in the area of multimedia and the respective web services and prepared a prototype which allows a user to advertise a web service using the ontology and look for a specific Web service. Meron Sahlemariam [6] has also attempted to look into the techniques of automatic classification and the study focuses in categorizing a given Amharic document into a predefined category by passing it through the preprocessing and classification processes.

Blogging is one component of digital ecosystem and there are multiple researches made in this area and one of them is by Cayzer [5] with the focus on semantic blogging and decentralized knowledge management which develops architecture for importing, export, view, navigate and query the ontology repository. This architecture is the starting point for our digital business ecosystem architecture. The author also described the impossibility of future blogging without semantics.

A project called Digital Ecosystem for Agriculture and Rural Livelihood (DEAL) in [1] was an implementation of web based initiative that coordinates back-end infrastructure, media technology, and knowledge in order to make agricultural content accessible through multiple channels in rural India.

The development of DBE throughout EU and Asia is aggressive as compared to that of Africa. As seen from the related works, its value for the development of SME in the developing countries is priceless.

4. The Proposed Solution

4.1 Architecture of Semantic Blogging

The major goal of this research work is to make use of the content related semantic metadata which are relevant in the DBE environment of blogging and specially in Amharic language. Figure 1 shows the architecture of the proposed semantic blogging web-based system. The semantic view is the part of the architecture which is used to view the blog content in the way that helps the semantically enriched blog metadata to the user.

The semantic navigation part of the architecture helps to efficiently find blog items of our interest. For ease of navigation, explorer or tree type interface is used to browse related items and users can also follow labeled links.

One part of this architecture is semantic query. We used SPARQL as a language to test the retrieval and manipulation of the data stored in OWL or RDF format.

In our consideration of import and export for a blog, we took the implementation through RSS (Really Simple Syndication) with the help from W3C schools RSS standards [8], which is used to distribute up-to-date web content from one web site to thousands of other web sites around the world. The advantage is that it will reduce the time spent on sites looking for news, events, and updates.

The option of using RSS as a file to import and export the blog is possible through companies called aggregators which need registration and this searches the registered web sites for RSS documents. The aggregator also verifies the link and displays information about the feeds so clients can link to documents that interest them.

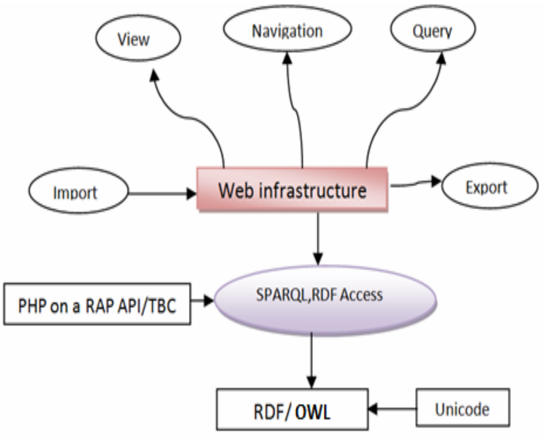


Figure 1: Revised Architecture of Semantic Blogging

4.2 Architecture of semantic Amharic blog Searching in Digital Business Ecosystem

Blog searching system architecture of a specific language requires a crawler, preprocessing, query analyzer and indexing as basic components. This session will cover the part which explains the architecture of Amharic blog searching and display.

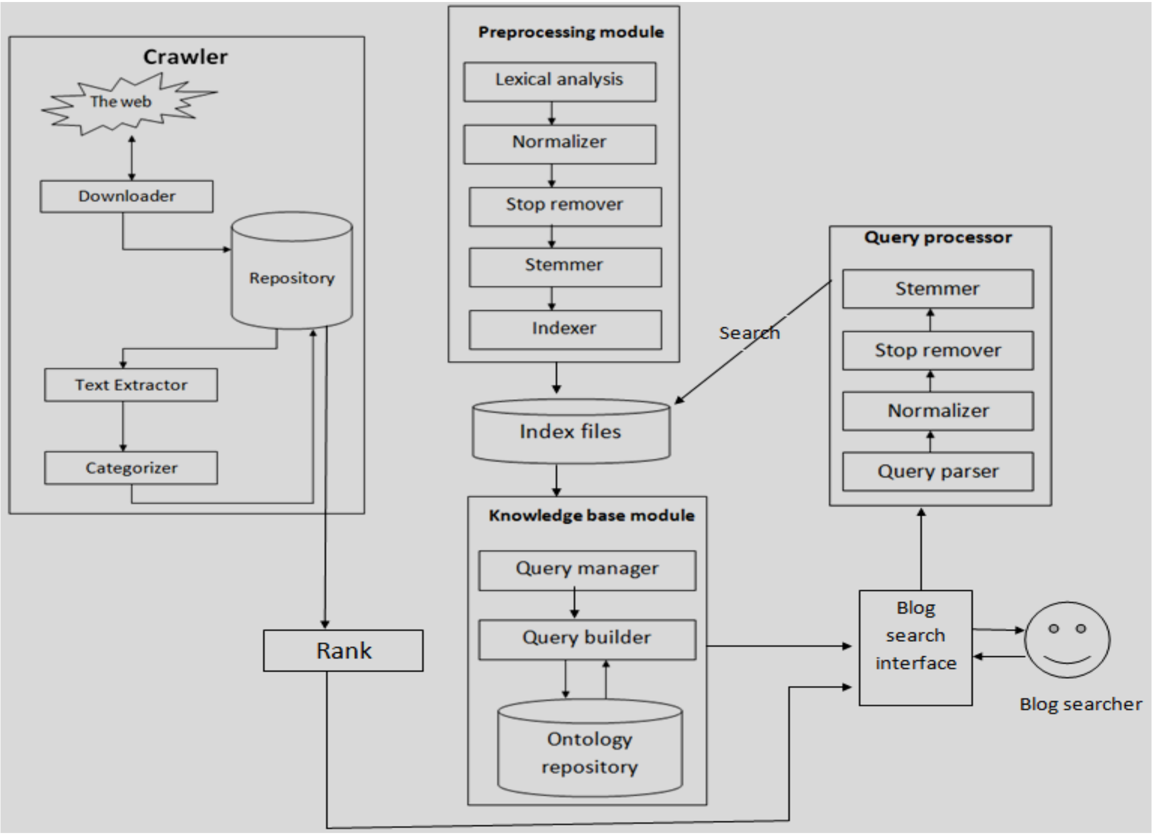


Figure 2: Semantic Amharic Blog Searching Architecture

The Semantic Amharic Blog Searching Architecture is shown in Figure 2. The Preprocessing module handles most of the Amharic language unique features. The input texts from blog users and the crawler are passed through different processes before they are stored in the index. The character streams are tokenized into words by taking white space and Amharic punctuation marks as word demarcations. Shorter forms of a word are replaced by its expanded form. Then, the word is checked for stop words. If the word is a stop word or if it is a variant of a stop word, it is removed.

The remaining non-stop words are stemmed to reduce them to their common form. The final result of all these processes is stored as an index in a structure that is appropriate for fast searching. The process of preprocessing module includes tokenization handling the repetitive alphabets, short words and stemming.

The tokeniser splits the text into very simple tokens such as numbers, punctuation and words of different types. In Amharic language, there are multiple punctuation marks that are used in between and at the end of sentences. For instance, colon ":" in English is called "Hulet Netib" in Amharic which is used to separate words. The other mark is ፣ "Netela Serez" which is used as a coma in English language where in case of Amharic language it demarcates a list of things, items, etc. Two colons side by side "::" "Arat Netib" is used at the end of a sentence and demarcates the end of a sentence. This tokenization component and the algorithm are adopted from Tesemma's work [7] which is designed to split words by identifying punctuation marks and white spaces.

The Amharic indexer has two inputs. One of the inputs is for the crawler component which passes through the normalizer, stop word remover, stemmer, and indexer. The other is for the input text form the blog searcher which passes through similar components.

The index repository is the component that stores (indexes) a word (term) with some necessary information such as Term Frequency (TF), Inverse Document Frequency (IDF), etc. We also used Lucene for this purpose and Lucene has an open source IR library which plays a pivotal role in the indexing process. One of the tasks in indexing component of our blog search is to integrate the

Amharic language features with that of Lucene library with the application of GATE tool.

The query processor module accepts input queries from users through an interface. This module lets the user enter queries in Amharic and retrieve the relevant documents for the given query.

Lucene has an important role to play in this module. One of the tasks in the indexing component of our blog searching is to integrate the Amharic language features with that of Lucene library. GATE has a data store to store language resources. Data stores are an abstract model of disk-based persistence, which can be implemented by various types of storage mechanisms. Lucene's data store is a full-featured annotation indexing and retrieval system. It is provided as part of an extension of the serial data stores.

The knowledgebase module serves as a knowledgebase to define the ontological structures of the business of SMEs. Knowledge representation in the ontology is mainly based on the concepts. The concepts are structured in the ontology as class, sub-classes, and relationships between classes. The ontology is like a database which constitutes the concepts, words, and categories and how they relate to each other.

The ontology contains the represented knowledge in the domain and this is designed in a way to access and use the knowledgebase in the ontology. We, therefore, mapped the concepts and the relationships between them with textual expressions. Using this relationship, we can query the semantic relations found among concepts.

The crawler is directly adopted from Tesemma's work [7]. The advantage of adopting the work is that the author developed a language-focused crawler which has the capacity to traverse through a country domain (ccTLD) which is a way to get web pages of a certain language. The crawler component traverses through the World Wide Web and fetches those web page blogs with Amharic content. The crawler has two sub components: the Crawler and Categorizer. The crawler is responsible for downloading a page from a web site and the categorizer will categorize the web as Amharic blogs or not. The Amharic web blogs are processed and the rest are discarded.

5. Prototype

GATE tool is used to customize the crawler. It is based on Websphinx, a JAVA-based, customizable, multi-threaded web crawler, which can be customized based on our requirement.

The basic idea is to specify a source URL of a blog web site (or set of documents created from web URLs) and a depth and maximum number of documents to build the initial data upon which further processing could be done. The PR itself provides a number of other parameters to regulate the crawling. In our experiment, we used blog pages that we developed through Dream Weaver and we also used Wamp server as our web server.

There are no well organized blogging sites in Amharic language so we needed to include blog web construction by our own. The blog web construction is based on the data we collected from web sites written in Amharic language like www. ethiopianreporter.com/, http://www. waltainfo.com/, and www.mcit.gov.eg/.

The knowledgebase is constructed based on the documents collected from IT-based SMEs with a focus on the organizational structure, services, and products.

Based on the above example, we have made our experiment on 21 blogs with a total of 3014 words and 26 sentences. Test was made for 8 queries.

5.1 Evaluation

Evaluation of the blog searching is done with the evaluation parameter that compares the number of web blogs which are classified correctly and incorrectly. Typically, the comparison is done based on the blogs which are retrieved based on the search term correctly and incorrectly.

Precision and recall, which are the evaluation parameters of Information Retrieval, are used in web based blog searching. In pattern recognition and information retrieval, precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are, therefore, based on an understanding and measure of relevance. When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is 20/30 = 2/3 while its recall is 20/60 = 1/3.

Precision =Tot-Ret-Rel ÷ Tot-Ret

Recall = Tot-Ret-Rel ÷ Tot-Rel

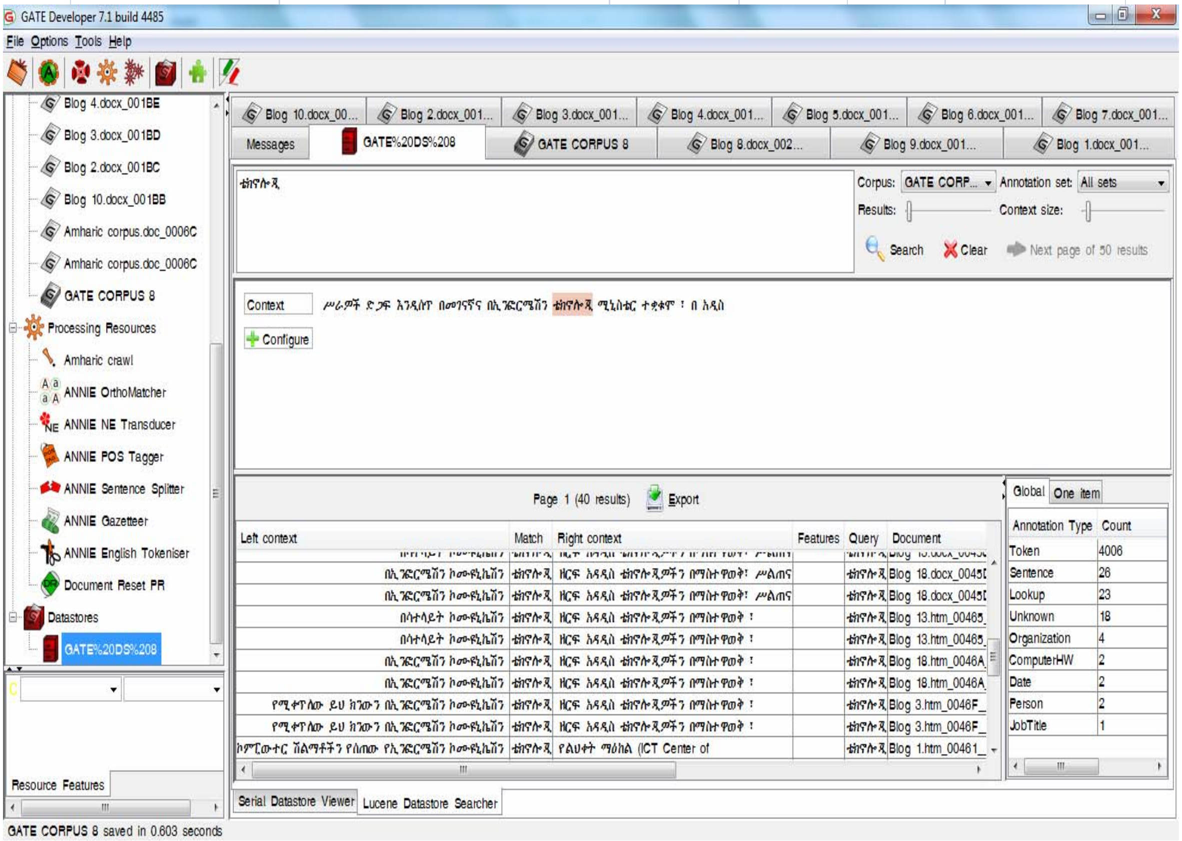
where Tot-Ret-Rel is the total number of returned documents that are relevant to the query, Tot-Ret is the total number of documents that are retrieved by the system, and Tot-Rel is the total number of documents that are relevant to the query

5.2 Result

The experiments are carried out on the Amharic web blog documents collected from the crawler and we used Amharic search terms and recorded the result of the retrieval.

Table 1: Blog Search Result

Query	Tot-Rel	Tot-Ret	Tot-Ret-Rel	Tot-Ret-Ire	Precision	Recall	Accuracy Percentage %
ኢንፎርሜሽን	5	5	5	0	1	1	100
ኢንፎርሜሽን ኮሙዩኒኬሽን ቴክኖሎጂ	4	7	2	5	0.28	0.5	58
ሞባይል ስልክ	4	9	3	6	0.33	0.75	60
ሶፍትዌር መፍጠር	1	0	0	0	0	0	0
ሳተላይት	5	4	4	0	1	0.8	100
አዲስ አበባ ዩኒቨርሲቲ	6	6	5	1	0.84	0.84	85
ኤሌክትሮኒክስ	2	2	2	0	1	1	100
አይሲቲ ልቀት ማዕከል	3	5	3	2	0.6	1	71
ቴክኖሎጂ	12	14	9	5	0.64	0.75	73
ፀረ ኮምፒውተር ቫይረስ	5	5	5	0	1	1	100
አፍሪካ	10	11	9	2	0.81	0.9	85
Average							75.63636364



ቴክኖሎጂ

Corpus: GATE CORP... Annotation set: All sets

Results: Context size:

Context ሥራዎች ድጋፍ እንዲሰጥ በመገናኛና በኢንፎርሜሽን ቴክኖሎጂ ሚኒስቴር ተቋቁሞ ፣ በ አዲስ

Page 1 (40 results)

Left context	Match	Right context	Features	Query	Document
በኢንፎርሜሽን ኮሙዩኒኬሽን	ቴክኖሎጂ	ዘርፍ አዳዲስ ቴክኖሎጂዎችን በማስተዋወቅ፣ ሥልጠና		ቴክኖሎጂ	Blog 18.docx_0045D
በኢንፎርሜሽን ኮሙዩኒኬሽን	ቴክኖሎጂ	ዘርፍ አዳዲስ ቴክኖሎጂዎችን በማስተዋወቅ፣ ሥልጠና		ቴክኖሎጂ	Blog 18.docx_0045D
በሳተላይት ኮሙዩኒኬሽን	ቴክኖሎጂ	ዘርፍ አዳዲስ ቴክኖሎጂዎችን በማስተዋወቅ ፣		ቴክኖሎጂ	Blog 13.htm_00465_
በሳተላይት ኮሙዩኒኬሽን	ቴክኖሎጂ	ዘርፍ አዳዲስ ቴክኖሎጂዎችን በማስተዋወቅ ፣		ቴክኖሎጂ	Blog 13.htm_00465_
በኢንፎርሜሽን ኮሙዩኒኬሽን	ቴክኖሎጂ	ዘርፍ አዳዲስ ቴክኖሎጂዎችን በማስተዋወቅ ፣		ቴክኖሎጂ	Blog 18.htm_0046A_
በኢንፎርሜሽን ኮሙዩኒኬሽን	ቴክኖሎጂ	ዘርፍ አዳዲስ ቴክኖሎጂዎችን በማስተዋወቅ ፣		ቴክኖሎጂ	Blog 18.htm_0046A_
የሚቀጥለው ይህ ክንውን በኢንፎርሜሽን ኮሙዩኒኬሽን	ቴክኖሎጂ	ዘርፍ አዳዲስ ቴክኖሎጂዎችን በማስተዋወቅ ፣		ቴክኖሎጂ	Blog 3.htm_0046F_
የሚቀጥለው ይህ ክንውን በኢንፎርሜሽን ኮሙዩኒኬሽን	ቴክኖሎጂ	ዘርፍ አዳዲስ ቴክኖሎጂዎችን በማስተዋወቅ ፣		ቴክኖሎጂ	Blog 3.htm_0046F_
ኮምፒውተር ሽልማቶችን የሰጠው የኢንፎርሜሽን ኮሙዩኒኬሽን	ቴክኖሎጂ	የልህቀት ማዕከል (ICT Center of		ቴክኖሎጂ	Blog 1.htm_00461_

Annotation Type	Count
Token	4006
Sentence	26
Lookup	23
Unknown	18
Organization	4
ComputerHW	2
Date	2
Person	2
JobTitle	1

GATE CORPUS 8 saved in 0.603 seconds

Figure 3: Query result for ቴክኖሎጂ

The experiment was made on 45 blogs with a total of 7114 words (see Table 1 and Figure 3 for a screenshot of an example query). The test was made for 11 queries. Based on result, out of 11 queries 4 of the results were found to have 100% precision and 3 of the queries were observed to have precision less than 50%. Ten of the recall values show a query result of 50% and above.

Average precision and average recall for combination of words which are associated with the "AND" operator is 51% and 68%, respectively. For single word query the average precision and recall is 89% and 89%, respectively.

6. Conclusion

A digital ecosystem uses the working principles of nature's ecosystem which will have a network of co-existing elements that depend on each other in order to survive. Every Small-to-Medium Enterprise is part of this business ecosystem working together with many other elements in the system like other SMEs and clients.

There are two architectures in our design. One of the architecture is the proposed semantic blogging web-based system with major components of import, export, view, navigation, and query. The other architecture is Blog searching system architecture for Amharic which has crawler, preprocessing, query analyzer and indexer as basic components. Amharic language has many distinct features that affect the information retrieval of the language's documents on the Web.

An experiment is conducted using GATE tool by customizing the crawler to gather the required documents. The experiment uses blog pages that we developed and we also used Wamp server as our web server. The blog searching system is found to have a better retrieval than general purpose blog search engines.

References

[1] Francesco Nachira and Andrea Nicolai, "Digital Business Ecosystems", The European Commission, 2010.

[2] Sudhir Batra and F. H. Salzburg, "Semantic Web Services in Digital Ecosystems", *Proceedings of the Digital EcoSystems and Technologies Conference*, Canada, 21-23 Feb 2007.

[3] Elizabeth Chang and Harold Boley, "Digital Ecosystems: Principles and Semantics", *in Proceedings of the 2007 Inaugural IEEE International Conference on Digital Ecosystems on Technologies*, Aug 14, 2007.

[4] Surafel Lemma, "Semantic Description of Multimedia Content Adaptation Web Services", Unpublished Master's Thesis, Addis Ababa University, 2005.

[5] Steve Cayzer, "Semantic Blogging and Decentralized Knowledge Management", *Communications of the ACM Journal*, Vol. 47, No. 12, December 2004.

[6] Meron Sahlemariam, "Concept-Based Automatic Amharic Document Categorization", Unpublished Master's Thesis, Addis Ababa University, January 2009.

[7] Tessema Mindaye, "Design and Implementation of Amharic Search Engine", Unpublished Master's Thesis, Addis Ababa University, July 2007.

[8] "RSS basic," Retrieved from http://www.w3schools.com/rss/rss_intro.asp, last accessed on Oct 04, 2012.