# A survey of Arabic language Support in Semantic web

Majdi Beseiso, Abdulrahim Ahmad, Roslan Ismail
Tenaga National University (UNITEN)
Km 7, Jalan Kajang-Puchong, 43009 Kajang, Selangor, Malaysia.
Email: majdibsaiso@yahoo.com

***Abstract*** **- Information availability is the key most important factor in considering the acquisition of knowledge and awareness. The access of information either in the general area or even in the most specific ones like sciences, languages, and religion become more elaborately wide, especially that the World Wide Web is able to use semantics.**

**Semantic Web technologies basically assist the acquiring information in a way that it can create processes that would match any available characters that would link of one information to the other. This highly digitalized result of technological advancement is dedicated to processing Latin family scripts but the studies that deal with Arabic script support in these technologies remained silent and have not been an object of discussion in many fields of study.**

**This paper, therefore, would like to account the support of Arabic in some of the existing Semantic Web technologies, and determine the ability to applying Semantic Web for Arabic applications. Aside from which, there is also an importance of establishing a multilingual support for this new technologies.**

## I. INTRODUCTION

The world today evolves as fast as we can imagine. Its direction is consistently towards the advancement and progress as long as human can possibly move towards with it. These many advances allowed greater possibilities and chances to get to acquire information and knowledge in the most basic to the complex manner.

One of the many wonders that the world was able to create was dawn of the Internet. All around the globe, mankind from the different walks of life would get to know each other and have the great access of the many possibilities of interaction. The global village, as Mcluhan would refer it, is much experienced in today's world. Aside from the establishment of this global village, access to information can also be made possible in the context of the Semantic Web.

Al- Kalifa & Al-Wabil indicated that the Semantic Web intends to improve the existing Web with a layer of machine-interpretable metadata (i.e., data about data) so that a computer program can understand what a Web page is about, and therefore draw conclusions about the Web page [1]. This Tim Berners-Lee's innovation of the web content medium that can match with the software agents allowed information to be found, shared and integrated. With the semantic web, there is a general sharing and exchange of data and with these data, the availability of knowledge can be become more accessible and easy to discover and researched.

The Semantic Web is an example of how fast technology can change. It is not just a tool for people to use but it can assist in evolution of knowledge as well. The Semantic Web, when it opens itself for public use worldwide, would be exposing new concepts and would let everyone exchange expressions. Its use of a unified logical language will enable the computer to connect the world to a universal Web. By this connections and links, humans can have access to a wide array of knowledge and ideas. In this way, all people can live together, work together, and learn together. It opens a lot of possibilities for the next generation of technology users.

That is why it is very important for the Semantic Web tools and applications to prepare themselves to support all languages such as Arabic in order to fulfil the Semantic Web goal of connecting the world into one network.

## II. SEMANTIC WEB AND ARABIC LANGUAGE

### *I. Importance of Arabic language*

The Arabic is an integral part of people living in the Middle-East. As a language that distinct them from other countries, it is a symbol that manifests their faith and perceptive. Arabic is the official language of hundreds of millions of people in twenty Middle East and northern African countries, and is

the religious language of all Muslims of various ethnicities around the world [5]. Also, Arabic is a Semitic language with 28 alphabet letters. Its writing orientation is from right-to-left. Arabic is also considered one of the six official languages of the United Nations and the mother language of more than 330 million people [6]. Prophet Muhammad started his mission when he was forty years old and continued preaching until his death at the age of sixty-three. The inspirations (revelations) which formed Muhammad's preaching discourse over the twenty-three-year period constitute the Qur'an, which means 'the recital' or the proclamation [7]. Thus, Arabic is the language for Holy Quran, a book that professes a Muslim's faith where they are also known for.

### *II. Task Difficulties*

On the contrary, the language has presented various task difficulties which might have affected the creation of such web tool. Arabic Language has many particularities like short vowels, absence of capital letters, complex morphology, etc. Again, the thesis of Saleh and Al-Khaliffa's precisely indicate that the Arabic language is composed of nouns, verbs and particles; wherein these are morphemes and derived from a closed set of around 10,000 roots [7]. Also, the same article highlighted the idea that "Arabic is highly inflectional and derivational, which makes morphological analysis a very complex task" and "Capitalization is not used in Arabic, which makes it hard to identify proper names, acronyms, and abbreviations" [7]. These presentations of query suggest the difficulties one might encounter in the idea of having an Arabic Semantic Web.

### *III. Arabic Language & Semantic web research*

There are various studies conducted by many that link the Arabic and Semantic values. A survey on the many research attempts that presented data's on Arabic language and semantic web may help more for further understanding. One example is the paper by Zaidi, Laskri and Bechkoum: The system that we propose to improve the Arabic information retrieval on the Web in the legal domain is situated in a general architecture of an Arabic search engine supporting the translation in English or French queries. The aim is to return documents written in Arabic, French or English. [8] Another paper suggested by Vossen, Pease and Fellbaum talks on the Arabic Word Net project that: AWN will be constructed according to the methods developed for EuroWordNet (EWN;Vossen 1998) and since applied to dozens of languages around the world. The EuroWordNet approach maximizes compatibility across wordnets and focuses on manual encoding of the most complicated and important concepts. The basic criteria for selecting synsets to be covered in AWN are connectivity, relevance, generality, from English to Arabic and from Arabic to English. [9] Hammo included these surveys in his study on "Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents" Aljlayl et al. (2002), built an Arabic–English IR system based on a machine translation approach. AbdulJaleel and Larkey (2003), proposed a statistical transliteration approach for Arabic–English IR.

Grefenstette et al. (2005), described the changes required to modify their cross language IR system, which has been designed for European languages to integrate Arabic language. Abdelali et al. (2006), described how precision can be improved in query expansion using LSI. Finally, Semmar and Fluhr (2007), presented a new approach to align Arabic–French sentences retrieved from a parallel corpus based on a cross-language IR system. This approach is basically based on building a database of sentences of the target text and considering each sentence of the source text as a query to that database. [10]

Guo and Ren cite that the NLP technology is one branch of the linguistics, which uses the computer technology to realize human language processing effectively. Its ultimate objective is to understand automatically human language supported by the artificial intelligence technology. Therefore, it is called as the natural language understanding sometimes is transformed to Semantic Web data. Traditional information retrieval also turns into knowledge discovery. [11] Al-Khalifa, Hen, Al-Yahya, Bahanshal and Al-Odah proposed a framework for representing a semantic opposition in the Holy Quran using Semantic Web Technologies. Their inputs include: Most previous research done in the field of Computers and the Holy Quran can be classified into six categories, namely: Information Retrieval, Speech Recognition, Optical Character Recognition, Morphology Analysis, Semantic checking and Educational Applications. Little -if any- has been conducted toward leveraging semantic web technologies for serving the lexical semantics of the Holy Quran. However, in this section we will highlight some of the latest attempts conducted in this field. [12]

Hammo, Abu-Salem and Lytinen stressed main goal of the QARAB system to identify text passages that answer a natural language question. The task can be summarized as follows: *Given a set of questions expressed in Arabic, find answers to the questions under the following assumptions:*

• The answer exists in a collection of Arabic newspaper text extracted from the Al-Raya newspaper published in Qatar.

• The answer does not span through documents (i.e. all supporting information for the answer lies in one document)

• The answer is a short passage. [13]

These are just a few studies conducted for Arabic Semantic

Based on the above information, it is to be concluded that Arabic language in the Semantic Web is still quite vague to be made possible since there is only few available Arabic semantics like those that are used in the Quran.

## IV. ARABIC ONTOLOGY

Arabic ontology is said to be the foundation of the creation of Semantic Web designs. The basic categorization of terminologies and meanings can basically give out semantics. The interrelationship between one word to the other words that matches to its meaning can also actually result to the stems and branches of semantics in the World Wide Web. The goal of ontology learning is to (semi-)automatically extract relevant concepts and relations from a given corpus or other kinds of data sets to form Ontology [11].

There are six parts in the means of creating the life cycle of ontology which are the following: Ontology Creation, Ontology Population, Ontology Validation, Ontology Deploy, Ontology Maintain and Ontology Evolve [14].

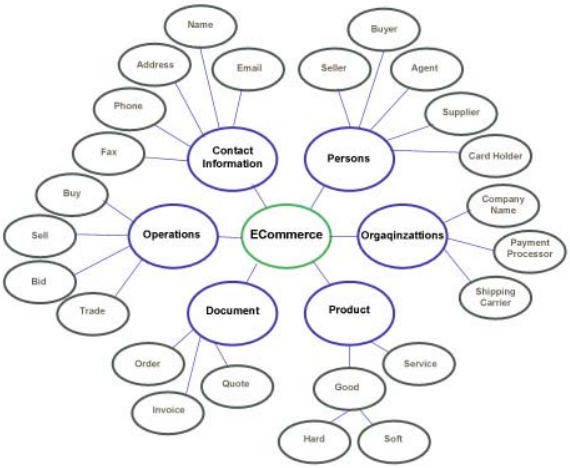
The ontology learning process into six steps can also be subdivided into Extract Term, Discover Synonyms, Obtain Concepts, Extract Concept Hierarchies, Define Relations among Concepts, Deduce Rules or Axioms. These processes are used in order to make the ontology matching become possible and that the related branches of topics would be available to any users.

Using the different languages in the study of Ontology can also be a challenge to the many attempts of the Web designs to cater the thousands of users in the World Wide Web.

Web information is usually language dependent; and the availability of information related to the language that would be much preferable according to the user would be an increasing need of today. Much is the need of the Arabic language since the ontology in English cannot be translated to Arabic.

Since there are no standard ontologies for Arabic language to apply our test for, we used our e-commerce ontology proposed in (Beseiso & Abdulrahim, 2010) with some refinement based on English e-commerce ontology proposed by Geller. Figure 1 shows the proposed ontology for e-commerce domain. Different languages have contained the specific linguistic environment and the cultural context, which has caused the need to develop different ontology for different information language.

Figure 1 : E-COMMERCE ONTOLOGY




## V. SYSTEMS TO BE EVALUATED

Ontology is one of the basic and the preliminary sources in order to start the process of Semantic Web; and in order for these to work, there are different systems used in the application of the said innovation for the information availability. Among these systems to be evaluated would include the traditional Ontology Management Systems such as Java's Protégé and Jena, Sesame, and KOAN. These systems have been designed and engineered in order for the ontology to work with the related needed information.

Protégé is basically an Ontology Visual Editor. This is a graphical ontology editor and development framework in providing the necessary manipulations and query from ontology.

Jena is another web system used to provide a programmatic environment for RDF, RDFS, OWL, SPARQL and includes a rule-based inference engine [1]. It is also a program development framework for Ontology manipulation and query [15].

Sesame is a Resource Description Framework that also allows ontology manipulation and query. This is an open-source RDF database with support for RDF Schema inference and querying [16].

KOAN, on the other hand, is also an ontology management that could create ontology aside from the manipulation, inference and query. KOAN is

basically unique as compared to the other given systems since it offers a Relational Database Management schema that would create an easier access for the availability of the OWL. Table1 shows a summary description for the four semantic tools.

**TABEL 1 : SUMMARY OD SEMANTIC WEB TOOLS**

| Tool | Creator | Functionality | Standards |
|---|---|---|---|
| Protégé | Stanford Center for Biomedical Informatics Research | Graphical ontology editor and knowledge base framework for ontology manipulation & query | RDF RDFS OWL SPARQL |
| Jena | Hewlett-Packard Development Company | Framework for ontology manipulation and query | RDF RDFS OWL SPARQL |
| Sesame | Aduna in cooperation with NLnet Foundation | Framework for storage, inferencing and querying of RDF data | RDF RDFS OWL SeRQL |
| KAON2 | Research Center for Information Technologies | Suite of ontology management (Create, Manipulate, Infer) tools | RDF RDFS OWL |

According to Pan, et.al, while these engineering ontology tools provide a stack of Ontology management support, they also show certain limitations in supporting large scale software engineering projects [15]. Therefore, the need to study and evaluate each of this given tools is an utmost importance, especially if the Ontology in the Arabic language has to be considered and proposed.

## VI. THREE DIMENSIONS FOR EVALUATING ARABIC

The Resource Description Framework Generation is the common model for the data to be made available over the web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed [17].

On the other hand, Ontology Web Language Generation, which is considered as the most effective model in terms of generating information, facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. [18]

Querying Tools such as SeRQL, OWL-QL, RDQL and SPARQL are also the needed system tools to be evaluated that would usually allow users to indicate different query for the needed information that would give out results to the given query [11].

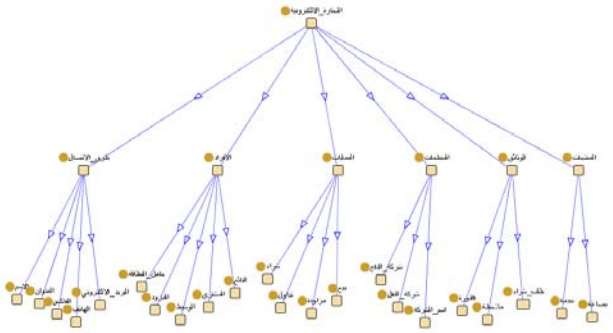
All three are related dimensions to determine as to whether these would be helpful in the coming up of the different needed information in the Arabic language.

## VII. RESULTS AND DISCUSSION

In the series of studies being conducted as to evaluate the different frameworks and the systems, the following results and discussions are presented:

Protégé can basically create & display ontology in Arabic, jambalaya plug in success to display Arabic text as shown in figure 2. Protégé is an applicable tool to build and manage conceptual terminology in ontology [19]. This system uses the RDF standard that also utilizes the UTF-8 encoding, that is compatible with null-terminated strings [19]. However, it could display numeral literal instead of Arabic characters as shown in figure 3, but when the preview of RDF/ OWL file it will be appear in Arabic as shown in figure 3. The use of the Wordnet is available only for english which would indicate that the use of lexical resources in order connectivity to be made faster and accessible is crucial. Once there is enough connectivity of the lexical resources such as the synsets, the availability of the needed resources can more likely to give our relevant results.

**Figure 2: Protégé Jambalaya**



Protégé Query tools support Arabic text query; however, without diacritics or stemming, Arabic language would not be supported unlike the processing in English text.

The Jena system can also build RDF/OWL File in Arabic as shown in figure 5. Many APIs can integrated with Jena query engine for English

language processing but nothing available to support Arabic, so we can query Jena only by exact Arabic word.

RDF Sesame, on the other hand, would use numeral literals to store Arabic characters but it is unable to read or query Arabic ontology.

**Figure 3: Protégé OWL GENERATED**

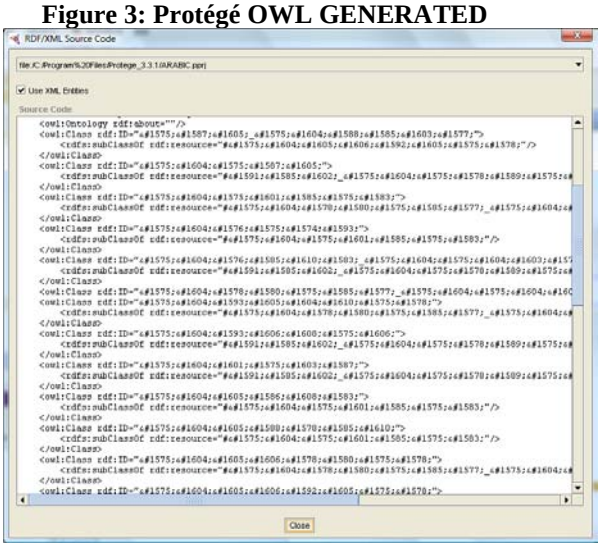


**Figure 4: OWL Preview**

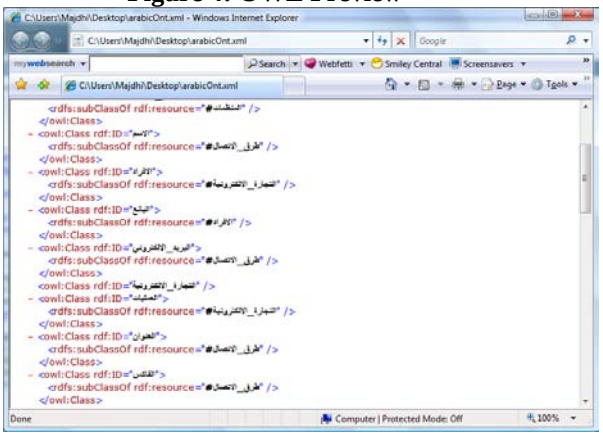
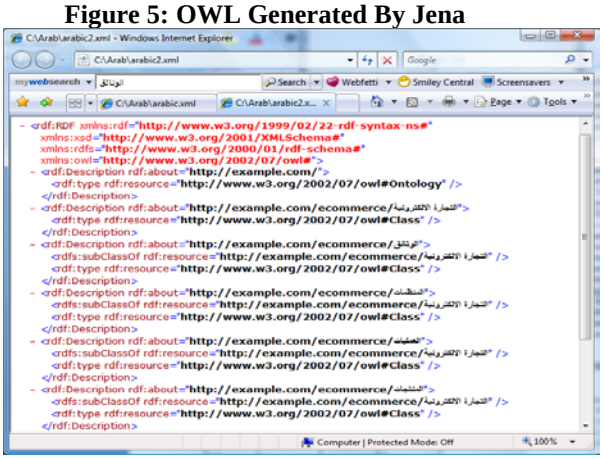


**Figure 5: OWL Generated By Jena**



KAON2 does not support Arabic at all, although UTF-8 encoding is already being used.

TABEL 2: ARABIC SUPPORT SUMMARY

| Tool | RDF | OWL | Query |
|---|---|---|---|
| **Protégé** | Support Arabic | Limited Support | Limited Support |
| **Jena** | Support Arabic | Support Arabic | Limited Support |
| **Sesame** | Limited Support | Limited Support | NO Support for Arabic |
| **KAON2** | NO Support for Arabic | NO Support for Arabic | NO Support for Arabic |

All evaluated systems don't support Arabic language processing or diacritics. There are certain conditions that are required in order to attain this goal. In the given systems being evaluated, no Arabic language or characters are not supported, which is why the need to develop the appropriate system tool that can generate and provide the information in the Arabic language to be made available is urgent and essential.

The study of Hammo in the diacritics is also quite an excellent observation to the process of semantic retrieval of data through ontology. Hammo is convinced that most researches in the field of Arabic Information Retrieval (AIR) did not pay much attention to the problem of searching and retrieving diacritisized text [10]. Most of the studies would actually focus on the tools to breakdown and filter all the necessary semantics in order to retrieve the needed data. Hammo actually sees it differently that his study would result to the citing of the importance of the prefixes and the needed diacritics in every verse or sentence, like for instance, in the Quran. The use of the "bag of words" can actually even create a more problem to the desired results since matches

would be low if the prefix words are neglected or taken for granted.

According to Guo & Ren, the NLP technology is one branch of the linguistics, which uses the computer technology to realize human language processing effectively. The Semantic Web is one of the ultimate and amazing results to this innovation. Through the use of NLP, Ontology was born. The cycling and layering of the different syntax in order to relate the information shared and needed by many internet users is made available. Also, with NLP, semantic data store and retrieval, and multilingual ontology mapping was made possible.

Given this proposition of Guo & Ren, NLP would be one way of finding out how to discover and perhaps design the tools that would support the Arabic Language. The relationships of the NLP and ontology can provide the chance that Arabic characters would match the results in a given query.

## VIII. CONCLUSION & FUTURE WORK

Arabic is the one of the widest spoken language in the world, with over 200 million speakers, utilized by twenty four countries. The need for information in the related language is dramatically high and so there are number of semantic systems to test whether Arabic characters would give out in the event of using the tools.

In this study, the evaluated tools like the Protégé and Jena, Sesame, and KOAN resulted to weak support of the Arabic language and thus, the need for new tools supporting ARABIC NLP is crucial. Moreover, there is a must for developing and designing semantic tools that support Arabic language processing & encoding.

The World Wide Web has created enumerable opportunities that are unimaginable to human. One of these wonders is the chance for human to be able to be understood by the machine. The establishing of the NLP allowed the birth of many possibilities like Ontology, data retrieval and storage. The Natural Language Programming gave way to the mentioned possible actions that the Semantic Web is able to do; and this possibility could help even the users of the World Wide Web who belongs to the Arab countries.

## REFERENCES


[1] Hend S. Al-Khalifa and Areej S. Al-Wabil. The Arabic language and the semantic web: Challenges and opportunities.

[2] http//www.w3.orgW3C - The World Wide Web Consortium [Last accessed 05/09/2010]

[3] http//www.w3.orgW3C - The World Wide Web Consortium [Last accessed 05/09/2010]

[4] Berners-Lee, T. 1998. Semantic Web Road Map. DOI= http//www.w3.org/DesignIssues/Semantic.html

[5] Rodriguez, Horacio,et al. Introducing the Arabic Wordnet

[6] Saleh, L. & Al-Khalifa, H. 2009. AraTation: An Arabic Semantic Annotation Tool

[7] Abu-Hamdiyyah, Mohammad.2000. The Qur'An: An Introduction

[8] Zaidi, S. et al. A Cross-language Information Retrieval: Based on an Arabic Ontology in the Legal Domain

[9] Vossen, P. et al. Introducing the Arabic WordNet Project

[10] Hammo, B. 2009. Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents

[11] Guo and Ren. Towards the Relationship Between Semantic Web and NLP

[12] Al-Khalifa, Al-Yahya, et al. SemQ: A Proposed Framework for Representing Semantic Opposition in the Holy Quran using Semantic Web Technologies

[13] Hammo, Abu-Salem & Lytinten. QARAB: A Question Answering System to Support the Arabic Language

[14] Buitelaar, P. Human Language Technology for the Semantic Web. http://agamemnon.uni.lu/ILIAS/ai.talks/Slides.Paul_Buitelaar.pdf, 2005.

[15] Pan, et. Al. IBM Research Report. An MDA Based System for Ontology Engineering.

[16] http://www.w3.org/RDF/

[17] http://www.w3.org/TR/owl-features/

[18] Zahari, F. ONTOLOGY APPLICATION FOR THE AL-QURAN.

[19] http://www.fileformat.info/info/unicode/utf8.htm