# A Survey of Arabic Language Support in Semantic Web

Majdi Beseiso
Tenaga National University
Selangor, Malaysia

Abdul Rahim Ahmad
Tenaga National University
Selangor, Malaysia

Roslan Ismail
Tenaga National University
Selangor, Malaysia

## ABSTRACT
Information availability is a key factor in the acquisition of knowledge. Access to information either in the general area or even in more specific ones like sciences, languages, and religion become wider since the use of semantics in World Wide Web.

Semantic Web technologies assist in the acquiring of information by creating processes that link information to another. However, the technology supports mostly languages using Latin family scripts. Arabic is still not well supported.

This paper, reports on the survey of the support for Arabic in some of the existing Semantic Web technologies, and give future scenario in applying Semantic Web for Arabic applications. Finally, multilingual support for these new technologies is also discussed.

## General Terms
Ontology, RDF, Semantic Query, Semantic Tools, Arabic Language Processing.

## Keywords
Semantic web, Arabic Language, Natural Language Processing, of the manuscript

## 1. INTRODUCTION
The world today evolves as fast as we can imagine. Its direction is consistently towards advancement and progress. These advances allow greater possibilities to acquire information and knowledge from the most basic to the more complex. One of the many wonders of the world is the Internet. Through it, human from different locations and walks of life around the world, are able to know each other and interact. The global village, as McLuhan refers to, is already a reality experienced today. Aside from that, access to information is also already possible in the context of Semantic Web.

Semantic Web can improve the existing Web with a layer of machine-interpretable metadata (i.e., data about data) that allows a computer program to understand what a Web page is about, and therefore draw conclusions about the Web page [1]. This innovation by Tim Berners-Lee's on the web content medium can work with the software agents to allow information to be found, shared and integrated. With the semantic web, there is also sharing and exchange of data that makes knowledge more accessible and easy to discover and research.

The Semantic Web assists the evolution of knowledge as well, when open for use worldwide. New concepts can be exposed and this would enable everyone to exchange expressions using a unified logical language allowing computers in the world to connect to a universal Web. By these connections humans can have access to a wide array of knowledge and ideas. In this way, people can live, work, and learn together. It opens a lot of possibilities for the next generation of technology users. Due to this it is important for the Semantic Web tools and applications to be able to support all languages of the world, one of which is Arabic in order to fulfil the Semantic Web goal of connecting the world into one network.

## 1. 2. SEMANTIC WEB AND ARABIC LANGUAGE

### 2.1 Importance of Arabic language
Arabic language is integral to majority of the population of the Middle-East. The language distinct them from countries in other regions and it is also a language manifest in their faith. 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]. It 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]. The Arabic Quran which means 'the recital' or the proclamation [7] was revealed to Muhammad, the Prophet of Islam making the use of Arabic wider among the Muslims, those who profess Islam.

### 2.2 Task Difficulties
The Arabic language is a difficult language that may hinder the development of the tools for Semantic Web in that language. Arabic Language has many particularities like short vowels, absence of capital letters and complex morphology. Arabic language is composed of nouns, verbs and particles; wherein these are morphemes and derived from a closed set of around 10,000 roots. Arabic is also highly inflectional and derivational, which makes morphological analysis a very complex task [7]. There is no capitalization in Arabic, which makes it hard to identify proper names, acronyms, and abbreviations.

## 2.3 Arabic Language & Semantic web research

There are various studies conducted on Arabic language in Semantic Web. Zaidi, Laskri and Bechkoum [8] proposed to improve the Arabic information retrieval on the Web in the legal domain by an Arabic search engine supporting the translation of Arabic queries into English or French queries. The aim was to return documents written in Arabic, French or English. Vossen, Pease and Fellbaum [9] worked on Arabic Word Net (AWN) based on the methods developed for EuroWordNet (EWN) and since applied to dozens of languages around the world. The EuroWordNet approach maximizes compatibility across Word Nets and focuses on manual encoding of the most complicated and important concepts. The basic criteria for AWN are connectivity, relevance, and generality, from English to Arabic and from Arabic to English. Hammo [10] surveys on enhancing retrieval effectiveness of search engines for diacritised Arabic documents by building an Arabic–English IR system based on a machine translation approach. AbdulJaleel and Larkey [20], proposed a statistical transliteration approach for Arabic–English IR.

Grefenstette et al. [21], 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. [22], described how precision can be improved in query expansion using LSI. Finally, Semmar and Fluhr [24], 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 [11] highlighted the use of Natural Language Processing (NLP) technology as a significant component in Semantic Web tool. NLP is one branch of the linguistics, which uses the computer technology to realize human language processing effectively. Its ultimate objective is to automatically understand human language with the support of artificial intelligence technology. It is also called as natural language understanding and sometimes is used to transform information to Semantic Web data. Traditional information retrieval also can be turned into knowledge discovery. Al-Khalifa, Hen, Al-Yahya, Bahanshal and Al-Odah [12] proposed a framework for representing a semantic opposition in the Holy Quran using Semantic Web Technologies. Previous research 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. Very little work has been done toward using semantic web technologies for serving the lexical semantics of the Holy Quran.

Hammo, Abu-Salem and Lytinen [13] developed a system QARAB whose main goal is to identify text passages that answer a natural language question. The tasks in QARAB 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 directly or indirectly in Semantic web in Arabic language. Based on the information gathered, it can be concluded that work in Arabic language for Semantic Web is still in the infancy. Due to that, it is possible to progress further besides the current available Arabic semantics like those that are used in the Quran.

# 3. ARABIC ONTOLOGY

Arabic ontology is the foundation of the creation of Semantic Web in Arabic language. Basic categorization of terminologies and meanings in a domain give the semantics. The interrelationship between one word to the other words that matches to its meaning can also result to the stems and branches of semantics. Ontology can be built by using domain experts or learned from information available in a corpus of the domain. The goal of ontology learning is to automatically extract relevant concepts and relations from the given corpus or other kinds of data sets to form Ontology [11]. There are six parts in the life cycle in the development of ontology: Creation, Population, Validation, Deployment, Maintenance and Evolution [14]. The 6 parts above can also be subdivided into the following: extracting terms, discovering synonyms, obtaining concepts, extracting concept hierarchies, defining relations among concepts, deducing 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. There is a strong need for Arabic language support since the ontology in English cannot be translated to Arabic. Since there are no standard ontology in the Arabic language to apply in our test, we used our e-commerce ontology proposed in [24] with some refinement based on English e-commerce ontology proposed by Geller. Figure 1 shows the 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.

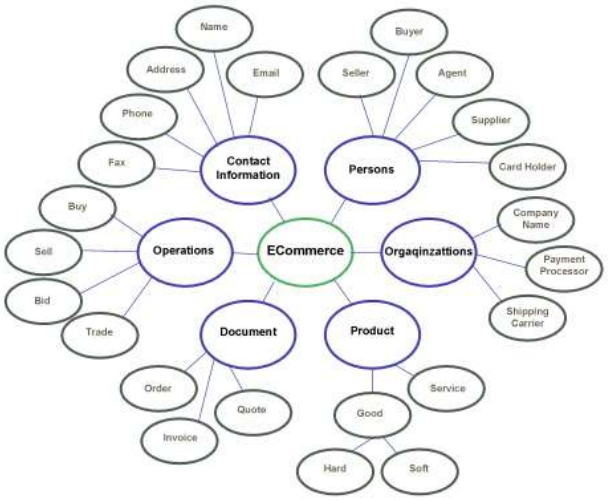


Figure 1 : E-Commerce Ontology [24]

## 4. SYSTEMS TO BE EVALUATED

Ontology is one of the basic and the preliminary sources in order to start the process of building Semantic Web. To build ontology, there are different systems that can be used. In this paper, evaluation of these tools is made. Among the systems evaluated are the traditional Ontology Management Systems such as Java's Protégé, Jena, Sesame, and KOAN. These systems have been designed and developed to assists in the creation of 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 that provides 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.

Pan, et.al, [15] noted that while these engineering ontology tools provide a stack of Ontology management support, they also show certain limitations in supporting large scale software engineering projects and in languages other than English. Therefore, there is a need to study and evaluate each of the given tools before deciding on the choice to use for development of ontology, especially if the Ontology is in the Arabic language.

**Table 1 : Summary of 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 |

## 5. THREE DIMENSIONS FOR EVALUATING ARABIC IN SEMANTIC WEB

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 needed. They 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.

## 6. RESULTS AND DISCUSSION

A series of studies were conducted to evaluate the different frameworks and the systems available. The following results and discussions are presented:
Protégé can basically create & display ontology in Arabic, jambalaya plug in is successful in displaying Arabic text in the form 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 in 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 the use of lexical resources in order that the connectivity is made faster and accessibility is wider. Once there is enough connectivity of the lexical resources such as the synsets, the availability of the needed resources can more likely give relevant results.

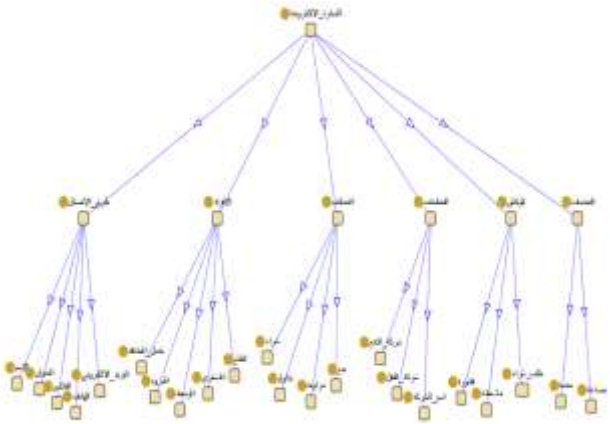

**Figure 2: Protégé Jambalaya**

Protégé Query tools support Arabic text query; however, without diacritics or stemming, Arabic language would not be well 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 integrate with Jena query engine for English language processing but nothing is available as yet 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.

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

All evaluated systems do not 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, there is no Arabic language or characters that are supported. That is why there is a 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.

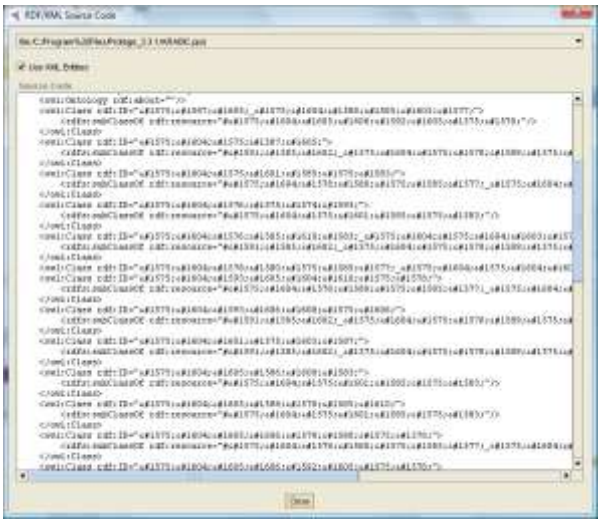

**Figure 3: Protégé OWL Generated**

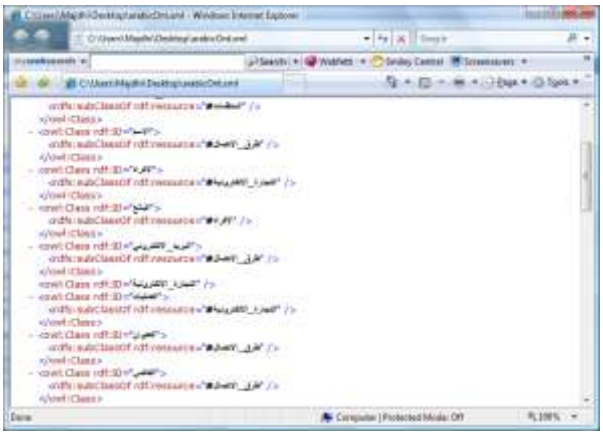

**Figure 4: OWL Preview**

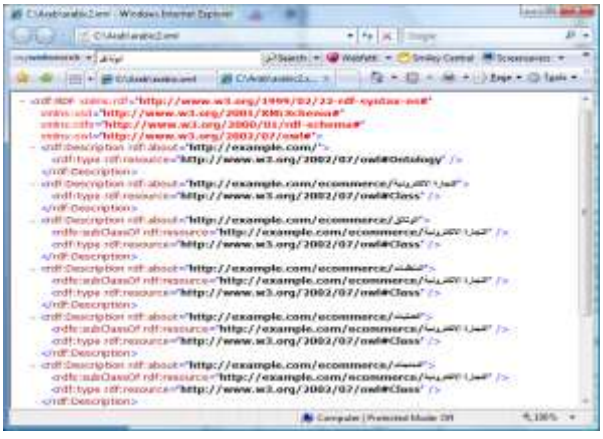

**Figure 5: OWL Generated By Jena**

**Table 2: Summary of Arabic Support**

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

The study of Hammo [13] in the diacritics is also quite an excellent observation on 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 diacritized text [10]. Most of the studies were actually focusing 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 create more problem for the desired results since matches would be low if the prefix words are neglected or taken for granted.

According to Guo & Ren [11], 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.

## 7. 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 quite 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 showed weak support of the Arabic language and thus, the need for new tools to be developed in supporting NLP for Arabic is crucial. Moreover, it is a must for development and design of 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 machine. The establishment of the Natural Language Programming allowed the birth of many possibilities like Ontology, data retrieval and storage. The NLP 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.

## 8. 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.org W3C - The World Wide Web Consortium [Last accessed 05/09/2010]

[3] http://www.w3.org W3C - 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

[20] AbdulJaleel, N. and Larkey, L. (2003). Statistical transliteration for english-arabic cross language information retrieval. Proceedings of CIKM, pages 139–146.

[21] Yan Qu, Gregory Grefenstette, David A. Evans: The Use of Monolingual Context Vectors for Missing Translations in Cross-Language Information Retrieval. IJCNLP 2005: 22-33

[22] Grefenstette, G., Semmar, N., and Elkateb-Gara, F. 2005. Modifying a natural language processing system for European languages to treat Arabic in information processing and information retrieval applications. In Proceedings of the ACL Workshop on Computational Approaches To Semitic Languages (Ann Arbor, Michigan, June 29 - 29, 2005). ACL Workshops. Association for Computational Linguistics, Morristown, NJ, 31-37

[23] Abdelali, A., Cowie, J., and Soliman, H. S. 2007. Improving query precision using semantic expansion. Inf. Process. Manage. 43, 3 (May. 2007), 705-716.

[24] Semmar, N., Fluhr, C., Arabic to French Sentence Alignment: Exploration of A Crosslanguage Information Retrieval Approach, 5th Workshop on Important Unresolved Matters, pages 73–80, Prague, Czech Republic, June 2007

[25] Beseiso, M., Ahmad, A.R, Jais, J., Semantic Arabic Search Tool, STAKE2010, Kuching, Sarawak,