# The Arabic language and the semantic web: Challenges and opportunities

Hend S. Al-Khalifa
*University of Southampton, Southampton*
Areej S. Al-Wabil
*City University, London*

***Keywords*** *: Semantic Web, Arabic language, Semantic Web tools, Arabic Ontology, Arabic Script.*

***Abstract*** *Since the advent of the Semantic Web in the late 90's, many Web applications were created to benefit from the capabilities provided by Semantic Web technologies. These capabilities include intelligent reasoning over data, semantic search and data interoperability. However, most Semantic Web technologies are dedicated to processing Latin family scripts, thus, an apparent lack of Arabic script support in these technologies kept the research in the Semantic Web for the Arabic language untackled. This paper reports on a pilot experiment to test the support of Arabic in some of the existing Semantic Web tools, and examines the challenges and opportunities encountered when implementing Arabic Semantic Web applications.*



## Introduction

The problem with the current Web is that it is only understandable by humans. Machines cannot interpret information on the Web as people do. To illustrate this problem, suppose a person wants to search the Web for the term 'apple', by which (s)he means the fruit. The search engine will return results with no semantic relations. It may supply links to Web pages on 'apple' as in the computer company, the fruit, or even an online shop. This ambiguity in the results needs to be resolved by semantically annotating resources on the Web so that intelligent results can be retrieved. This can be done by adding an extra layer of semantics to the current Web to enable machines to understand what a Web page is about.

Semantic Web (SW) technologies have proved their success in multiple domains, such as Medicine, e-Commerce, e-Learning and Biology. To extend this success to the Arabic language, a set of SW tools and applications needs to be created to fulfill the requirements of the Arabic language.

Thus, the research questions we are trying to answer in this paper are as follows:

- Can existing Semantic Web tools facilitate building Semantic Web applications that support the Arabic language?
- What are the challenges faced by users of the Arabic language when trying to embrace Semantic Web technologies?
- What opportunities will this new technology bring to users of the Arabic language?

To share our understanding of the answers to these questions and to shed some light on the potential research in this field, this paper is organized as follows: in section 2, we present some background information about the SW and discuss its basic concepts. In section 3, we outline the main components of a typical SW Annotation tool. In section 4, we present an overview of SW research in Semitic languages. In section 5, we construct a simple SW application to evaluate the support of the Arabic language in some of the open-source SW tools. Finally, we conclude the paper with some of the main reasons why research conducted for Arabic has not contributed to SW research, and we provide a discussion of the potential challenges and opportunities that are awaiting researchers in this domain.

## Background

The Semantic Web intends to improve the existing Web with a layer of machine-interpretable meta-data (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. The SW as defined by its creator Tim Berners-Lee (Berners-Lee et al., 2001), implies "… an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation."

To add a layer of semantics to the existing Web, three challenges need to be achieved (Harmelen, 2004):

- A syntax for representing metadata,
- Vocabularies for expressing the metadata, and
- Metadata for lots of Web pages.

The SW includes the following technologies (Antoniou and Harmelen, 2004):

- ***Explicit Metadata:*** the SW does not rely on text-based manipulation, but on machine-processable metadata.
- ***Ontologies:*** an ontology can be defined as an explicit and formal specification of a conceptualization.
- ***Logic and Inference:*** where automated reasoners can infer conclusions from the given knowledge.
- ***Agents:*** are computer programs that work autonomously on behalf of a person. They receive tasks to accomplish, make certain choices and give answers.



**A Typical Semantic Web Application**

A classic SW application consists of two main parts: the semantic annotation process and the semantic query and reasoning process.

The semantic annotation process, as depicted in Figure 1, requires three components: an extraction engine, an ontology and an annotation generator. First, the extraction engine is responsible for finding the entities of interest in a document using knowledge extraction techniques (e.g., Wrappers, Information Extraction 'IE', etc.). Second, an ontology describes the domain of interest. Third, the annotation generator is responsible for explicitly adding semantic meaning to the extracted entities using the ontology.

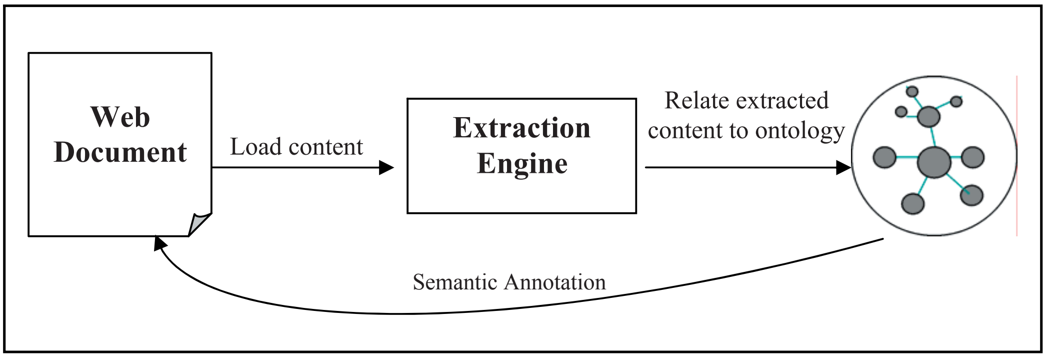


**Figure 1.** The three main components of the semantic annotation process

The semantic query and reasoning process, on the other hand, requires tools that enable 'intelligent' searching through SW data rather than uninformed string matching, which is currently employed in traditional Information Retrieval systems. These tools use the layer of descriptive semantics as a medium for communication between agents and as a way to inference over data and derive conclusions.

**Right-to-Left Languages and the Semantic Web**

Arabic script constitutes 8.9% of the world's languages, as shown in Table 1. Whatever the numbers, it is critical to understand that this high percentage of the Arabic language, when compared to other languages such as Hebrew, raises the following question: Why is there SW research in the Hebrew language and yet we do not have the same application for the Arabic language? Moreover, why do we find Farsi SW applications, as Farsi is from the same script family, and we do not find the same application for the Arabic script?

Studies on language distribution over the Web content show English is the prevailing language for documents with 56% presence-share [1], which makes most SW developers design their tools and applications with Latin-based scripts in mind (i.e., left-to-right script). Nevertheless, there are some emerging attempts to process right-to-left scripts from Middle Eastern languages. There are some applications that have been built to facilitate SW technologies for right-to-left scripts, such as Farsi and Hebrew.

Hasti is an automatic ontology building system (Shamsfard and Barforoush, 2002). It extracts lexical and ontological knowledge from Persian (Farsi) texts. The system starts from a small ontology kernel and automatically constructs the ontology through an understanding of the text. The results of the system have proved its success in handling the Farsi language compared to other ontology learning algorithms (Shamsfard and Barforoush, 2004).

In the case of Hebrew, Klamma et al. (2002) developed a Talmudic tractate that transcribes an original printed edition of The Babylonian Talmud into a structured electronic version, thus providing much easier ways of studying the text with hypertextual annotations and dynamic knowledge level-dependent features.

The previously mentioned cases use proprietary tools to build their SW applications. In this study, our aim was to extend the success of other right-to-left language applications into Arabic using the existing SW tools. To accomplish this objective, a pilot test was carried out to envision the possibilities of building a Semantic Web application that handles the Arabic language with the tools currently available.

| Script | Latin | Cyrillic | Arabic | Hanzi | Indic | Others* |
|---|---|---|---|---|---|---|
| Million users | 2,238 | 451 | 462 | 1,085 | 807 | 129 |
| % of Total | 43.3% | 8.7% | 8.9% | 21.0% | 15.6% | 2.5% |
| Key languages | Romance (European) Slavic (some) Vietnamese Malay Indonesian | Russian Slavic (some) Kazakh Uzbek | Arabic Urdu Persian Pashtu | Chinese Japanese Korean | Hindi Tamil Bengali Punjabi Sanskrit Thai | Greek Hebrew Georgian Assyrian Armenian |

**Table 1.** Clustering of the world's languages based on their script family [source: Paolillo, 2005]

**Arabic Research in the SW domain**

To the best of our knowledge, only a few published studies have employed SW technologies in developing Arabic language applications. Existing research exploring the potential of SW applications for Arabic is mainly in the Cross-Language Information Retrieval (CRIL) domain.

In a study by El-Helw and Aly (2004), an intelligent database application was developed using the techniques and structures of the SW. The integrated system developed in that study was capable of extracting data from unstructured and structured documents available in Arabic newspapers. Formal definition rules in the specified domain were developed to allow for querying this data for implicit information. The study falls short in that only one key technology of the SW was employed in the integrated system, namely, the inferencing rules. Incorporating key SW technologies, such as ontologies, to represent the data would have added an interesting dimension to their approach.

Zaidi and Laskri (2005) have developed a Web-based multilingual tool for Arabic information retrieval based on ontology in the 'legal' domain. They used Protégé to develop their ontology and they have also used the query engine built in Protégé for retrieval purposes. One important shortcoming of their application is that the produced ontology relies on the RDF language, which lacks important axiom constructs found in more expressive languages such as OWL. Another shortcoming is that the reported study focused only on how to build the legal ontology without mentioning how they annotated legal documents with semantic metadata.

Another similar use of SW for the Arabic language is an experiment carried out by Abdelali et al. (2003), in which the researchers retrieved Arabic documents from a multilingual corpus. They used ontologies instead of the usual lexical-based information retrieval technique for retrieving Arabic documents. Despite the promising results that the ontology-based information retrieval technique has provided, the full power of SW technologies was not demonstrated in their research.

**Semantic Web Tools: A Pilot Test for Arabic Support**

For the semantic annotation and query to take place, a set of tools were designed to facilitate the process of annotating resources on the Web (i.e., attaching semantic metadata to a resource) and exploiting their semantic relationships. These tools provide the described resources with a semantic metadata in a machine-readable, machine-understandable and usable form to anyone interested in using them.

Tools used for constructing SW applications vary in their openness and functionality. Most well-known SW tools are open-source and free for public use. Commercial SW tools also exist, which include Oracle's 10g RDF database management platform [2] and Intellidimension's RDF Gateway [3].

Table 2 shows some SW tools that have been classified based on their functionality. From the comprehensive listing of 250 SW tools found in (Bergman, 2006), we can roughly estimate that 12% of the various tools (ontology editors, repositories, API's, reasoners, interpreters and inference engines) support Arabic. This assumption was based on the programming language the tool was built on.

**Table 2.** Listing of tools used in building Semantic Web applications

| Function | Tool | Description |
|---|---|---|
| Ontology Editor | Protégé | Visual ontology editor written in Java with many plug-in tools. |
| Ontology Repository | Sesame | An open-source RDF database with support for RDF Schema inferencing and querying. |
| Information Extraction | GATE | Open-source infrastructure that provides language processing capabilities. Plug-ins for processing Arabic documents are available at: http://gate.ac.uk/gate/doc/plugins.html#arabic |
| Reasoners and Processors | Jena | A Java framework to construct SW applications. It provides a programmatic environment for RDF, RDFS, OWL, SPARQL and includes a rule-based inference engine. |

Although these tools were designed with Latin script processing in mind, further investigations need to be carried out to test their support for Arabic. Thus, we have conducted a pilot test to evaluate Arabic character processing and encoding support in selected SW tools.

**Pilot Test**

Our pilot test framework consisted of two well-known open-source tools that are usually used for building SW applications, namely Protégé [4], an editor used for building ontologies and creating semantic metadata, and Jena 2 API [5], an application programming interface library used within Java for reasoning over the semantic metadata.

Our aim for the creation of a simple SW application is to verify whether some of the widely used existing tools can fulfill the requirements of building applications that can process the Arabic language without any problems.

To construct our semantic application a simple scenario was adopted. The scenario was about conceptualizing the parental relationship for a fictional family. Figure 2 shows an ontology scheme with its associated instances for this relationship. It should be noted that the ontology scheme represents a high level abstraction of the different concepts in our sample domain, which include the following {إنسان، ذكر، انثى}. The ontology instances represent the actual individuals in our domain, namely {محمد، مريم، سالم، ريم}, along with the different relationships (i.e., properties) between these instances, for example, {زوج، زوجة، ابن، اخت}.

When creating this scheme in Protégé, the concepts pane was capable of displaying the hierarchical listing of the concepts and their instances without any problems as shown in Figure 3. However, when trying to use one of Protégé visualization plug-ins to view the complete ontology, the plug-in failed to display the appropriate Arabic characters; question marks were displayed instead (as shown in Figure 4).

Another hurdle we faced was viewing the generated ontology in OWL format [6]. The displayed text appeared as unintelligible code instead of in the proper Arabic text. However, when saving the file into OWL format, the Arabic characters were preserved. While it was possible to save and process Arabic script in OWL format, it was not possible to display Arabic characters in the ontology correctly.

When we finished building the ontology using Protégé, a simple pilot program was written in Java with Jena 2 API to test the support for Arabic processing in Jena. The rationale behind building the program was to illustrate the use of Jena's semantic reasoner and to show how Jena can deal with Arabic text.

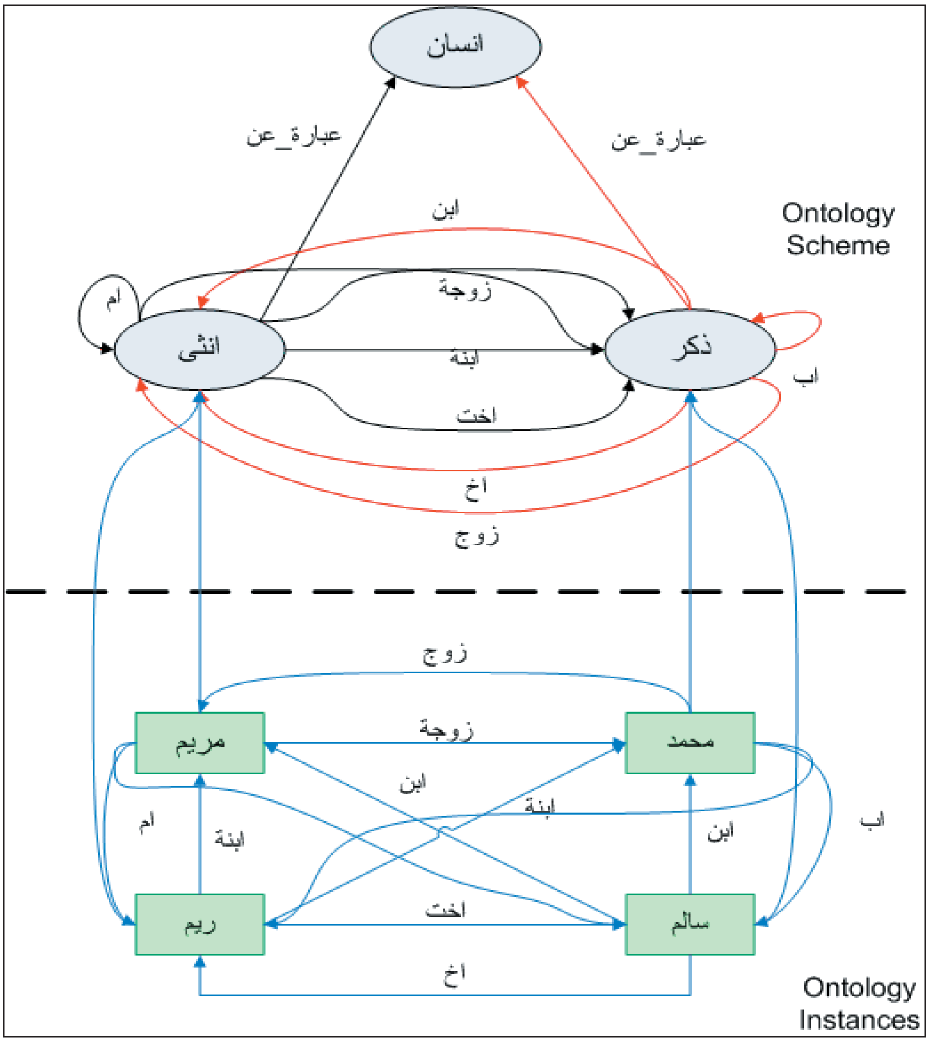


**Figure 2.** The fictional family relationship

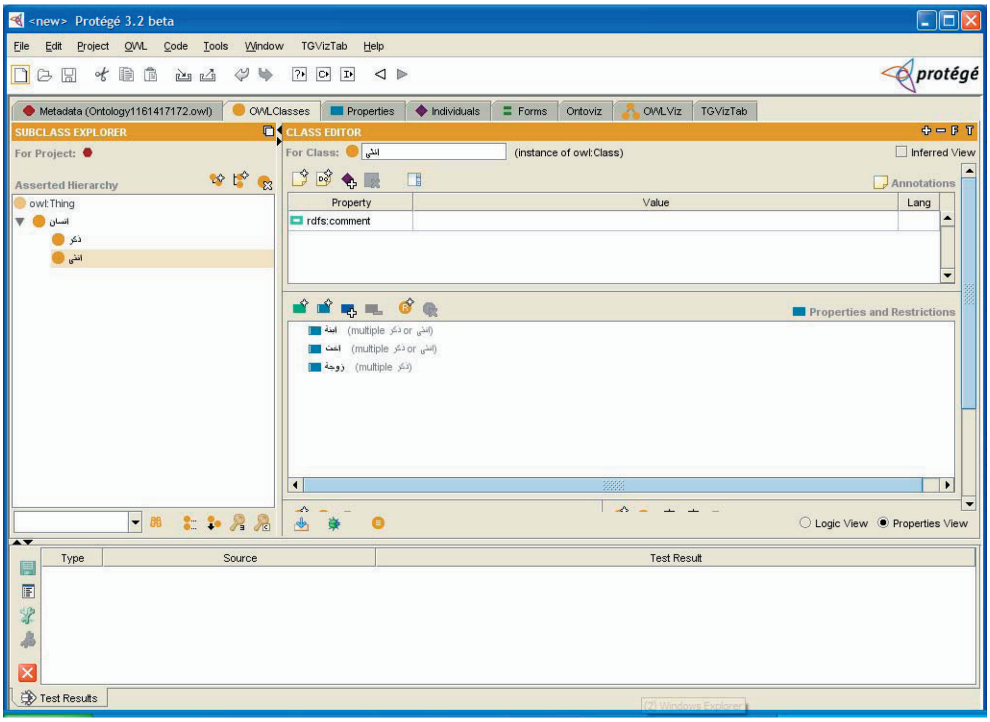


**Figure 3.** Protégé editor showing the different concepts (left pane) and their associated relationship (i.e., properties, which is in the right-middle pane)

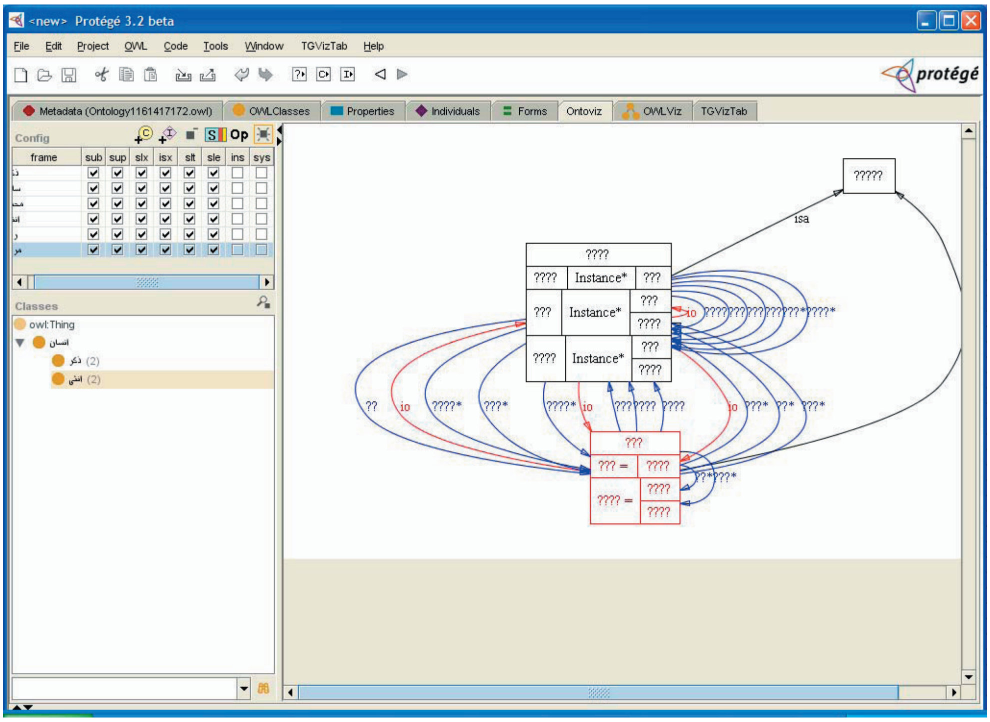


**Figure 4.** The garbled appearance of Arabic characters when using the Ontoviz visualization plug-in

## Discussion

Protégé provided us with a visual editor for creating our sample ontology along with its associated instances. The visual editor was capable of displaying Arabic text without any problems, and we were able to use the SPARQL [7] query engine in Protégé to test for some sample queries such as:

- Who is ‘سالم’’s father?
- Who are ‘ريم’’s parents?
- What is ‘محمد’’s gender?

Unsurprisingly, the SPARQL engine exploited the strength of OWL expressiveness and provided us with an interface to inference the different relationships between the ontology instances (see Figure 5). Moreover, the engine was capable of processing Arabic text without any glitches.

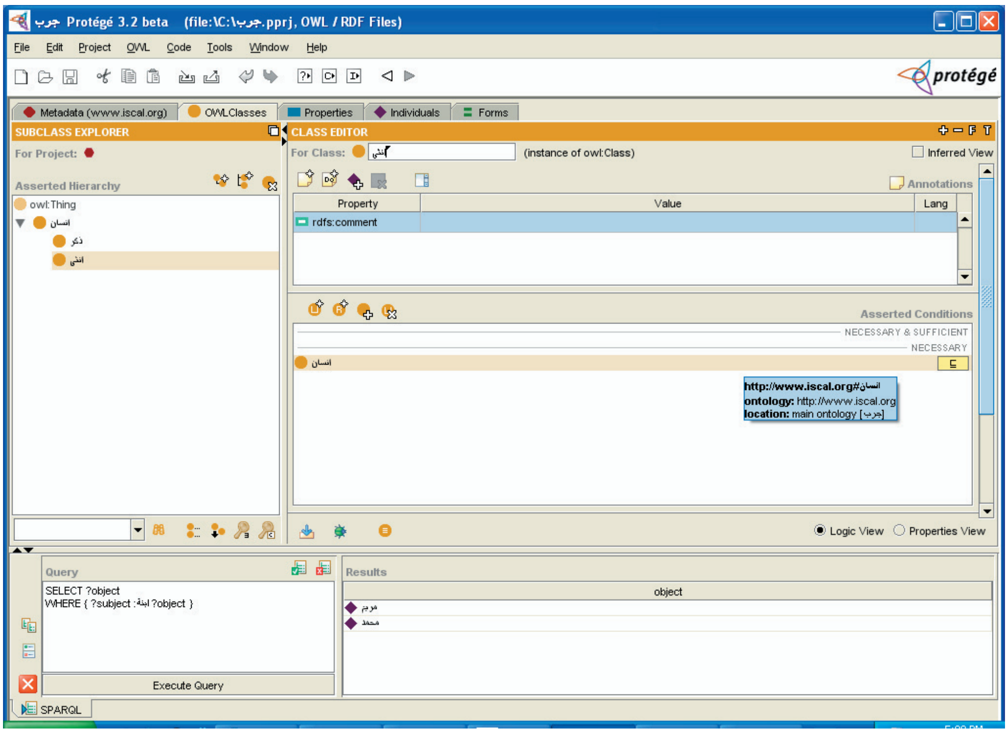


**Figure 5.** The result after executing the (Who are ‘ريم’’s parents?) query is shown in the lower-right pane

The results of the sample program that we created to use with Jena 2 API was encouraging. Jena was able to load and read our Arabic ontology and process the queries accordingly. This result refuted the claim that Hatem, Neagu and Ramadam (2006) reported concerning Jena's lack of support for the Arabic language.

**Challenges and Opportunities**

Our preliminary investigation in the field of the SW revealed a number of issues that suggest reasons for the lack of Arabic research, namely the lack of technology support and adequate resources.

***1) Lack of Arabic support in existing Semantic Web tools***

A specific problem with SW tools processing Arabic text concerns encoding. Different encoding of Arabic script exists on the Web; dominant encodings include UTF-8, Windows-1256 and ISO-8859-6 [8]. Moreover, most of the SW tools we have encountered were built using Java, which supports internationalization. Therefore, there is a strong need to consolidate the different Arabic encoding or simply adhere to one encoding schema when representing Arabic text (i.e., Unicode). Typical SW developers' tools use Unicode throughout (Carroll, 2005); hence this might solve part of the support problem.

***2) Lack of Arabic Semantic Web applications***

Another evidence of the lack of Arabic in the Semantic Web world is a recent statistic provided by the OntoSelect [9] ontology library, which shows that 49% of the ontologies in the library are created in English. The distribution of languages used in creating ontologies, depicted in Figure 6, highlights the problem of limited Arabic SW applications. This problem could be attributed to the lack of tools and software development environments that process Arabic script in all steps of the semantic annotation process. Few means are available to generate SW annotations routinely and effortlessly at the point of content use or creation

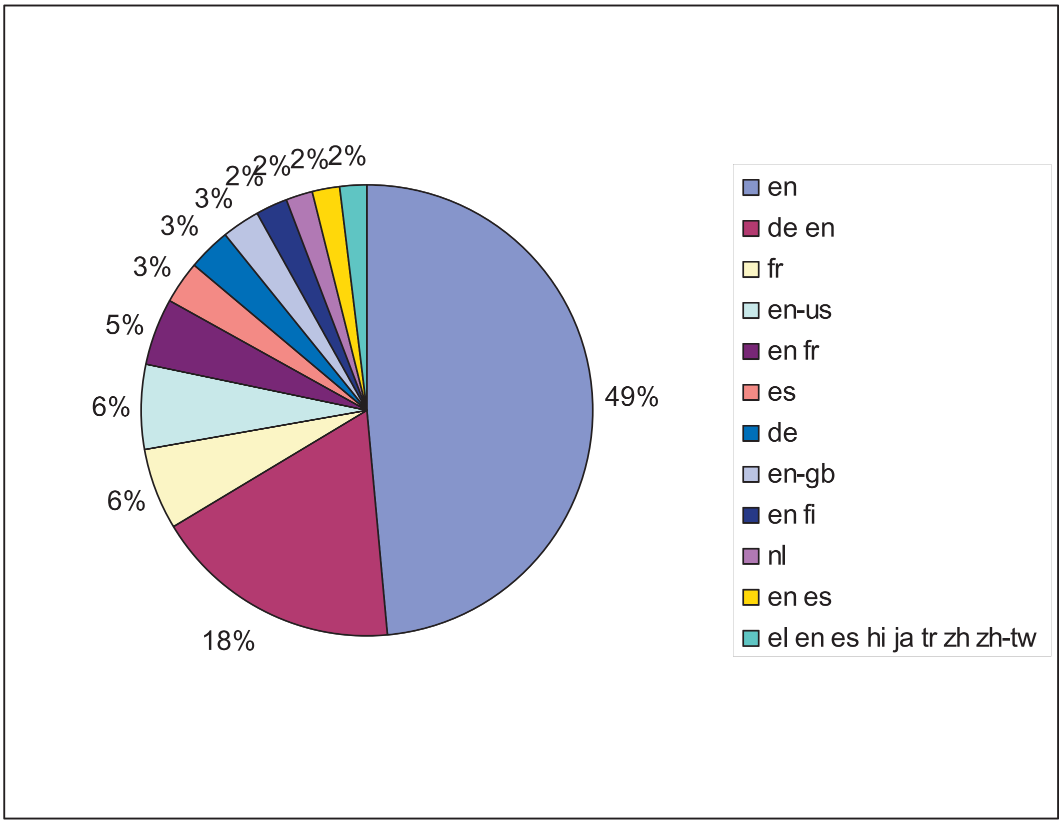


**Figure 6.** A pie chart showing the distribution of languages used in creating ontologies stored in the OntoSelect library

***3) Limited support for Arabic research in the field of Semantic Web technologies***

Most of the SW research is a result of investment from both grant bodies and academic research centers. SW tools, such as Protégé, GATE, and Jena to name just a few are all products of successful investment in the SW field. For the particular case of Arabic, the limited research problem can be attributed to the lack of adequate resources in terms of skills, funding and interest in this emerging field of Web research. The allocation of research funding, the provision of resources, and interest from a committed practice community are essential if we are to overcome this problem. Experience with similar Web technologies suggests that a focused community of enthusiastic early adopters, no matter how small, is an essential prerequisite for success. Key Arabization interest groups and developer networks, such as Arabeyes and Unicode consortium, may be able to contribute to this effort. In terms of financial resources, a major concern -particularly for communities with a small user base such as Arabic users- is the cost of SW application development and maintenance. In some well-structured areas such as scientific, commercial and government applications, the potential benefits in productivity gain and profit will outweigh the cost of developing and maintaining an ontology/application. The cost, in terms of time and effort required, will decrease as the user base increases.

There are many potential challenges and opportunities that users of the Arabic language may encounter with the commencement of SW solutions. Some of these challenges include the following:

1. The need for the development of controlled vocabularies and ontologies to explicitly define machine-processable semantics. These controlled vocabularies and ontologies would help people and machines in communicating concisely and support the semantic and syntax exchange between them.



Such promising efforts are being developed; they include the construction of an Arabic WordNet (Elkateb et al., 2006)

2. Computational processing of Arabic text differs from its English counterpart. Arabic language encompasses more complex morphological, grammatical and semantic aspects that existing Natural Language Processing (NLP) algorithms used for the English language cannot be directly re-purposed for the use of the Arabic language. This problem also has its roots in the SW domain. Information extraction, one of the SW's main processes, relies extensively on extracting concepts from Web documents; this process requires the analysis of the document content - either morphologically, grammatically or semantically - to be able to relate the extracted instances to a pre-defined ontology concept. Thus, more tailored NLP tools such as GATE [10] are needed to process the Arabic language accordingly.

As for opportunities, the SW provides rich opportunities for users of the Arabic language to process data at many levels. In terms of data, large and currently growing repositories of Arabic content in business, science, government documents and email messages exist in the current Web. The SW provides users with a common framework to integrate and derive new meaning, value and intelligence from these existing Arabic repositories. An important element of the SW is the language used in recording how the data relates to real-world concepts. Arabic content on the Web can be networked in meaningful ways in semantic-enhanced applications, as the concept of a multilingual SW supporting Arabic content is attainable. In terms of language barriers, the SW offers users of the Arabic language, and machines alike, the chance to move from one part of the Web to another based on a related meaning regardless of the language used to represent that meaning.

Opportunities can also be highlighted from local and regional perspectives. In the era of knowledge economies, modern organizations and governments are increasingly turning to knowledge management as a differentiating asset to generate greater productivity and create new value. As nations in the region begin the era of e-Government and e-Commerce, shared data understanding and representation for Arabic content is essential. This can be achieved by using SW technologies. Thus computer-mediated Arabic content is attainable by employing SW tools and applications that aim to provide machine-processable data with a universal representation.

**Conclusions**

This paper presented an overview of the Semantic Web, its underpinning technologies, and a preliminary investigation into Arabic support offered by Semantic Web applications. This study reported the results for a pilot test that demonstrated the limited support for processing Arabic script in selected Semantic Web tools.

We explored issues behind the lack of Arabic support in Semantic Web tools and applications and presented suggestions for overcoming these problems. We identified challenges faced by users of the Arabic language, particularly the development of controlled vocabularies and ontologies to define machine-processable semantics explicitly, and tailored NLP tools for processing Arabic text. We also highlighted opportunities for potential Arabic research in this emerging field of Web research

**References**

Abdelali, A., Cowie, J., Farwell, D., Ogden, W., and Helmreich S., (2003) Cross-Language Information Retrieval using Ontology. TALN et multilinguisme, June 2003. Batz-sur-Mer, France.

Antoniou, G. and F. v. Harmelen (2004). A Semantic Web Primer, The MIT Press.

Bergman, M. (2006). Comprehensive Listing of 250 Semantic Web Tools. Available online: http://mkbergman.com/?p=291 (Accessed October 2006)

Berners-Lee, T., J. Hendler and O. Lassila (2001). The Semantic Web. Scientific American 284(5): 34-43.

Carroll, J.J. (2005). An Introduction to the Semantic Web, Considerations for building multilingual Semantic Web sites and applications. HPL-2005-67 technical report. Available online: http://www.hpl.hp.com/techreports/2005/HPL-2005-67.html (Accessed October 2006)

El-Helw, A. and Aly, H. (2004). An Intelligent Database Application for the Semantic Web, in proc. CSITeA-04 conference.

Elkateb, S., Black, W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., (2006). Building a WordNet for Arabic. In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006).

Harmelen, F. v. (2004). The Semantic Web: What, Why, How, and When. IEEE Distributed Systems Online vol. 05 (no. 3).

Hatem M.S., Neagu D.C. and Ramadan H.A.. (2006). Towards Personalization and a Unique Uniform Resource Identifier for Semantic Web Users within an Academic Environment. International journal of Instructional Technology and Distance Learning.

Available Online: http://www.itdl.org/Journal/Jun_06/article01.htm (Accessed October 2006)

Klamma R., Jarke M., Hollender E., D. Börner-Klein. (2002). Enabling Communities by Constructed Media: The Case of a Web-Based Study Environment for a Talmudic Tractate. In: J. Fong, C.T. Cheung, H.V. Leong, Q. Li (Eds.): Advances in Web-Based Learning, Proceedings of the First International Conference, ICWL 2002, Hong Kong, China, August 17-19, Springer-Verlag, Berlin Heidelberg, LNCS 2436, p. 275.

Paolillo J., (2005). Language Diversity on the Internet. pp. 43-89, in John Paolillo, Daniel Pimienta, Daniel Prado, et al., Measuring Linguistic Diversity on the Internet, UNESCO Publications for the World Summit on the Information Society 2005, 113 pp. See http://www.uis.unesco.org/template/pdf/cscl/MeasuringLinguisticDiversity_En.pdf.

Shamsfard M. and A.A. Barforoush. (2002). An Introduction to Hasti: An Ontology Learning System. Proceedings of Artificial Intelligence and Soft Computing. Banff, Canada.

Shamsfard M. and A.A. Barforoush. (2004). Learning ontologies from natural language texts. International Journal of Human-Computer Studies. 60(1), 17-63.

Zaidi, S., Laskri, M. (2005). A cross-language information retrieval based on an Arabic ontology in the legal domain. The International Conference On Signal-Image Technology & Internet–Based Systems (SITIS'05), Morocco

**Endnotes**

[1] http://www.isoc.org
[2] http://www.oracle.com/technology/tech/semantic_technologies/index.html
[3] http://www.intellidimension.com/pages/site/products/rdfgateway.rsp
[4] http://protege.stanford.edu/
[5] http://jena.sourceforge.net/
[6] Web Ontology Language (OWL) is a markup language for publishing and sharing data using ontologies on the Internet.
[7] A query language and data access protocol for the Semantic Web. http://www.w3.org/TR/rdf-sparql-query/
[8] For a full list of other Arabic legacy encoding the reader is referred to http://www.i18nguy.com/unicode/codepages.html
[9] http://olp.dfki.de/OntoSelect/index.php?mode=home
[10] http://gate.ac.uk/