International Journal of Computing and Digital Systems

A Proposed Framework for Arabic Semantic Annotation Tool

Ahmed N. El-ghobashy, Gamal M. Attiya, and Hamdy M. Kelash

Computer Science & Engineering Department, Faculty of Electronic Engineering, Menoufia University, Egypt.
E-mail address: ahmed.elghobashy@gmail.com, gamal.attiya@yahoo.com

Received 3 Sep 2013, Revised 9 Oct. 2013, Accepted 13 Nov 2013, Published 1 Jan 2014

Abstract: Semantic Web is an approach to facilitate communication by making the web suitable for computers. To enhance current Web, a semantic layer should be added to the web pages to enable computers understand them. Recently, some annotation tools have been developed to make machine understandable data on the web. However, little number of tools are concerned with the Arabic Language, although, this language is the mother tongue of more than 293 million of people in 23 countries. This paper first presents an overview of the existing Semantic Web concerning the Arabic Language in the domain of Ontology building. Then, some examples of the tools that can be used in the semantic annotation process are discussed. Finally, a framework is proposed to develop a semantic annotation tool for supporting Arabic contents.

Keywords: Semantic Web; Annotation Tools; Software Architecture; Arabic Language; Ontology.

I. INTRODUCTION

Tim Berners-Lee, the inventor and the director of the World Wide Web Consortium (W3C) expresses it as *"a web of data that can be processed directly and indirectly by machines"* [1]. With the advancement of the Internet and the web content, the direction is to make the web suitable for computers, i.e., make web content meaningful to computers. This facilitates the communication between human and the Internet. To achieve this vision, a machine interpretable metadata (i.e., data about data) layer should be added to the existing web pages. This layer allows a computer program to understand what a Web page is about, and therefore draw conclusions about the web page [2]. By doing this, different users can smoothly interact, share, and exchange knowledge that represented in a machine-readable format in the same way they dealing with the computer in their specific native tongue and their own style of expressing [3].

Researchers found that Semantic Web (SW) has a very promising future in reshaping the Web and the way of dealing with it. It will open many opportunities for the next generation of internet technology by allowing better ways to acquire information and knowledge from the further complex web.

Lots of countries and communities start researching in the SW field to develop Semantic Web tools that adapt web pages to support their Natural Languages (NL). However, little SW tools are developed to support Arabic language although this language represents the mother tongue of 23 countries and more than 293 million of people [4]. Indeed, Arabic is one of the most robust, richest and most languages able to articulate in the world.

This paper organized as follows. Section 2 focuses on the Semantic Web annotation and methods for creating annotations. Section 3 presents an overview of related work and a summary for the tools. Section 4 concerns with some of the open source SW tools that support the Arabic language and the differences of each one. The proposed framework for developing Arabic annotation tool is presented in section 5. Finally, the concluding remarks and outline of future work are listed in section 6.

II. SEMANTIC WEB ANNOTATION

Semantic annotation is "*the process of labeling Web Pages with the semantics of their contents*" [3]. It can also be defined as the process of mapping data instances to ontological concepts. The purpose of semantic annotation is to enable computers to understand human language so that they can perform tasks that are more intelligent.

In overall annotation process, a note is created while reading any text. This may be as simple as underlining or highlighting passages. Creating these notes or comments,

i.e., a few sentences long, creates a summary for Web content and expresses the significance of each source. In other words, an annotation is further information in a document that identifies or expresses the semantics of a part of that document. Accordingly, it attaches sense tags, names, characteristics, remarks, explanations, etc., to a document or to a selected part in a text. This process helps to associate the ambiguity of the natural language when expressing notions and their computational representation in a formal language, by telling a computer how data items are linked and how these relations can be evaluated automatically [5, 6].

Compared with normal annotation that speeds up searching and helps you find related and specific information in a document, semantic annotation goes one level deeper. That is, it enhances the unstructured or semi-structured data with a context that has further linked to the structured knowledge of a domain. In addition, it allows result that has not obviously related to the original search to be reachable.

A usual semantic annotation procedure needs a number of basic preconditions and tools to provide the designated resources with a semantic metadata in a machine readable, machine understandable and usable form to anyone to use them for the representation of semantic annotations, which can be:

- An Information Extraction (IE) module
- A semantic annotation module using ontology.

The purpose of IE is to identify Named Entities (NE) with respect to a specific domain and finding the entities of importance in a document using knowledge extraction techniques. While, the semantic annotation module is responsible for approving the meaning of the words and the semantic relationships of the context by adding semantic meaning to the extracted entities using the ontology [7].

For semantic annotation tool, annotations create a relationship between Uniform Resource Locators (URLs) and construct a network of data. Creating semantic annotations of Web resources can follow one of these methods:

- Manual Annotation
- Semi-Automatic Annotation
- Automatic Annotation methods

A. Manual Annotation

The furthermost basic annotation tools allow users to manually add annotations to Web pages or other resources, and share those. With the alteration of existing syntactic resources into interlinked knowledge structures that represent relevant underlying information [8]. An example of annotation would relate the text “Cairo” to ontology, classifying it as a city and as capital of Egypt.

B. Semi-Automatic Annotation

Semi-automatic annotation tools rely on human intervention at some point in the annotation process. The tools vary in their architecture, information extraction and methods, initial ontology, amount of manual work required to perform annotation, performance and other features, such as storage management [8]. General distinguish between different kinds of semi-automatic annotation mechanisms:

- Wrapper Generation**:** Particularly in the case of annotating Web pages that mainly be made up of Hyper Text Mark-up Language (HTML) tables, one may annotate the first row of the table and automatically enumerate over the remaining rows of the table.
- Pattern Matching: Consistency of word expressions may be captured by consistent expression based patterns. Patterns are stored with the models of the domain ontology.
- Information Extraction: The complex mechanism for semi-automatic annotation is full-fledged ontology based information extraction based on a trivial text processing strategy.

C. Automatic Annotation

Annotation includes robotics components, which deliver recommendations for annotations. The most straightforward kind use rules or wrappers written by hand that try to capture identified patterns for the annotations.

For automatic annotation, two types of systems that learn how to annotate are used: supervised systems and unsupervised systems. The supervised systems learn from sample annotations marked up by the user. The drawback of this method is that selecting sufficient good examples is a non-trivial and error-prone task. The unsupervised systems engages a variety of tactics to learn how to annotate without user supervision, but their accuracy is imperfect [9]. The completely automatic creation of semantic annotations is an unanswered problem.

Table I summarizes the advantages and disadvantages of the semantic annotation methods.

III. RELATED WORK

This section presents an overview of some annotation tools to mark the important information. There is more than one group that can be categorized into methods that support automatic, semi-automatic and manual creation of semantic annotations on Web content.

TABLE I. Advantages and disadvantages of semantic annotation methods

Annotation Method	Advantages	Disadvantages
Manual	A very accurate manner of annotating resources. can support the needs of different users.	A costly process, and often does not consider that multiple perceptions of a data source, requiring multiple ontologies.
Semi-automatic	Acceptable speed of annotation with intermediate accuracy.	Annotations need to be reviewed to make sure it is annotation procedure is correct.
Fully automatic	Have multiple perspectives of a data source in respect with fast speed annotating process.	Several are still limited to usage by experts while others are appropriate for understanding workers. User interface design concerns associated with reducing intrusiveness while get the most out of accuracy.

For annotating the content manually, there is a set of tools based on annotation frameworks (e.g. Annotea [10]) enable users to add metadata to content, or some of them developed distinct.

Amaya [11]**:** It is a user-friendly interactive Web browser and editor built on the Annotea framework, which can mark-up Web documents in extensible Mark-up Language (XML) and HTML. The user can make annotations in the same tool they use for browsing and for editing text, making Amaya a good example of a single point of access environment. It has facilities for manual annotation of Web pages but does not contain any features to support automatic annotation.

One Click Annotation (OCA): What You See Is What You Get (WYSIWYG) Web editor for Web browsers that allows for annotating words and phrases with references to ontology concepts and for creating relationships between annotated phrases by enriching content with Resource Description Framework in Attributes (RDFa) annotations. An intuitive user interface hides the complexity of creating semantic data. To process and to store the semantic content as well as to answer queries about resources occurring in the edited content OCA interacts with a server. The best thing about OCA is they consider non-experts having little or no knowledge of semantic technologies as the primary target group, which is a new way to the success of the SW annotation because it depends on accomplishment a massive corpus of users creating and consuming semantic content [12].

AraTation is an Arabic semantic annotation tool for semantically annotating Arabic News content on the Web. Implemented as a desktop application, this tool constructed using the Java programming language and Web Ontology Language (OWL) ontology to produce Resource Description Framework (RDF) metadata for Web pages. The RDF standard will make the annotated Arabic Web pages reusable and machine process-able on the Web [13].

A few numbers of systems use semi-automatic and automatic to annotating content. These include; ***KIM*** [14, 15] and ***GATE*** [16, 17].

KIM platform afford a knowledge and information management infrastructure and services for automatic semantic annotation, indexing, and retrieval of unstructured and semi-structured content. Inside the process of annotation, KIM also performs ontology population. As a base line, KIM examines texts and identifies references to entities (like persons, organizations, locations, dates), then it attempts to match the reference with a known entity, having a unique Uniform Resource Identifier (URI) and description in the knowledge base. KIM is a platform that offers a server, web user interface, and Internet Explorer plug-in, and equipped with upper-level ontology (KIMO) of about 250 classes and 100 properties. Additional, a knowledge base (KIM KB), pre-populated with up to 200,000 entity descriptions [14, 15].

GATE is the most widely known system over the last 13 years. It is used for mass organization and text annotation. It is a desktop application written in Java and can be run under nearly any Operating Systems (OS). It offers many of functionality everyone may want [16, 17].

Table II presents some examples of tools and systems contribute to the revelation of the SW that open the field for Arabic Semantic Web study.

TABLE II. Summary of some tools properties

Tool	Annotation type	Annotation Storage	Annotation Method
Amaya	RDF(S) XLink, XPointer	Local or annotation server	Manual
OCA	XHTML+ RDFa	Annotation server	Manual
AraTation	RDF	Local	Manual
KIM	RDF(S), OWL	RDF(S) knowledge base	Semi-automatic and automatic
GATE	RDF(S), OWL	Local or server	Semi-automatic and automatic

It is clear that, there is a lot of work in the field of Semantic Web annotation, but a lack of Arabic annotation tool, which is a sign of the lack of Arabic effort in the field of SW.

IV. SEMANTIC APPLICATION AND ARABIC LANGUAGE SUPPORT

This section presents some tools that support Arabic language. From these tools, we can say that a few studies deal directly or indirectly with SW in Arabic language. Based on the information gathered, this can remark some of the application according to their domains.

Ontology Applications:

Ontology is one of the elementary and the major foundations in order to start the process of building SW. Finding tool to build Arabic ontology is the basis of the creation of SW in Arabic language, since they offer a well-defined and standardized form of interoperable, machine understandable repositories [18, 19]. There are different tools that can be used. Some of these tools such as ***Protégé*** and ***Jena*** are tested in this section. The need to study and evaluate each of the given system is necessary before deciding which to use for development of ontology, mainly if the Ontology is in the Arabic language.

Protégé is an open source freely obtainable ontology editor and knowledge base framework essentially an ontology visual editor, with a development framework that provides the crucial manipulations and query from ontology [20].

Jena is another Web system used to afford a programmatic environment for RDF, Resource Description Framework Schema (RDFS), OWL, and SPARQL includes a rule-based inference engine. It is also a program development framework for ontology manipulation and query [21].

A brief description of the two semantic tools; ***Protégé*** and ***Jena*** are given in Table III.

TABLE III. Description of the two semantic tools

Tool	Functionality and usage	Standards
Protégé 4.3 (build 304)	Graphical ontology and knowledge base framework with Visual editor written in Java with many plug-in tools for ontology manipulation & query.	RDF RDFS OWL2 SPARQL
Jena 2.10.1	A Java framework to construct SW applications, with programmatic environment for RDF, RDFS, OWL, and SPARQL includes a rule-based inference engine.	RDF RDFS OWL SPARQL

We concentrate in our investigation about the supporting capability of these tools for:

The RDF generation is the corporate model for the data to be ready accessible over the Web. RDF has structures that simplify data merging even if the underlying schemas differ, and it exactly supports the evolution of schemas over time without demanding all the data consumers to be changed. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data be mixed, exposed, and shared across different applications [22, 23].

On the other hand, OWL generation, which measured as the operative model in terms of creating information, accelerates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema by providing further vocabulary along with a formal semantics [22].

Last is Querying Language (QL) Tools such as SeRQL, OWL-QL, RDQL and SPARQL are also required. They allow users to specify dissimilar query for the needed information that would give out results to the given query. All three are associated measurements to determine as to whether these would be helpful in the coming up of the different needed information in the Arabic language [24].

After testing each tool, which was a simple file (RDF/OWL) to see how each system can handle Arabic on it, we get:

Protégé can ultimately create & show ontology in Arabic, This system uses the RDF standard that also makes use of the UTF-8 encoding. However, it might display numeral literal instead of Arabic characters but in general, the showing of RDF/OWL file will appear in Arabic as shown in Fig. 1.

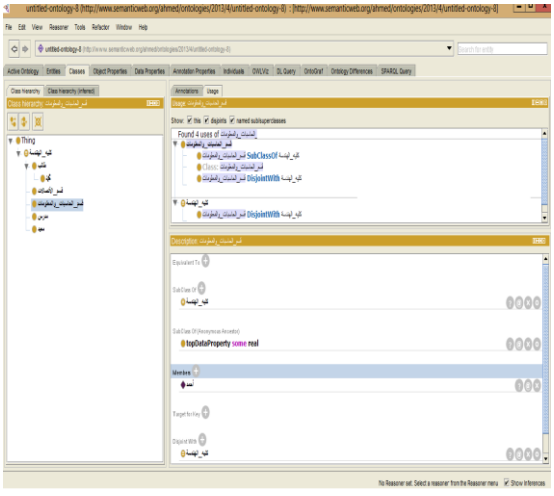

Figure 1 Arabic language handling and creating

But we get some complications when trying to use one of Protégé graph plug-ins to view the complete ontology, the plugin failed to display the appropriate Arabic characters; we got question marks displayed instead as shown in Fig. 2. Moreover, it possible to save and procedure Arabic script in OWL format, but not possible to display Arabic typescripts in the ontology correctly.

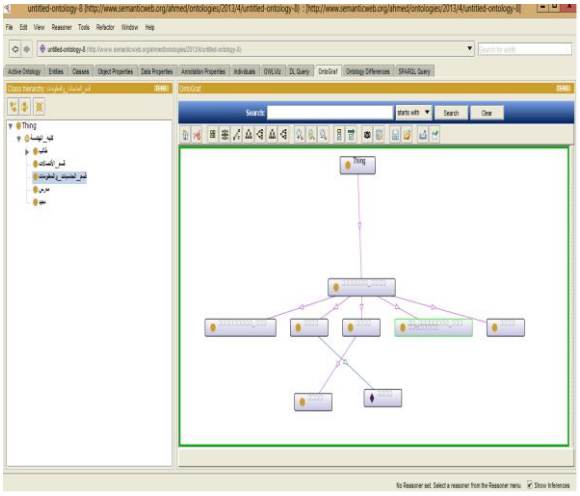

Figure 2 Protégé graph plugins and Arabic problem

Jena system can also construct RDF/OWL File in Arabic. Many Application Programming Interface (APIs) can incorporate with Jena query engine for English language processing but nothing is available yet to support Arabic, so we can query Jena only by strict Arabic word.

Table IV presents a short summary of the investigation study of the two semantic tools; ***Protégé*** and ***Jena.***

TABLE IV. Summary of test results

Tool	Standards	Discussion
Protégé	Support RDF Limited Support for OWL Limited Support for Query	Generating sample ontology capable of showing Arabic text without any problems, the SPARQL query engine in Protégé verified for some sample queries was proficient for processing Arabic text. However, Some problem with graphic plug-ins.
Jena	Support RDF Support OWL Support Query	Capable of load and read Arabic ontology and process the queries consequently without any problem declared.

V. PROPOSED TOOL FRAMEWORK

This section presents the proposed tool framework with some features that can be used in the developing of Arabic semantic annotation tool in the future. The tool must meet some requirements.

A. Tool requirements

Some perspectives that a tool must require are the ontologies, the documents, and the users of the tool. Each perspective suggests one or more requirements, each of which normally brings together several related needs.

1) Formats

Using the latest standard formats that recommended by W3C is favored, wherever possible. Two types of standard are required, one for describing ontology such as the OWL (OWL 2 is now recommended by W3C [25]), and the other for annotations such as the W3C's RDF schema [26].

2) Supported document formats.

Semantic Web standards for annotation accept the documents annotated in web formats such as HTML and XML, but the system should work with any standard without any problem (e.g. XHTML/HTML5). In addition, Documents may be in several formats including word processor files, Portable Document Format (PDF) files, but dealing with various document formats is essential for including annotation into existing work.

3) Annotation Storage.

Annotations can be stored separately from the original document as a bookmark and accessed through a server, or storing annotations as a part of those documents.

4) GUI design.

The annotation tool must use interfaces that make straightforward the annotation process to the user. A good method would be a single point of entry interface, so that the environment in which users annotate documents integrated with the one in which they create, read, share and edit them (e.g. WYSIWYG editors).

B. Tool structure

The general framework of the proposed Arabic semantic annotation tool require three major components, which are, a text preprocessing module, semantic annotation module using ontology, and annotation management module.

- Text Preprocessing module (TP)

The text-preprocessing mission is to clean and normalize the text, and it frequently done before text processing in any Arabic application. Due to characters nature in Arabic, sometimes the same word has different written forms. Therefore, a text-preprocessing module needed to decrease the effect from inconsistency [27].

- Semantic Annotation module (SA)

The semantic annotation module is the main part responsible of understanding the meaning of the words, and the semantic relationships of the context, after that the save operation of the annotated document [28].

- Annotation Management module (AM)

The annotation module used to update the Knowledgebase (KB) of annotations on requests from the user to do so.

In addition, the other needed component depends on the tool implementation like IE module, which used for processing and handling the Arabic language NE, and User Management (UM) module that provides the management of users and access rights to the annotations.

C. Tool implementation

Two ways of implementing the proposed framework of the annotation tool are developed. They are:

1) The first proposed framework

The first proposal is shown in Fig. 3. In this framework, the tool is located on a specific server (annotation server) autonomously of the client. Then, a Web server (proxy server) acts as an interface between the client and annotation server that manages pages with annotations on its base.

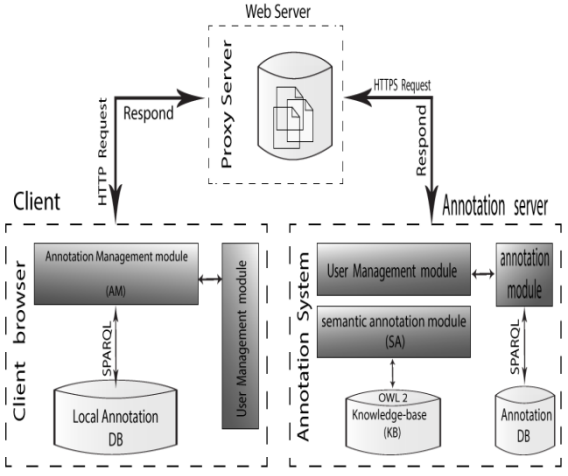

Figure 3 First proposed framework

The problem of this framework is the slow response time. To annotate a document, all tasks done by the proxy server, like requesting pages, the abstraction of annotations, the addition of these annotations and return the response to the browser.

2) The second proposed framework

The second framework is shown in Fig. 4. In this framework, the tool is developed as a plug-in. The purpose is to enhance the browser functions to handle the annotations of a web page. The used browser must follow the web standard's (e.g. Firefox or Google Chrome).

The second framework has further advantages than the earlier, as the ability to annotate web documents stored locally, avoiding the slow request and response at the server. However, it remains dependent on the type of browser used and the opportunities of distributing annotations.

VI. CONCLOUSION AND FUTUR WORK

There are many existing SW application ease the building of semantic annotation tool. However, a rare number of tools that is demanding on the target of leveraging SW technologies to support the Arabic language, and produce semantically annotated Web documents. There may be some task difficulties to take Arabic language in respect. Arabic language is a challenging language that may delay the development of the tools for SW in that language because Arabic language has much discrimination like short vowels, nonexistence of capital letters and composite morphology.

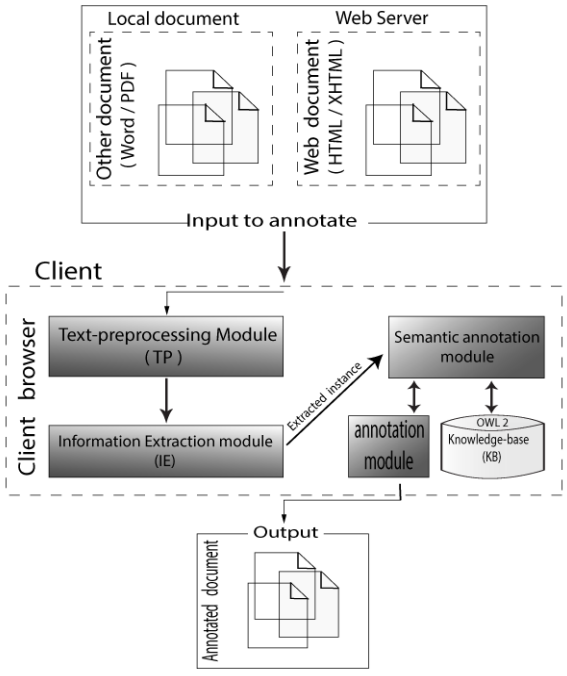

Figure 4 Second proposed framework

In the future work, the main concern is to develop an Arabic annotation tool by applying one of the proposed frameworks that discussed earlier and compare it with other tools in the same field.

ACKNOWLEDGMENT

I would like to thank Dr. Haytham T. Al-Feel, Department of Internet and Multimedia, IAEMS, Egypt, for his kind assistance.

REFERENCES

[1] http://www.w3.org W3C - The World Wide Web Consortium [Last accessed September 30, 2013].

[2] Berners-Lee, T. 2005. Keynote paper in BCS Workshop on the Science of the Web, London.

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

[4] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers [Last accessed September 30, 2013].

[5] Handschuh, S. (2005). Creating ontology-based metadata by annotation for the semantic web . Doctoral dissertation, Karlsruhe, Univ., Diss., 2005.

[6] Cardoso, J. (2007). The semantic web vision: Where are we?. IEEE Intelligent Systems, 22(5), 84-88.

[7] Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 36(3), 306-323.

[8] Dingli, A. (2011). Annotation for the Semantic Web. In Knowledge Annotation: Making Implicit Knowledge Explicit (pp. 19-24). Springer Berlin Heidelberg.

[9] Cunningham, K. B. H. (2011). 3 Semantic Annotations and Retrieval: Manual, Semiautomatic, and Automatic Generation. Handbook of Semantic Web Technologies, 1.

[10] Crawley, S., Chernich, R., & Hunter, J. (2010, August). Beyond Annotea. IneResearch Australasia 2010.

[11] URL: http://www.w3.org/Amaya/ [Last accessed June 3, 2013].

[12] Heese, R., Luczak-Rösch, M., Oldakowski, R., Streibel, O., & Paschke, A. (2010, February). One click annotation. In Proceedings of the Workshop on Collaborative Construction, Management and Linking of Structured Knowledge (CK). Ed. by Tania Tudorache, Gianluca Correndo, Natasha Noy, Harith Alani, and Mark Greaves. CEUR Workshop Proceedings..: http://CEUR-WS.org (Vol. 514).

[13] Saleh, L. M. B., & Al-Khalifa, H. S. (2009, December). AraTation: an Arabic semantic annotation tool. In Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services (pp. 447-451). ACM.

[14] Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., & Goranov, M. (2003). KIM–semantic annotation platform. In The Semantic Web-ISWC 2003 (pp. 834-849). Springer Berlin Heidelberg.

[15] URL: http://www.ontotext.com/kim [Last retrieved June 5, 2013].

[16] Cunningham, H., Maynard, D., & Bontcheva, K. (2011). Text processing with gate. Gateway Press CA.

[17] URL: http://gate.ac.uk/ [Last accessed June 5, 2013].

[18] Hitzler, P., Krotzsch, M., & Rudolph, S. (2011). Foundations of semantic web technologies. Chapman and Hall/CRC.

[19] Jarrar, M. (2011, April). Arabic ontology engineering-challenges and opportunities. In Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications (p. 2). ACM.

[20] URL: http://protege.stanford.edu/ Source downloaded in May 21, 2013.

[21] URL: http://jena.apache.org/ Source downloaded in May 22, 2013.

[22] Pan, J. Z. (2009). Resource description framework. In Handbook on Ontologies (pp. 71-90). Springer Berlin Heidelberg.

[23] URL: http://www.w3.org/RDF/ [Last retrieved May 30, 2013].

[24] Beseiso, M., Ahmad, A. R., & Ismail, R. (2010). A survey of Arabic language support in semantic web. International Journal of Computer Applications, Vols.

[25] http://www.w3.org/TR/owl2-overview [Last accessed May 30, 2013].

[26] http://www.w3.org/RDF [Last retrieved September 30, 2013].

[27] Xiang, B., Nguyen, K., Nguyen, L., Schwartz, R. and Makhoul, J. 2006. Morphological Decomposition for Arabic Broadcast News Transcription. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006 Proceedings.

[28] Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2011). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, *2*(1).

Ahmed N. El-ghobashy graduated in 2009 and obtained his B. Sc. degree in computer engineering and information technology. His main interests include web development, web design, information architecture, and web-standards.

Gamal M. ATTIYA graduated in 1993 and obtained his M.Sc. degree in computer science and engineering from the Menoufia University, Egypt, in 1999. He received his PhD degree in computer engineering from the University of Marne-La-Vallée, Paris-France, in 2004. His main research interests include distributed computing, task allocation and scheduling, computer networks and protocols, congestion control, QoS, multimedia networking and Image processing.

Hamdy M. Kelash is professor in computer science and engineering department, Faculty of Electronic Engineering, Menoufia University, Egypt. His main research interests include computer vision, computer aided design, and Image processing.