International Journal of Intelligent Computing and Information Science

# AN APPROACH FOR AUTOMATIC ARABIC ONTOLOGY GENERATION

D. Fadl S. Abas M. Aref

Department of Scientific Computing, Faculty of Computer and Information Sciences,Ain Shams University,cAIRO
Dalia_sayed_43@hotmail.com safiaabas@yahoo.com[2]

***Abstract.*** *The increasing interest in Ontologies for many natural language applications in the recent years has led to the creation of ontologies for different purposes and with different feature systems. Manual ontology building is a time consuming activity that requires a lot of effort. In order to overcome these problems many methods have been developed to generate ontologies. There are various studies conducted on Arabic language in Semantic Web. This paper purposes an approach for the generation of an Arabic ontology from semi-structured data (xml document). This approach takes xml document and generates Arabic Ontology. First the system generates xml schema for the xml document. Using this schema it develops the xml schema graph XSG, translate this graph into Arabic and then start the generation of the ontology and the relations using the graph and the xml document. Finally, the system is going to evaluate the generated Arabic ontology using data driven ontology measures.*

***Keywords:*** *ontology learning, Arabic ontology, Natural language applications*

## 1. Introduction

Ontologies are of basic interest in many different fields, largely due to what they promise: they are a shared and common understanding of some domain that can be the basis for communication across the gaps between people and computer. Ontology approaches allow for sharing and reuse of knowledge bodies in computational form [1, 2]. Language is the way to vehicle information and knowledge. So the need for linguistic data is very important in all research fields. Arabic Language is the mother tongue for 23 countries and more than 350 million persons. Moreover, since it is the language of the Holy Quran, many other Islamic countries, like Pakistan, teach Arabic as a second language. Nevertheless, it is noticed that the Arabic content on the web is less than what should be expected. The evolution of the semantic web (SW) added a new dimension to this problem [3].

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 words that matches to its meaning can also result to the stems and branches of semantics [3]. The ontology development process is not a linear process but a refinement one where each activity can be repeated several times. To build the taxonomy of concepts, several approaches have been exposed in literature. In this paper Arabic ontology development life cycle from Semi-structured data is going to be discussed.

We present an approach to generate Arabic ontology from semi-structured data (xml file). The approach can be used to generate ontology to any language. It also provides automatic evaluation to the generated Arabic ontology using cosine similarity measures. The rest of this paper is organized as follows. Section 2 gives an overview on ontology learning process. Section 3 reviews the related work. The Arabic ontology framework is presented in section 4 in details. Section 5 presents a case study to show how the proposed approach works. Finally, section 6 concludes the paper and suggests new future research work.

## 2. Ontology Learning

The ontology learning process is useful for different reasons. First of all, it accelerates the process of knowledge acquisition. Second, it reduces the time for the updating of an existent ontology. Finally, it accelerates the whole process of ontology building [3, 4]. There are few types of ontologies which have different roles. In some cases, discussion goes to a mess because of the ignorance of what type of ontology is under consideration. Some say "ontology is domain-specific like a knowledge base". Others say "No, it isn't. Ontology is very generic and hence it is widely applicable and sharable". Both are correct because they are talking about different types of ontology

There are two main classes of Ontologies: the first would be the one that is employed to explicitly capture "static knowledge" about a domain. The second that provides a reasoning point of view about the domain knowledge (problem solving knowledge).

In the first class a distinction between types is made on the basis of the level of generality, as summarized below:

1. Domain Ontologies: Designed to represent knowledge relevant to a certain domain type, e.g. medical,
   mechanical etc.
2. Generic Ontologies: Can be applied to a variety of domain types.

3. Representational Ontologies: These formulate general representation entities without defining what should be represented. For the problem solving knowledge class, two types may be found:

1. Task Ontologies Provide terms specific for particular Tasks.
2. Method Ontologies Provide terms specific to particular Problem Solving Methods.

The absence of free usable lexical and syntactic resources and tools for Arabic makes it a "pi- language" (poorly informative). This constitutes a real difficulty in the process of transferring technology into Arabic. There is a strong need for Arabic language support [5, 6]. Ontology has 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 ontology tools and applications needs to be created to meet the requirements of the Arabic language and that of technologies. In Arabic ontology basic categorization of terminologies and meanings in a domain give the semantics. The interrelationship between 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 [7, 8].

## 3. Related work

### 3.1 Ontology learning from textual documents

The main functionality in this work is that the noun phrases appearing in the headings of a document as well as the document's hierarchical structure can be used to discover the concepts and is-a relation between them in the documents' domain [6]. They implemented and applied the system on a set of Arabic agricultural extension documents. The system takes as input a root concept, analyzes all input documents' heading structure, extracts concepts from headings and builds a taxonomical ontology. The resulting ontology was verified against a modified version of AGROVOC ontology, which is a hand-made ontology developed by Food and Agriculture Organization of the United Nation (FAO).

### 3.2 Building a framework for Arabic ontology learning

This paper presents the ArOntoLearn a Framework for Arabic Ontology learning from textual resources [9]. It supports the Arabic language and using domain knowledge or previous knowledge in the learning. It uses Probabilistic Ontology Model (POM) in the learned ontology, which can be translated into any knowledge representation formalism, and implements data-driven change discovery. Therefore it updates the POM according to the corpus changes only, and it helps user to trace the evolution of the ontology with respect to the changes in the underlying corpus. The system matches Arabic textual resources to Arabic Lexico-syntactic patterns in order to learn new Concepts and Relations. They developed a framework for incremental ontology learning, using Arabic natural language processing, machine learning and text mining techniques, in order to extract ontology from Arabic textual resources.

### 3.3 Arabic Word Net

Arabic Word Net (AWN) is a free lexical resource for modern standard Arabic [4]. It is based on the design and contents of Princeton WordNet (PWN) and can be mapped onto PWN as well as a number of other Word Nets, enabling translation on the lexical level to and from dozens of other languages. Moreover, the mapping of WordNet to the Suggested Upper Merged Ontology (SUMO) provides opportunities to use the semantic side in some Arabic NLP (natural language processing) applications. Constructing AWN presents challenges not encountered by established WordNet.

### 3.4 Building Ontologies form xml data source

They proposed a tool, called X2OWL that aims at building OWL (web ontology language) ontology from an XML data source [10]. This method is based on XML schema to automatically generate the ontology structure, as well as, a set of mapping bridges between the entities of the XML data source and the created ontology, mapping bridges contribute into query translation between OWL and XML. This approach addresses simple cases and complex cases that arise from the reuse of global types and elements that are used to create XML schema. XML schemas can be modeled using different styles.

## 4. The proposed Arabic Ontology Approach

In this section, the proposed Arabic ontology approach is going to be described. The approach generates an Arabic ontology from XML document. The proposed method uses the same notations used in [10]. It consists of four phases: Extraction, Xml schema parsing, Generation, Refinement and Evaluation. As shown in figure 1, the developed Arabic ontology approach takes a XML document and generates Arabic ontology. This XML document contains the information needed to generate the final ontology. XML schemas and ontologies in a given domain are somehow related. In general, schemas are built in a domain before ontologies. To benefit from preexisting schemas, the proposed method and tool is going to derive ontology (i.e., a concept hierarchy with concepts properties and main relationships) from a set of XML schemas.The aim is to provide a general view of the automation aspect of the ontology generation; thus, Deriving Ontologies from XML Schema. The proposed method uses the same notations used in [10] with some modifications to apply on Arabic xml data sources. The automatic ontology generation life cycle is defined as a process composed of four main steps necessary to achieve our goal. These steps represent the main tasks of the process for building ontologies starting from an existing corpus, like XML schemas. The approach mainly describes what is expected from each task. The whole process is depicted in figure 1. The four steps are:

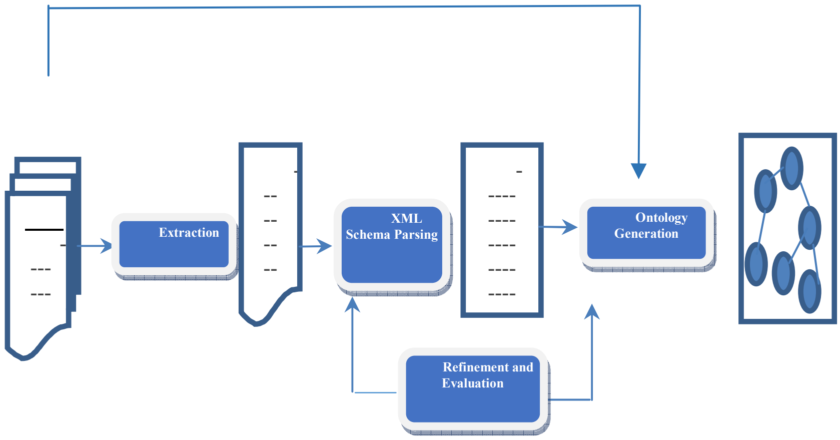

Figure 1. Arabic Ontology Framework

### 4.1 Extraction

This step deals with the acquisition of information needed to generate the ontology (concepts, attributes, relationships and axioms) starting from an existing corpus. Input resources are xml files. The approach goes through the XML file to extract the Arabic words and the tags connected to it. Then it will connect the tags extracted from every page to an xml document and creates xml schema for it. The approach may merge more than one xml document to each other to create our XML schema. The input selected from (http://www.saudiwildlife.com/site/home/) web site, on (10/1/2016).

### 4.2 Xml schema parsing

In this phase the input XML schema is going to be parsed. The result of the parsing is generated by two tasks. The first task is XML schema graph XSG. This graph is a tree which has one root, containing the complex types and elements which are structured in a hierarchic shape describing the document. The first version of the XSG will be in English language, the XSG will be translated into Arabic language. The second task is Auxiliary mapping of derived complex types which connects the classes with each other and this task is going to be used in next phases. After that, the information from the xml file will be retrieved to fill the ontology classes with information. Figure 2 shows the steps of this phase. The input to the ontology is selected from the source documents using similarity measurements. If the input matches with 80% of this measures this input will be used otherwise this input will be eliminated from our ontology.

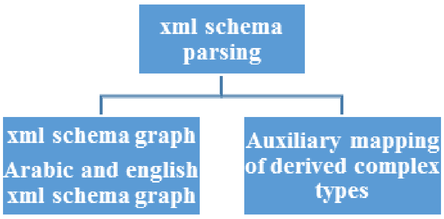
xml schema parsing
xml schema graph Arabic and english xml schema graph
Auxiliary mapping of derived complex types

Figure 2. XML schema parsing

## 4.3 Ontology Generation

This step deals with the production of a first version of the target ontology based on the tool formal meta-model, i.e., in a universal language interpretable by other applications, such as OWL and RDF/S. The created ontology is described in OWL language. The generated ontology is instance-free and only contains the description of concepts and properties. This process is based on some mapping rules that indicate how to convert each component of the XML schema to the corresponding ontology component. Figure 3 shows the three types of the generated ontology classes.

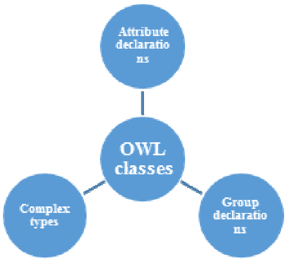

Figure 3. OWL classes

The algorithm that applies our suitable mapping rules on the XML schema to generate OWL ontology is shown in figure 4 below. The ontology generation process consists of the following steps:

1. Generation of classes.
2. Generation of properties and mapping rules.

**1) Complex types**
We can distinguish two kinds of complex types:
a. Global, named complex types.
b. Local anonymous complex types.
Both cases are mapped to OWL classes. However,
A class generated from a global named type will have the name of that type.
A class generated from local anonymous type will have the name of its surrounding element.
**2) Element groups and Attribute declarations are mapped to OWL classes.**
**3) Inheritance XML schema supports two types of inheritance: extension and restriction.**
Both of these inheritance mechanisms are translated to the class inheritance mechanism of OWL, using rdfs:subClassOf.
When a complex type is defined as an extension or a restriction of another base complex type, then the class corresponding to this type is set as subclass of the class corresponding to the base type.

Figure 4. Ontology generation algorithm

For the ontology generation, the vertices of the Arabic XSG are checked. A class C for complex type will be created, element group or attribute group vertices. Then an association (v;C) will be added between the generated class C and the vertex v to the auxiliary mapping classMap. In the case of complex type vertex, the name of the generated class C is the name of the complex type if it is global. However in case of local complex type, the name of the generated class C is that of its surrounding element (the source vertex of the incoming edge). In the case of an element group vertex and attribute group vertex, the name of the generated class C is the name of that element group or attribute group respectfully. After the creation of classes the generated ontology are going to be refined and evaluated.

## 4.4 Refinement and Evaluation

All previous steps may introduce wrong concepts or relationships, thus a refinement phase is needed. Conversely, a refinement task can be introduced at the end of each previous step. Validation is often done by hand, but can sometimes be automated. Ontology is not a static description of a domain, but with the evolution of applications, in quality and number, the ontology may also require some changes. Data driven evaluation will be used to evaluate the ontology generated. In this technique the ontology is evaluated by comparing it with existing data about the ontology domain. This existing data can by a corpus or the source documents of the ontology. To achieve this, one could, for example, perform automated term extraction on the documents and simply count the number of terms that overlap between the ontology and the documents [11].

There are a variety of ways in which one could attempt to extract the information content of the documents in order to correlate that with the ontology. In the evaluation phase Term-based distance measure is applied. A document is commonly represented as a vector of terms in a vector space model (VSM). The basis of the vector space corresponds to distinct terms in a document collection. Each vector represents one document. The components of the document vector are the weights of the corresponding terms that represent their relative importance in the document and the whole document collection [11]. The mutual information between two terms *t*1 and *t*2 can be calculated on the basis of the *ontology-based VSM*. Here, *cosine similarity* was adopted to measure the term mutual information between their corresponding vectors:

$$\begin{aligned}\cos(\angle(t_1, t_2)) &= \frac{t_1 \cdot t_2}{\| t_1 \| \cdot \| t_2 \|} \\ &= \frac{\sum_{j=1}^{n} \tilde{x}_{j1}\tilde{x}_{j2}}{\sqrt{\sum_{j=1}^{n} \tilde{x}_{j1}^2}\sqrt{\sum_{j=1}^{n} \tilde{x}_{j2}^2}}\end{aligned} \tag{1}$$

Where *~xj*1 and *~xj*2 represents the term weights of *t*1 and *t*2 in the document *X~j* in the *ontology-based VSM*. According to the above cosine measure, the similarity of each pair of terms in the given document can be computed. Distance measure with Term Mutual Information (*TMID*) since the mutual information between terms was calculated, it is better to take advantage of them in clustering process. The term mutual information can be expressed by a matrix called Mutual Information Matrix (*MIM*) [11]. The similarity measures between the input xml documents and the selected terms for the ontology generation applied.

## 5. Mammals Case study

In the case study, an xml documents contain information about mammals are going to be used. The first phase in our approach is extraction. In the extraction the tags containing the Arabic words in the every xml webpage will be selected. Then an xml document will be created to connect these tags. Finally the XML schema for the xml file will be generated. Figure 6 show the XSG for the insectivores xml file.

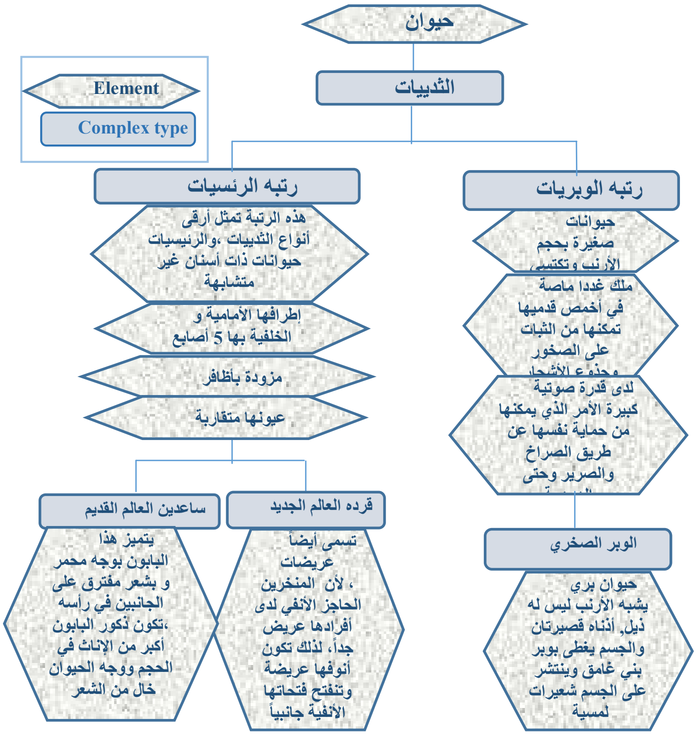
حيوان
Element
Complex type
الثدييات
رتبه الرئسيات
رتبه الوبريات
هذه الرتبة تمثل أرقى أنواع الثدييات ،والرئيسيات حيوانات ذات أسنان غير متشابهة
إطرافها الأمامية و الخلفية بها 5 أصابع
مزودة بأظافر
عيونها متقاربة
حيوانات صغيرة بحجم الأرنب وتكتسي
ملك غددا ماصة في أخمص قدميها تمكنها من الثبات على الصخور وجذوع الأشجار
لدى قدرة صوتية كبيرة الأمر الذي يمكنها من حماية نفسها عن طريق الصراخ والصرير وحتى
ساعدين العالم القديم
قرده العالم الجديد
الوبر الصخري
يتميز هذا البابون بوجه محمر و بشعر مفترق على الجانبين في رأسه ،تكون ذكور البابون أكبر من الإناث في الحجم ووجه الحيوان خال من الشعر
تسمى أيضاً عريضات ، لأن المنخرين الحاجز الأنفي لدى أفرادها عريض جداً، لذلك تكون أنوفها عريضة وتنفتح فتحاتها الأنفية جانبياً
حيوان بري يشبه الأرنب ليس له ذيل, أذناه قصيرتان والجسم يغطى بوبر بني غامق وينتشر على الجسم شعيرات لمسية

Figure 6. XSG for XML file

The second phase in our model is xml schema parsing which contain two steps the first step is xml schema graph XSG which will be generated in English and translated into Arabic for our schema as shown is figure 7 and figure 8. Our mammals case study it contains 6 complex type classes, 10 element and the Auxiliary mapping of derived complex types. In the parsing phase all the terms and relations that fit with the ontology will be extracted. The extracted terms go through similarity measures to make sure that they fit with our ontology. If the input matches with 80% of this measures used, this input will be eliminated from the ontology.

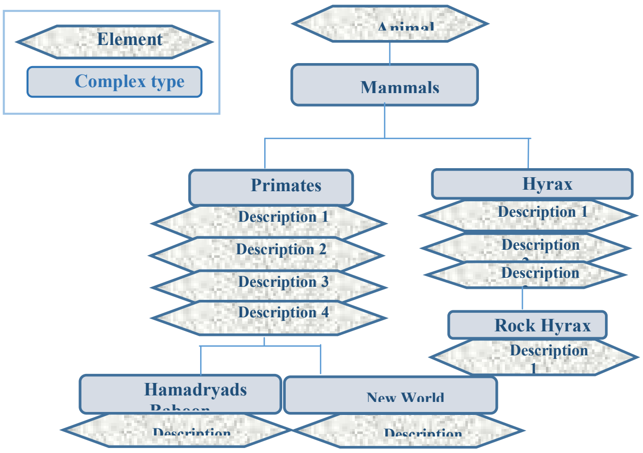

Figure 7. English XSG of the running insectivores' animal case study

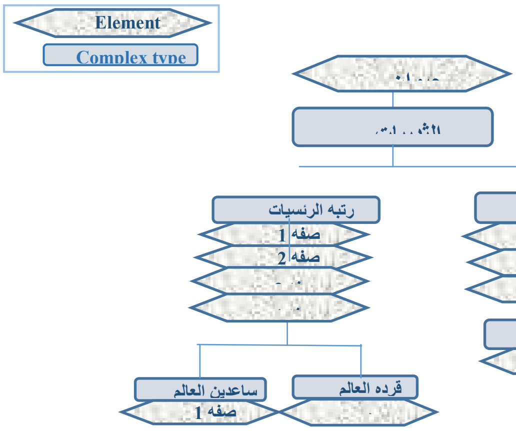

Figure 8. Arabic XSG of the running insectivores' animal case study

In the generation phase, OWL classes are generated according to the rules that introduced in figure 4. This step is based on the Arabic XSG and the derived Types Map. That is, the Arabic XSG is traversed to retrieve complex type vertices, for each one of these vertices, an OWL class is generated. Then, the derived Types Map is used to set up sub Class Of relationships between classes. The case study contains 6 complex types. Six classes will be generated to represent these complex types and indicated the sub class of relationship, the object properties and data type properties. The final output of the approach is an Arabic ontology with its relations and properties like in figure 9 below. The generated ontology is going to be evaluated using data driven evaluation method. In the method the output of our ontology are going to be compared with the source document using equation (1). After applying the similarity measure on our case study the best obtained result was 70%. The evaluation was done based on the cosine similarity measure function. The Arabic ontology case study contained 70% of the source xml document. This result can be refined more than one time to reach satisfying results.

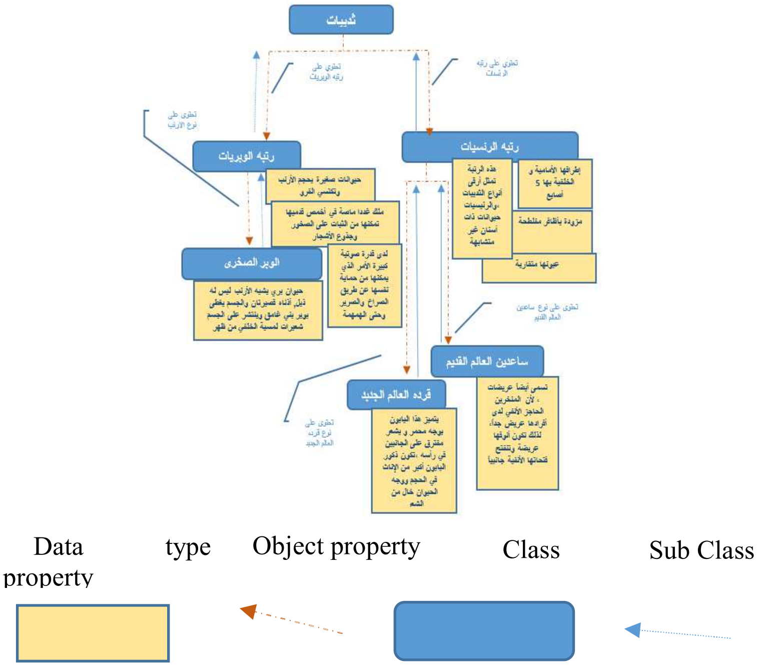

Figure 9. Mammals' ontology

## Conclusion

The increasing interest in Ontologies for many natural language applications in the recent years has led to the creation of ontologies. Ontology for linguistic purpose is very important to enrich languages. The Arabic language is the language of millions of people around the world and yet the work in Arabic ontology is very poor and facing a lot of problems. In our study we are trying to improve the Arabic information retrieval on the web. The proposed Arabic ontology approach from semi-structured data (xml document) was introduced. The approach takes xml document and generates Arabic ontology. The proposed approach tried to create automate the process of creation Arabic ontology. The output of the ontology was evaluated using data driven evaluation and achieved acceptable similarity measures. This paper is part of an ongoing research to develop a framework for building an Arabic ontology. Further work to enhance the similarity measures for the evaluation and apply to different case studies would be beneficial for a more a robust ontology approach.

## References

1. P.Saariluoma , K. Nevala, “From Concepts to Design Ontologies”, Cognitive Science, University of Jyväskylä, Finland, 2009
2. C.e Roche, “ONTOLOGY: ASURVEY”, University of SavoieEquipeCondillac - Campus Scientifique,73 376 Le Bourget du Lac cedex – France,2002
3. C. Brewster, F. Ciravegna, and Y. Wilks. “User-Centred Ontology Learning for Knowledge Management”. In Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers,2002
4. Sojka, Choi, Fellbaum and Vossen eds. “Introducing the Arabic WordNet Project”, in Proceedings of the Third International WordNet Conference, 2006
5. L. Al-Safadi, M. Al-Badrani, M. Al-Junidey, “Developing Ontology for Arabic Blogs Retrieval”, International Journal of Computer Applications (0975 – 8887)Volume 19– No.4, April 2011
6. Maryam Hazman, Samhaa R. El-Beltagy, Ahmed Rafea “Ontology Learning from Textual Web Documents”, INFOS2008, March 27-29, 2008 Cairo-Egypt.
7. Aya M. Al-Zoghbyaa, Ahmed Sharaf Eldin Ahmedb and Taher T. Hamzac “Arabic Semantic Web Applications – A Survey” 2013 Journal of Emerging Technologies in Web Intelligence, Vol 5, No 1 (2013), 52-69, Feb 2013
8. Zhan Cui, Dean Jones and Paul O’Brien “Intelligent Business Systems Research Group Intelligent Systems Lab BTexact Technology Issues in Ontology-based Information Integration” , 2002
9. N. Ghneim, W. Safi, M. Al Said Ali,” Building a Framework for Arabic Ontology Learning”, Damascus University, Damascus, Syria, 2008.
10. R. Ghawi, and N. Cullot, "Building Ontologies from XML Data Sources", In 1st International Workshop on Modelling and Visualization of XML and Semantic Web Data (MoViX '09), held in conjunction with DEXA'09 (Linz, Austria, September 2009).
11. Maedche, Alexander and Steffen Staab, 2002. “Measuring similarity between ontologies. In Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web”, Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02, volume 2473 of LNAI. Berlin: Springer Verlag.

12. Maryam Hazman, Samhaa R. El-Beltagy, Ahmed Rafea "A Survey of Ontology Learning Approaches", International Journal of Computer Applications (0975 – 8887)Volume 22– No.9, May 2011.
13. M. Attia, M. Rashwan, A. Ragheb, M. Al-Badrashiny, H. Al-Basoumy, "A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields".2008
14. Maha Al-Yahya*, Hend Al-Khalifa , Alia Bahanshal, Iman Al-Odah and Nawal Al-Helwah " an ontological model for representing semantic lexicons: an application on time nouns in the holy QURAN", *The Arabian Journal for Science and Engineering, Volume 35, Number 2C, December 2010.*
15. HassinaAliane, ZaiaAlimazighi, Ahmed CherifMazari, "Al - Khalil: The Arabic Linguistic Ontology Project. In proceeding of: Proceedings of the International Conference on Language Resources and Evaluation", LREC 2010, 17-23 May 2010, Valletta, Malta
16. Hend S. Al-Khalifa, Areej S. Al-Wabil, "The Arabic language and the semantic web: Challenges and opportunities", The 1st International Sysmposium on Computers and Arabic Language & Exhibition 2007.

.