# *Advanced Search in the Qur'an using Semantic modeling*

Aimad Hakkoum
FSTG, Cadi Ayyad University
Marrakesh, Morocco
imad.hakkoum@ced.uca.ma

Said Raghay
FSTG, Cadi Ayyad University
Marrakesh, Morocco
s.raghay@uca.ma

***Abstract*—The Qur'an is the religious text of Islam, distinguished by its miraculous style, it is considered as the basic reference for all Islamic sciences, and therefore it's very sensitive to model its content for fear to make bad assumptions and axioms. In recent years a number of researches has been done to facilitate the retrieval of knowledge from the Qur'an, but most of the available researches are using human readable data resources and therefore cannot be reused and linked using semantic web technologies, this is why in this project we will adopt an approach that enables humans and computers to understand the Qur'an knowledge throughout the creation of a Qur'anic ontology. The goal of the ontology is to build a computational model capable of representing as much as possible of the concepts mentioned on the Qur'an and the relationships between them using Protégé-OWL. The ontology can be queried using SPARQL queries, we also developed a search engine that will parse user's questions by extracting and stemming the keywords, and then it generates SPARQL queries using the concepts defined by the ontology.**

***Keywords: Qur'an, semantic web, ontology, knowledge extraction***

## I. INTRODUCTION

Ontology is one of the emerging specialty of research in computer science and semantic web, it can be defined as «an explicit specification of a conceptualization» [1]. Ontologies explicitly structure and represent domain knowledge in a machine-readable format so they can be incorporated into computer-based applications and systems to facilitate automatic annotation of web resources, domain representation and reasoning task, decision support, and natural-language processing and serve as an integral part of the Semantic Web [2].

In recent years the Qur'an was the subject of numerous researchers in the field of computer science, most of them are taxonomy, hierarchy or tree structure to present and classify the Qur'an knowledge, these approaches still effective to answer most of user's queries but cannot be reused and linked using web semantic technologies, this is why in this project we will adopt an approach that enables humans and computers to understand the Qur'an knowledge throughout the creation of a Qur'anic ontology. The ontology will be created using Protégé, we will also use Jena Framework to manipulate and query the ontology, both of these tools support Arabic language to write and display RDF data [3].

We will start by presenting the existing work done so far in the knowledge representations of the Qur'an, we will focus on the researches that are going to be used to achieve our task. After that we will discuss the methodology used to extract and model the content of the concepts mentioned on the Qur'an and the relationships between them, then we will present the advanced search tool, and finally we will address what left to be done in this project.

## II. RELATED WORK

### *A. Ontology-based researches*

Semantic web technology is still lacking a critical mass of RDF data online and up-to-date terms and ontologies are missing for many application domains especially for Islamic sciences, in order to address this lack the Qur'an became in recent years to target of interest for studies in the field of Semantic web technologies. Because of the complexity of the task and the necessary time to extract all the knowledge contained on the Qur'an, several researchers tried to cover a specific topic from the Qur'an like prayer (Salat)[4], faith (Iman) and deed (Akhlaq) [5] or Umrah [6].

One of the important ontologies developed so far in the subject is the Semantic Qur'an dataset[7] , the namespace of the ontology is QVOC (Qur'an vocabulary), and it consists of a multilingual RDF representation of translations of the Qur'an. The resulting RDF data encompasses 43 different languages. The Semantic Qur'an dataset is published at http://datahub.io/dataset/semanticquran, it contains over 15 million RDF triples, the Qur'an words are linked to 7718 word from dbpedia and 18655 from wikitionary.

Another useful project is the Qur'anic Arabic Corpus (QAC) and discussed in Kais Dukes PhD Thesis [8], it's an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Qur'an, it contains also an ontology of 300 concepts, it was developed using Knowledge Interchange Format (KIF) who is a standard to describe knowledge among different computer systems to facilitate its exchange, KIF is less used than OWL to create ontologies according to [9]. This ontology was translated to OWL and enhanced by designing more relationships and restrictions using sources from the Qur'an, hadith and Islamic websites[10].

### *B. Text mining researches*

The web contains a lot of resources that provide the Qur'an text and other Islamic books in different formats (HTML, Text, SQL dump and XML) and enables to do keyword search in an advanced way using lemmas, roots, word proximity and Boolean search, these resources are available in different languages especially in Arabic and English.

*1) Tanzil Project*

The Tanzil Project[1] was launched in early 2007 with the aim of producing a curated Unicode version of the Arabic Qur'an text that can serve as a reliable standard text source on the web. The Qur'an text from Tanzil project is widely used by a number of web sites and groups, and it was validated by different entities like "King Fahad Quran Complex"[2].

*2) The Qur'an Annotation for Text Mining*

This resource[3] contains several useful tools to understand the Qur'an, especially two important ones that we will use in our work. The first one is QurAna [11] which annotate the antecedent of every pronoun in the Qur'an, it relays on QAC to extract all the pronouns from the Qur'an then they perform a Manuel annotation for over a year, and finally the result was put on the QAC website for further validation by users. The results are available at www.textmaining.com.

The other valuable tool from this project is QurSim [12] which provides a dataset of related verses, it was based not only on common words or roots but also on Ibn Kathir commentary (Tafsir) of the Qur'an where he cited some relative verses when commenting on a verse.

*3) Qurany*

This research[13] proposes a tool[4] that categorize the topics discussed in the Qur'an verses to a comprehensive index that covers nearly 1100 topics in the Qur'an, it classifies the Qur'an into fifteen main themes and subdivides the main themes into sub themes and sub sub themes and so on.

## III. ONTOLOGY DEVELOPMENT

At the present time there is no consensus on the best practices to follow when developing an ontology. There are more than 33 methods of ontological engineering [14]. There is obviously no method which is the best. We are going to follow the methodology discussed in [15] by adopting an iterative approach to ontology development with the six steps: define the ontology domain and scope, review existing ontologies, enumerate important terms in the ontology, define the classes and the class hierarchy, define the properties of classes and finally create instances.

### *A. Domain and scope*

The ontology will cover the Qur'an knowledge, the ontology must allow semantic indexing of the Qur'anic content and the relation between the extracted concepts.

We will cover the following subjects: Qur'anic chapters and verses, each word of the Qur'an and its root and lemma, we will not cover words morphology but we will add links to QVOC ontology where it's covered, however we will cover the pronouns in order to define their antecedents. In a second step we will add additional concepts like: people, events and places cited on the Qur'an.

### *B. Ontology reuse*

We are going to reuse the two Qur'anic ontologies: Semantic Qur'an (QVOC) and QAC ontology with the OWL format. We can see ontology reuse according to two different points of view: building an ontology, by assembling, extending, specializing and adapting, other ontologies, or building an ontology, by merging different ontologies on the same or similar subject into a single one that unifies all of them [16].

### *C. Enumerate important terms in the ontology*

There are two approaches to create the Qur'an ontology: verse by verse extraction and topic extraction.

In verse by verse extraction we have to analyze each verse and build the ontology progressively in a linear way, this will be an incredibly time consuming process and it will require that we cover all the Qur'an otherwise the resulting model will be incoherent because we will cover a fragment of each subject and it won't be very useful.

The other approach is to cover only some topics by only analyzing their related verses, this way after adding a topic the ontology will be in a coherent form and can be used and published.

We will use the second approach with the following topics: chapters, verses, words, verse topics, pronouns antecedents, people, events and places cited on the Qur'an.

### *D. Define the classes and the class hierarchy*

We used Protégé and OWL to create the ontology because it's well maintained and contains a number of useful plugin that we can add to facilitate reasoning tasks and visualizing the model using diagrams and matrices.

The different classes created in the model are as follow:

TABLE1: ONTOLOGY CLASSES DESCRIPTION

| Class name | characteristics |
|---|---|
| Topic | Represents a topic discussed in a Verse |
| Chapter | Represents a chapter in the Qur'an |
| Verse | Represents a Verse in a chapter |
| Word | Represents a term in a verse, a word can be composed of several parts |
| PronounRef | Represents the relation between a pronoun and its reference. |
| Living Creation | Contains the following sub classes: Human, Angel and Jinn |
| Location | Contains the following sub classes: Geographical Location, Afterlife Location, Divine Location and Construction. |

### *E. Define the properties of classes and their facets*

We defined the relation between the ontology classes using object properties as described in the following diagram:

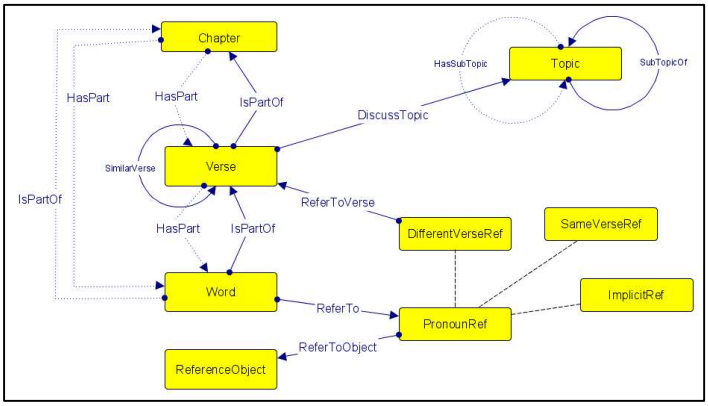


Fig1: Ontology classes and object properties

The properties using dotted lines are obtained using inference, meaning that we won't have a corresponding triple in the data but it will be calculated by the reasoning tool from other triples.

[1] http://tanzil.net/wiki/Tanzil_Project

[2] http://www.qurancomplex.org/

[3] http://www.textminingthequran.com/wiki/Main_Page

[4] http://quranytopics.appspot.com

### F. *Create instances*

To extract the instances of our ontology we used an automatic approach for the concepts: Chapter, verse, word, pronoun and topic, by extracting and validating data from the sources described in the second chapter, these sources were from different formats: OWL, XML and Text. For each extraction we used a program that parsed, validated and transformed the source data to RDF triples.

For the other concepts (Location, Living Creation and Event) we will use a manual approach using the previously cited research and especially the Tafsir books.

TABLE2: ONTOLOGY CLASSES STATISTICS

| class | Instances count |
|---|---|
| Topic | 1181 |
| Chapter | 114 |
| Verse | 6236 |
| Word | 77430 |
| PronounRef | 24674 |
| ReferenceObject | 1028 |
| Living Creation | 110 |
| Location | 65 |

## IV. EXPLOITING THE RESULTS

Protégé cannot load a big file of RDF triples, so we have to store the ontology in a triple database, there is already a great number of RDF triple store that support SPARQL query language [17], we will use Jena TDB with Fuseki server. After loading the data into Jena TDB, we can issue SPARQL queries against the database by using the server user interface or by developing a program using JENA API and the SPARQL ENDPOINT if we want to do more processing with the result, here is a sample query that the model can answer:

- Get the verses (top 5) that discuss the topic of Zakat without having the string "زكاة" in the verse text:

SPARQL query:

```
SELECT ?verse  ?text
WHERE {?verse rdf:type qur:Verse.
?verse qur:DiscussTopic ?topic.
?verse qur:SearchText ?text.
?topic qur:TopicCompleteName ?topicName.
FILTER (REGEX(STR(?topicName), "زكاة", "i")).
MINUS { ?verse qur:SearchText ?text.
   FILTER (REGEX(STR(?text), "زكاة", "i")).}}
```

Result:

| verse Id | Verse Text |
|---|---|
| quran63-10 | وأنفقوا من ما رزقناكم من قبل أن يأتي أحدكم الموت فيقول رب لولا أخرتني إلى أجل قريب فأصدق وأكن من الصالحين |
| quran57-18 | إن المصدقين والمصدقات وأقرضوا الله قرضا حسنا يضاعف لهم ولهم أجر كريم |
| quran2-263 | قول معروف ومغفرة خير من صدقة يتبعها أذى والله غني حليم |
| quran64-16 | فاتقوا الله ما استطعتم واسمعوا وأطيعوا وأنفقوا خيرا لأنفسكم ومن يوق شح نفسه فأولئك هم المفلحون |
| quran64-17 | إن تقرضوا الله قرضا حسنا يضاعفه لكم ويغفر لكم والله شكور حليم |

We can see that other words are used to speak about the topic of "Zakat" like: "أنفقوا", "صدقة", "تقرضوا الله"

## V. ADVANCED SEARCH TOOL

The advanced search tool will allow non-technical users to use the ontology without writing SPARQL queries, it will parse and analyze the user question written in Arabic, and build SPARQL queries using the ontology concepts and properties, Fig2 describes the different steps of the search algorithm:

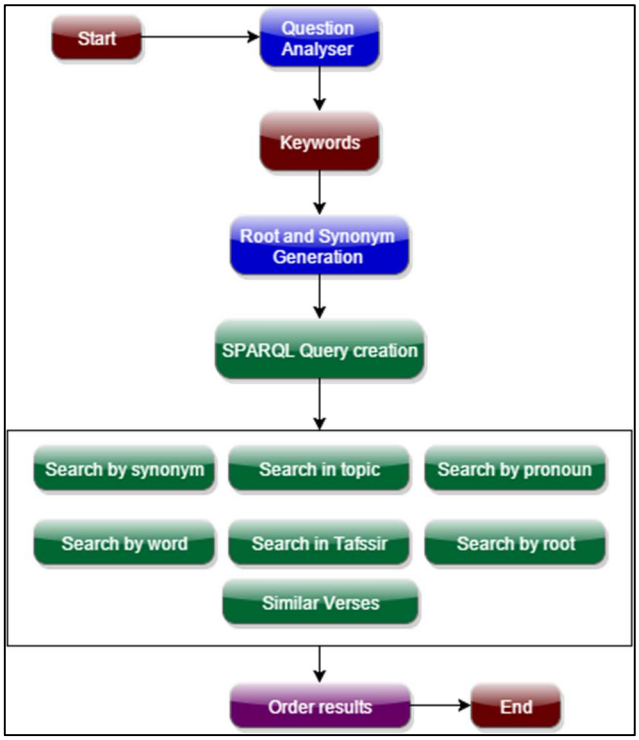


Fig2: Advanced Search algorithm

### A. *Question analyzer*

In this step we will parse the user query and stem each word using the stemmer developed by S.Khoj[18], Stemming is the process of removing all of a word's affixes to produce the stem or root.

The stemmer also detects stop words like "لكن, وهو, حتى …" stop words are very common words that appear in the text and do not carry an important meaning, so they will be ignored on the query builder.

### B. *Synonym generation*

For each keyword we will generate a set of synonym using the synonym dictionary at "AlMaany" and "SensAgent".

### C. *SPARQL query creation*

We will run in parallel the following queries:

- Search by word: search in the verse text using the keywords, we will search at first using all the keywords and then with one keyword at a time.
- Search by root: search using the word root, this will allow us to find the different forms of the word.
- Search by synonym: search in the verse text using the keywords synonyms.
- Search in topics: search in the topic text using the keywords and the synonyms.
- Search by pronoun: search in the pronouns references using the keywords.
- Search by Taffsir: search in the Taffsir (Jalalayn and Muyassar) of the verse.

We will get the top 100 verse for these queries and then add their related verses to have more relevant information.

### D. *Evaluation*

In order to evaluate the relevance of the search tool, we will start by analyzing a sample search query, then we will execute a set of statements and then evaluate the results by comparing them with other research tools in the Qur'an.

For the sample query we will use a two words query: «سفينة نوح» (noah's ark), we get the following result:

TABLE3: RESULT OF THE SEARCH QUERY «سفينة نوح»

| | |
|---|---|
| Search by word | 53 results with only one word |
| Search by root | 50 results with only one word |
| Search by synonym | 0 result |
| Search in topics | 21 results with only one word |
| Search by pronoun | 11 results with two words<br>Example: " وَقَالَ ٱرۡكَبُواْ فِيهَا بِسۡمِ ٱللَّهِ مَجۡرٖىٰهَا وَمُرۡسَىٰهَآۚ إِنَّ رَبِّي لَغَفُورٞ رَّحِيمٞ" |
| Search by Tafsir | 21 results with two words:<br>Example: " وَءَايَةٞ لَّهُمۡ أَنَّا حَمَلۡنَا ذُرِّيَّتَهُمۡ فِي ٱلۡفُلۡكِ ٱلۡمَشۡحُونِ " |

By analyzing the result we find that the query by pronoun and by Tafsir gives the most relevant result because the search by word or by root find the verses that contains only one word of the query therefore some of the returned verses talks about different subjects.

To see the benefit of the semantic approach, we will execute 5 different search queries and compare them with other search engines:

TABLE4: SEARCH RESULT BY QURAN SEARCH ENGINES

| Query / Search engine | سفينة نوح | السماوات والجبال | أهوال يوم القيامة | الزبور | تحريف التوراة |
|---|---|---|---|---|---|
| Our model | 50 | 105 | 120 | 16 | 44 |
| quranicresearcher | 0 | 0 | 0 | 1 | 0 |
| altafsir.com | 7 | 2 | 8 | 16 | 1 |
| quran.ksu.edu.sa | 0 | 0 | 0 | 1 | 0 |
| alawfa.com | 50 | 191 | 451 | 1 | 16 |
| quran.com | 22 | 0 | 90 | 11 | 2 |
| alfanous.org | 49 | 184 | 221 | 7 | 14 |

Quranicresearcher and quran.ksu.edu.sa are based on exact word search so they don't find any result when we use a query with multiple words, whereas Altafsir.com uses in addition of searching on Quran text it looks into different Quran explications.

Quran.com uses different forms of the words and it looks always for all the words using the Qur'an text and a large set of Quran explications, and finally Alfanous.org allows to include different forms and synonyms of the word on the search process.

We note that when we search with multiple words, we don't often find verses with all the keywords, in the other hand searching by pronoun, topic and Tafsir is more precise and returns the most accurate results.

## VI. PUBLISHING THE RESULTS

We chose the name "QuranOntology" as a name for our ontology, and we published it at the following URL: http://quranontology.com.

We used Jena TDB with Fuseki server as a database for the ontology and .NET framework, MVC and HTML5 to develop the web site.

The web site allows users to browse the content of the ontology as well as to use the advanced search tool.

## VII. CONCLUSION AND FUTURE WORK

The study of the existing work done on knowledge extraction from the Qur'an showed us that this subject is currently the point of interest of different research groups but until now there is no global ontology that represent the knowledge contained in the Qur'an.

In this work we created a Quran ontology that encompasses a set of concepts and the relations between them using OWL, this ontology can be used to answer complex queries and to describe about 11000 resources using over 1 million RDF triple.

We also developed an advanced search engine that takes advantage of the different concepts defined by the ontology to return the most accurate answer for a user query.

The next step of our research is to extract more concepts and knowledge from the Qur'an and to enhance the web site to allow users to interact with the ontology by adding new content to the ontology.

## REFERENCES


[1] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? International journal of human-computer studies, 43(5), 907-928.

[2] Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. Intelligent Systems, IEEE, 21(3), 96-101.

[3] Majdi Beseiso, Abdul Rahim Ahmad and Roslan Ismail. Article: A Survey of Arabic language Support in Semantic web. International Journal of Computer Applications 9(1):35–40, November 2010. Published By Foundation of Computer Science

[4] Saad, S., Salim, N., & Zainal, H. (2010). Towards context-sensitive domain of Islamic knowledge ontology extraction. International Journal for Infonomics (IJI), 3(1), 197-206.

[5] TA'A, Azman, ABIDIN, Syuhada Zainal, ABDULLAH, Mohd Syazwan, et al. AL-QURAN THEMES CLASSIFICATION USING ONTOLOGY. In: Proceedings of the 4thInternational Conference on Computing and Informatics, ICOCI. 2013. p. 28-30.

[6] Sharef, N.M., Murad, M. A. A., and Mustapha, A., Shishehchi, S., "Semantic Question Answering of Umra Pilgrims to Enable Self-Guided Education", 13th International Conference on Intelligent Systems Design and Applications (ISDA 2013), pp. 141-146, Kuala Lumpur

[7] Sherif, Mohamed Ahmed, Ngonga Ngomo, Axel-Cyrille: Semantic Quran: A Multilingual Resource for Natural-Language Processing. In: Semantic Web Journal (2014), S. 1-5

[8] Kais Dukes (2013). Statistical Parsing by Machine Learning from a Classical Arabic Treebank. PhD Thesis. University of Leeds.

[9] Marchetti, A., Ronzano, F., Tesconi, M., & Minutoli, M. (2008). Formalizing Knowledge by Ontologies: OWL and KIF. Relatório apresentado L'Istituto di Informatica e Telematica (IIT). Consiglio Nazionale delle Ricerche (CNR). Italia.

[10] Aliyu Rufai Yauri, Rabiah Abdul Kadir, Azreen Azman and Masrah Azrifah Azmi Murad, 2013. Quranic Verse Extraction base on Concepts using OWL-DL Ontology. Research Journal of Applied Sciences, Engineering and Technology, 6(23): 4492-4498.

[11] A.M. Sharaf and E. Atwell, "QurAna: Corpus of the Qur'an annotated with Pronominal Anaphora", in Proc. LREC, 2012, pp.130-137.

[12] Sharaf, A. B. M., & Atwell, E. (2012). QurSim: A corpus for evaluation of relatedness in short texts. In LREC (pp. 2295-2302).

[13] Abbas, N. H. (2009). Quran Search for a Concept Tool and Website. PhD Thesis. University of Leeds.

[14] Psyché, Valéry, Olavo Mendes, and Jacqueline Bourdeau. "Apport de l'ingénierie ontologique aux environnements de formation à distance." Revue des Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF) 10 (2003): 89-126.

[15] NOY, N. F. (2001). Ontology Development 101: A Guide to Creating Your First Ontology: Knowldege Systems Laboratory, Stanford University. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

[16] Pinto, H. S., & Martins, J. P. (2000). Reusing ontologies. In AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes (Vol. 2, No. 000, p. 7). AAAI Press.

[17] Stegmaier, Florian, et al. "Evaluation of current RDF database solutions." Proceedings of the 10th International Workshop on Semantic Multimedia Database Technologies (SeMuDaTe), 4th International Conference on Semantics and Digital Media Technologies (SAMT). 2009.

[18] Khoja, Shereen. "APT: Arabic part-of-speech tagger." Proceedings of the Student Workshop at NAACL. 2001.