
Proceedings Paper:
Alqahtani, M orcid.org/0000-0001-9403-1286 and Atwell, E orcid.org/0000-0001-9395-3764 (2018) Developing Bilingual Arabic-English Ontologies of Al-Quran. In: Proceedings of ASAR'2018 Arabic Script Analysis and Recognition. ASAR'2018 Arabic Script Analysis and Recognition, 12-14 Mar 2018, Alan Turing Institute, The British Library, London UK. IEEE , pp. 96-101.

Developing Bilingual Arabic-English Ontologies of Al-Quran

1st Mohammad M Alqahtani
Faculty of Computing and Information Technology
University of Jeddah
Jeddah, Saudi Arabia
scmmal@leeds.ac.uk

2ed Eric Atwell
School of Computing
University of Leeds
Leeds, United Kingdom
E.S.Atwell@leeds.ac.uk

***Abstract*—The main aim of developing a Quranic ontology is to facilitate the retrieval of knowledge from Al-Quran. Additionally, Quranic ontologies will enrich the raw Arabic and English Quran text with Islamic semantic tags. However, current Quran ontologies have different: scopes, formats, and entity names for the same concepts. Additionally, a single Quranic ontology does not cover most of the knowledge in Al-Quran. Therefore, these ontologies need to be increased, normalised, aligned and combined with other Quran resources such as Quran chapter and verse names, Quran word meanings, and other Quranic datasets. This paper reviews current Quran ontologies and datasets. Then, it presents several stages for developing Arabic-English Quran ontologies from different datasets related to Al Quran.**

Keywords—arabic ontology; semanitc taging; Al Quran;

I. INTRODUCTION

The Holy Quran is the most important resource for the Islamic sciences and the Arabic language. Many research studies have been built on ontologies to facilitate the retrieval of knowledge from Al-Quran. The term ontology is defined as an explicit specification of concepts, attributes and relations in a domain [1]. Common components of ontologies include classes (concepts), attributes, relations, function terms, restrictions, and axioms. These concepts are the entities of interest in a domain. They are structured into a taxonomy tree or un-taxonomy tree. Each tree node represents a concept that is a specialisation of its ancestor. The concept is related to a set of instances. Additionally, it is given a set of attributes. Relations refer to the ways in which concepts and instances can be connected.

An ontology can be evaluated against several criteria, such as the coverage of a certain domain and the size of the ontology. Additionally, ontologies can also be assessed in terms of the specific use cases, scenarios, requirements, applications and scope. This evaluation includes the consistency and completeness of the ontology and the representative modelling language. Moreover, the assessment of ontology covers the feasibility of alignment the ontology with other ontologies and improvements [2]. Ahmad [3], and Alrehaili [4] compared existing Quranic ontologies against nine and six criteria respectively such as number of concepts, availability, relation type, verification methods, coverage area, maturity level, and underlying format. These surveys concluded that these ontologies have unclear consensus on semantic annotation format and validation methods.

Therefore, these ontologies need alignment and normalisation. Ontology alignment is a process of finding one to one correspondence via the entities of both ontologies. The primary goal of ontology alignment is to integrate different ontologies of the same domain [5].

Moreover, the primary objective of merging Quranic Ontologies is to pioneer research enriching the raw Arabic Quran text with Islamic ontology. Additionally, this combined ontology might be used in semantic search tools to answer questions about Al-Quran. Moreover, aligning the Quranic ontologies will increase the coverage of the domain of Al-Quran in various capacities. Furthermore, the alignment will enhance the knowledge extraction from Al-Quran.

Three modules are used to align Quranic ontologies: normalisation, terminological approach and structural approach [6]. In normalisation process, all ontologies are reformatted to have the same file format. Terminological techniques are divided into string based and language based approaches. String based matches entities based on the similarity between letters in the two entities, for instance, author and authority are more similar than author and writer. However, the language-based technique aligns two entities that share the same meaning, for instance, paper and article. On the other hand, the structural approaches detect correspondences between entities depending on the internal structure of the entity and how it is connected to other entities. In other words, the structural method matches entities based on the ontology graph. Most of the existing alignment tools exploit terminological techniques as the initial step and then use the structural techniques to improve the outcomes ontology.

This paper aims to review the majority of ontologies and datasets that have been constructed for the Holy Quran.

This document is organised as follows. Section II is Literature review of qur'anic ontologies. Section III a methodology of aligning and combining Quranic Ontology. Finally, Section IV concludes the critical points in this paper.

II. Research Conducted on Quran Ontology Selecting

Sherif and Ngonga Ngomo [7] developed a Semantic Quran dataset in an Resource Description Framework[1] (RDF) format representing forty-two different Al-Quran translations. This dataset was built by merging data from two different semi-structured sources: the Tanzil project and the Quranic Arabic Corpus. An ontology of the Semantic Quran was constructed to demonstrate various multilingual data from Quranic sources with a hierarchical structure, which is a chapter, a verse, a word, and a lexical item. This ontology has 7,718 links to DBpedia[2], 18,655 links to Wiktionary, and 15,741,399 triples.

Khan [8] developed an ontology for Al-Quran in based upon the animals found in Al-Quran. This ontology was constructed using the Protégé. SPARQL was then used to search through it. This ontology provides 167 links to animals in Al-Quran, based on information from the book, "*Al-Hayawany Fi Al-Quran Al-Kareem*" [9].

Yauri [10] rebuilt the existing ontology created by Dukes [11] using the Protégé tool and Manchester OWL. He increased the number of relationships from 350 to approximately 650 based on Al-Quran, the Hadith and some online Islamic resources. This ontology covers some subjects that mentioned in Al-Quran, such as food, people, religions and life.

Yahya [12] created a bilingual ontology featuring the English and Malay languages, which was also based on that developed by Dukes. In the translation of Al-Quran into Malay language, 5,999 verses are assigned to concepts of Quranic ontology, while 237 verses are unrelated to any concepts. In the English translation, 5,695 verses related to concepts in this ontology, whereas, 541 documents were not allocated to any concepts.

Abbas [13] developed nearly 1,100 Quranic concrete and abstract concepts linked to all verses of Al-Quran. She used existing Quranic topics from the Islamic scholarly book, *Mushaf Al Tajweed* [14]. These concepts in the index have an aggregate relationship; the hierarchy of concepts is non-reflexive, non-symmetric and transitive.

Dukes [11] extracted 300 concepts and 350 relationships from Al-Quran. The relationship types connecting concepts using predicate logic are 'part-of' and 'is-a'. The ontology is based on a famous Al-Quran discerption book "*Tafsir Ibn Kathir*" [15].

Azman [16] created a Quranic ontology based on themes mentioned in "*Syammil Al-Quran Miracle the Reference*" [17]. This ontology was evaluated by several experts in Al-Quran knowledge, and was built using the Protégé tool in English-Malay languages.

Muhammad [18] developed an annotated dataset for Al-Quran covering of pronoun antecedents. This dataset consists of 1,050 concepts and more than 2,700 relationships. In addition, the relationship types connecting concepts are 'has-antecedent', 'has-concept' and, 'has-a-segment'. Additionally, he produced a dataset called QurSim, containing 7,600 pairs of related verses that contain similarity in the main topic. The scope of this dataset is the similarity of Quran verses [19].

Aldhubayi [20] unified three different Quranic datasets for the Arabic text of Al-Quran. These datasets were created by former researchers at the University of Leeds. These datasets are the Al-Quran Arabic Corpus [21], the Quran annotated with Pronominal Anaphor [QurAna] [22] and, the Qurany project [23]. These datasets are merged in one XML file, and then the file is uploaded to the Sketch Engine[3] as a unified Arabic Quranic corpus.

Abdelnasser [24] developed 1,217 leaf concepts for the Quranic ontology. This ontology was integrated from the Quranic Arabic Corpus Ontology (QCO) [21] and the Qurany topics [23]. In this ontology, each verse of Al-Quran is connected to at least one leaf concept. However, when the QCO and Quranic concepts were merged and manipulated, 621 verses did not link to any concepts.

Hakkoum and Raghay [25] developed a new Quran ontology by combining the Al-Quran Arabic Corpus, the Quran annotated with Pronominal Anaphor (QurAna), part of Quranic Arabic Corpus Ontology (QCO), and the Qurany project. These datasets were merged into one OWL file.

III. Methodology of developing Quran Ontologies

Developing Quranic ontologies from deferent Islamic and Quranic data resources required several sequential stages. These were: collection of Al-Quran datasets and ontologies; reviewing both datasets and ontologies; formatting and normalisation of datasets; aligning; merging; and, finally, storing in the ontology database.

A. Evaluation criteria of existing Quranic ontologies

This stage is aimed at designing criteria to review and evaluate most of the ontologies that are constructed for the Holy Quran. An ontology can be evaluated against several criteria. For example, ontologies can be assessed in terms of the specific use cases, scenarios, requirements, applications, triples size, and scope. Additionally, this evaluation includes the consistency and completeness of the ontology and the representation modelling language. Moreover, assessment of ontology covers the feasibly of the ontology alignment with other ontologies and improvement [2]. Alrehaili and Atwell [4] compared existing Quranic ontologies against nine measures including number of concepts, availability, relationship type, verification methods, coverage area, maturity level, and underlying format. The survey concluded that these ontologies provide unclear consensus on semantic annotation format and validation methods. In this paper, the evaluation of existing Quranic ontologies uses fourteen criteria as follows:

1. Scope:

a) Morphological

[1] https://www.w3.org/RDF/
[2] http://wiki.dbpedia.org/
[3] https://www.sketchengine.co.uk/

b) Translation

c) Quran topics

d) Antecedent pronouns

e) Animals

f) Time

g) Subjects

h) History

i) Prayer (Salaht)

j) Women

k) Similarity between verses

2. Types of relationships between concepts:

a) Taxonomy or Hierarchy: such as 'is_a', 'part_of' or, 'sub_class'.

b) Un-taxonomy: uses a verb to describe the relationship between two concepts.

3. Relationship numbers (triples)

4. Number of concepts

5. Semantic Ontology formats:

a) Not applicable (Text)

b) RDF

c) OWL

6. Ontology representation language:

a) Arabic

b) English

c) Malay

d) Dutch

e) More than four

7. Availability of a source file for the ontology:

a) Available to use

b) Not available to use

8. Validation techniques: methods of validating the ontology:

a) By domain experts: an Islamic scholar

b) Depending on existing Islamic resources such as Tafseer.

c) None

9. Coverage Domain:

a) Covers all Quranic verses

b) Covers almost all Quranic verses

c) Covers half of the Quranic verses

d) Covers only some verses

10. Is dependent on another ontology (dependency): this means that a new ontology is built based on a previous ontology:

a) No

b) Yes.

11. Is used by another ontology (Usability): this means that a new ontology is built based on a previous ontology.

a) Yes

b) No

12. Published on Linked Open Data:

a) Yes

b) No

13. Linked to another linked data: upper ontology, such as friend-of-friend ontology:

a) Yes

b) No

14. Is ontology used in application:

a) Yes

b) No

B. Review of Quran Ontologies

According to the review of Al-Quran ontologies in Appendix A, some deficiencies were found in most of these ontologies. For example, some ontologies were not evaluated by an Islamic scholar [8], [12], [24], [26]–[29], or not tested by an application. Moreover, most of these ontologies do not tag all Quranic verses with semantic tags. Furthermore, these ontologies were built in different structures and file formats, such as CVS, XML, RDF, OWL or text. Additionally, these ontologies are available in one or multi-natural languages, such as Arabic, English, or Malay. Moreover, these different datasets shared some similarity in concepts (overlapped). Additionally, the majority of Al-Quran ontologies are part of, or dependent upon, the Quranic Topics dataset (QT), the Arabic Quran Corpus (AQC), the Ontology of Quranic Concepts (OQC), or the QurAna dataset.

C. Resources to Build Al-Quran Ontologies

The current datasets AQC, OQC, QT, part of Semantic Quran dataset [26], and QurAna were selected to be used as resources for developing a new Quan ontology because: they are used in most existing Quran ontologies. Additionally, these datasets cover all Al-Quran verses, as illustrated in Table 1. Moreover, they are represented in both the Arabic and English languages, and cover many aspects of knowledge.

TABLE 1 THE SEMANTIC TAGS COVERAGE OF AL-QURAN BY SELECTED ONTOLOGIES.

Ontology	Total Tagged verses	Coverage %
OQC	1343	21.54
QurAna	5537	88.79
QT	5561	89.18
All	6202	99.45

Additionally, there are more valuable datasets we used as resources for Al-Quan ontologies. For example, the Arabic quranic word meanings dataset extracted from '*Mushaf Al Tajweed*' [14]. Moreover, the dataset of Al-Quan names, surah names, and some verse names, which was extracted from "*Names of Al Quran and its surah names and verses names*" [30]. Table 2 shows examples of verses names datasets.

TABLE 2 VERSES NAMES DATASETS

Verse name	English transliteration	Chapter name	Verse No
ادم	adm	الأعراف	189
الأخوة	al'akhwa	الحجرات	9
الإذن	al'izhn	النور	27
الإذن	al'izhn	النور	58
الإذن بالقتال	al'izhn baalqtaal	الحجرات	39
الإذن في خروج النساء	al'izhn fee kharouj annasaa'	الأحزاب	53
الأذى	al'azhaa	البقرة	222
الحبس	alhabs	النساء	15
الإرتداد	al'irtdaad	المائدة	54
الردة	arrada	المائدة	54
الإستئذان	al'isti'zhaan	النور	27

D. Similarity between Selected Quran Datasets

The similarities between concepts (entities) in the OQC, QT and QurAna ontologies were measured using two methods: exact match of 2 strings, and Simple Fuzzy String Similarity[31]. The exact match between concepts occurs if the concepts share the same Arabic, or English name. The results of these experiments are shown in Table 3.

TABLE 3 EXACT MATCH OF ARABIC OR ENGLISH CONCEPTS NAMES

Datasets	OQC	QurAna	-
OQC	-	67	-
QT	39	68	-
All	-	-	21

Table 4 shows examples of exact match between the OQC and QurAna.

TABLE 4 EXAMPLE OF SIMILAR CONCEPTS IN BOTH OQC AND QURANA

OQC Arabic concept	QurAna Arabic concept	OQC English concept	QurAna English concept
إسماعيل	إسماعيل	Ishmael	Ishmael
إلياس	إلياس	Elijah	Elias

The Simple Fuzzy String Similarity algorithm[31] was used to find similarities between concepts in which a pair of concepts from two ontology datasets are compared. The results of all similar concepts are illustrated in Table 5.

TABLE 5 SIMILAR CONCEPTS IN THE QT, QURANA AND OQC DATASETS

Dataset	No. of Concepts	No. Comparisons	No. Matching
QT & QurAna	1150, 1050	1194669	339
QT & OQC	1150, 300	285606	100
OQC & QurAna	300, 1050	253135	93

Subsequently, these results were reviewed manually to verify the matches between similar concepts. Then, each matched pair of concepts has the same Arabic or English names.

TABLE 6 EXAMPLES OF THE VERIFICATION OF CONCEPTS WITH MATCHED RESULTS

QT	OQC	Match%	Manual match
العمرة	عمرة	66.66	yes
الزبور	زبور	66.66	yes
السحاب	سحاب	66.66	yes
الحديد	حديد	66.66	yes
الجماعة	الجمعة	66.66	no
العرب	العرش	66.66	no
هارون	هاروت	66.66	no
هارون	قارون	66.66	no

E. Storing Quranic ontologies

After matching and merging Quranic ontologies using the Protégé[4] ; the new ontology is stored in a graph database. The Graph Database is a subject-predicate-object database server (triple-store). This is used to provide the protocol engine for other RDF query and storage systems. Apache Jena Fuseki[5], GraphDB[6] and Neo4j[7] are examples of the graph database systems. Both Fuseki and Neo4j were used to store the merged ontologies, because Neo4j has more features than Fuseki such as graphical presentation of concepts, powerful query language

[4] https://protege.stanford.edu/
[5] https://jena.apache.org/documentation/serving_data/
[6] https://ontotext.com/products/graphdb/
[7] https://neo4j.com/

called Cypher, and API with many programming languages such as Python and PhP.

IV. Conclusion and Future work

This paper reviews previous Quranic ontologies and compares them against fourteen criteria. According to this study, some deficiencies have been found in the majority of these ontologies, such as lacking evaluation by an Islamic scholar and tested by an application. Additionally, the most common datasets covering Al-Quran are: the Quranic topics (QT), Arabic Quran Corpus (AQC), and QuranA.

References

[1] T. Gruber, 'Ontology (Computer Science) - definition in Encyclopedia of Database Systems'. [Online]. Available: http://tomgruber.org/writing/ontology-definition-2007.htm. [Accessed: 20-Aug-2015].

[2] L. Obrst, B. Ashpole, W. Ceusters, I. Mani, and B. Smith, 'The Evaluation of Ontologies: Toward Improved Semantic Interoperability', in *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, New York: Springer Verlag, 2007, pp. 1–19.

[3] O. Ahmad, I. Hyder, R. Iqbal, M. A. A. Murad, A. Mustapha, N. M. Sharef, and M. Mansoor, 'A Survey of Searching and Information Extraction on a Classical Text Using Ontology-based semantics modeling: A Case of Quran', *Life Sci. J.*, vol. 10, no. 4, 2013.

[4] S. M. Alrehaili and E. Atwell, 'Computational ontologies for semantic tagging of the Quran: A survey of past approaches', *Lr. 2014 Proc.*, 2014.

[5] A. Zaeri and M. Nematbakhsh, 'A Semantic Search Algorithm for Ontology Matching', *semantic-web-journal.net*.

[6] P. Shvaiko and P. Shvaiko, 'Ten challenges for ontology matching ome Euzenat To cite this version : Ten Challenges for Ontology Matching', 2013.

[7] M. A. Sherif and A.-C. Ngonga Ngomo, 'Semantic Quran', *Semant. Web*, vol. 6, no. 4, pp. 339–345, Aug. 2015.

[8] H. U. Khan, S. M. Saqlain, M. Shoaib, and M. Sher, 'Ontology Based Semantic Search in Holy Quran', *Int. J. Futur. Comput. Commun.*, vol. 2, no. 6, pp. 570–575, 2013.

[9] Z. El-Naggar, *Al-Hayawanat FI Al-Quran AL-KAREEM*, First. Beirut: DAR El-Marefah, 2006.

[10] R. A. K. Aliyu Rufai Yauri Azreen Azman, Masrah Azrifah Azmi Murad, 'Semantic Web Application for Historical Concepts Search inAl-Quran.', *Int. J. Islam. Appl. Comput. Sci. Technol.*, vol. 2, no. 2, pp. 1–7, 2014.

[11] K. Dukes, 'Statistical Parsing by Machine Learning from a Classical Arabic Treebank', 2013.

[12] M. Al-Yahya, H. Al-Khalifa, A. Bahanshal, I. Al-Odah, and N. Al-Helwah, 'An ontological model for representing semantic lexicons: an application on time nouns in the holy Quran', *Arab. J. Sci. Eng.*, vol. 35, no. 2, p. 21, 2010.

[13] N. Abbas, 'Quran'search for a Concept'Tool and Website', *Unpubl. Diss. Leed*, 2009.

[14] Dar al-Maarifah, *Mushaf al Tajweed*, 4th ed. Damascus: Dar al-Maarifah, 1999.

[15] M. S. Abdul-Rahman, *Tafsir Ibn Kathir Juz' 1 (Part 1): Al-Fatihah 1 to Al-Baqarah 141 2nd Edition*. MSA Publication Limited, 2009.

[16] A. Ta'A, M. S. Abdullah, A. B. M. Ali, and M. Ahmad, 'Themes-based classification for Al-Quran knowledge ontology', *Int. Conf. ICT Converg.*, pp. 89–94, 2014.

[17] L. P. M. A.-Q. (Indonesia) and I. D. Agama, *Syaamil al-Qur'an miracle the reference: mudah, sahih, lengkap, dan komprehensif*. Sygma Pub., 2010.

[18] A. Muhammad, 'Annotation of conceptual co-reference and text Mining the Qur'an', 2012.

[19] A.-B. M. Sharaf and E. S. Atwell, 'QurSim: A corpus for evaluation of relatedness in short texts', 2006.

[20] N. Abbas, L. Aldhubayi, H. Al-Khalifa, Z. Alqassem, E. S. Atwell, K. Dukes, M. Sawalha, and M. Sharaf, 'Unifying linguistic annotations and ontologies for the Arabic Quran', in *Proc WACL2 Second Workshop on Arabic Corpus Linguistics*, 2012, p. 13.

[21] K. Dukes, 'Statistical Parsing by Machine Learning from a Classical Arabic Treebank', School of Computing, University of Leeds, 2013.

[22] A.-B. M. Sharaf and E. Atwell, 'QurAna: Corpus of the Quran annotated with Pronominal Anaphora', in *LREC*, 2012, pp. 130–137.

[23] N. H. Abbas, 'Quran "search for a concept" tool and website', 2009.

[24] H. Abdelnasser, R. Mohamed, M. Ragab, A. Mohamed, B. Farouk, N. El-Makky, and M. Torki, 'Al-Bayan: An Arabic Question Answering System for the Holy Quran', *ANLP 2014*, p. 57, 2014.

[25] A. Hakkoum and S. Raghay, 'Semantic Q&A System on the Qur'an', *Arab. J. Sci. Eng.*, 2016.

[26] M. A. Sherif and A.-C. Ngonga Ngomo, 'Semantic QuranA multilingual resource for natural-language processing', *Semant. Web*, 2009.

[27] H. S. Al-Khalifa, M. Al-Yahya, A. Bahanshal, I. Al-Odah, and N. Al-Helwah, 'An approach to compare two ontological models for representing quranic words', in *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services*, 2010, pp. 674–678.

[28] L. Aldhubayi, 'Unified Quranic Annotations and Ontologies', University of Leeds, UK, 2012.

[29] A. R. Yauri, R. A. Kadir, A. Azman, and M. A. A. Murad, 'Quranic Verse Extraction base on Concepts using OWL-DL Ontology', *Res. J. Appl. Sci. Eng. Technol.*, vol. 6, no. 23, pp. 4492–4498, 2013.

[30] B. Adam, *Assma Alquran Alkareem wa Assma Ayatih wa Suwarih*, 1st ed. Dubai: Markaz Jum'ah al-Mājid lil-Thaqāfah wa-al-Turāth, 2009.

[31] George Stragand, 'Simple Fuzzy String Similarity in Java - CodeProject', *The Code Project*, 2011. [Online]. Available: https://www.codeproject.com/Articles/147230/Simple-Fuzzy-String-Similarity-in-Java. [Accessed: 02-Oct-2017].

V. APPENDIX

A. Appendix A

Table 1: Evaluation criteria of existing Quranic Ontologies

Name of ontology	1. scope	2. Relationship types	3. NO. Triples	4. Number of concepts	5. File Format	6. Ontology Lang	7. Availability	8. Validation tech	9. Coverage Domain	10. Dependency	11. Usability	12. Published	13. use upper ontology	14. used in application
(Sherif & Ngonga Ngomo 2009)	a, b	b	>15m	6	b	e	a	c	a	b	b	a	a	b
(Abbas 2009)	c	a	11824	1150	d	a,b	a	b	a	a	a	b	b	a
(Al-Khalifa et al. 2009)	j	b	n/a	18	c	a	a	c	d	a	b	b	a	b
(Al-Yahya et al. 2010)	f	b	n/a	18	c	a	a	c	d	a	b	b	a	b
(Saad et al. 2010)	i	a	374	6	b	b	b	a	d	a	b	b	b	b
(Muhammad 2012)	d	a	24679	1050	d	a	a	b	a	a	a	b	b	a
(Aldhubayi, 2012)	a,d,c	a	128k	1050	d	a b	a	c	a	b	b	b	b	a
Azman, (2013)	c	a	n/a	n/a	c	c	b	A b	a	a	b	b	b	b
(Dukes 2013)	g ,c	a	350	300	a	a	b	a	a	b	a	b	b	a
(Khan et al. 2013)	e	a	n/a	n/a	c	b	b	c	d	a	b	b	b	a
(Yauri et al. 2013b)	c,g	a	650	300	c	b,c	b	c	a	b	b	b	b	a
(Yahya et al. 2013)	g	a	5695	300	a	b,c	b	c	a	b	b	b	b	a
(Abdelnasser et al. 2014)	c, g	a	n/a	1350	b	a	b	c	a	b	b	b	b	a