EasyChair Preprint

№ 7860

Automatic Mapping of Quranic Ontologies Using RML and Cellfie Plugin

Ibtisam Khalaf Alshammari, Eric Atwell and Mohammad Ammar Alsalka

April 28, 2022

Automatic Mapping of Quranic Ontologies Using RML and Cellfie Plugin

Ibtisam Khalaf Alshammari[1,2][0000−0002−7619−373X], Eric Atwell[1][0000−0001−9395−3764], and Mohammad Ammar Alsalka[1][0000−0003−3335−1918]

[1] University of Leeds, Leeds, United Kingdom
ml18ikfa, e.s.atwell, m.a.alsalka@leeds.ac.uk
[2] University of Hafr Al-Batin, Hafr Al-Batin 39524, Kingdom of Saudi Arabia
ikalshammari@uhb.edu.sa

Abstract. The text of the Qur'an has been analysed, segmented and annotated by linguists and religious scholars, using a range of representations and formats, Quranic resources in different scopes and formats can be difficult to link due to their complexity. Qur'an segmentation and annotation can be represented in a heterogeneous structure (e.g., CSV, JSON, and XML). However, there is the lack of a standardised mapping formalisation for the data. For this reason, this study's motivation is to link morphological segmentation tags and syntactic analyses, in Arabic and Buckwalter forms, to the Hakkoum ontology to enable further clarification of the Qur'an. For achieving this aim, the paper combines two mapping methods: the RDF (resources description framework) mapping language, which is an R2RML extension (the W3C level necessary when mapping relational databases into RDF), and Cellfie plugin, which is a part of the Protégé system. The proposed approach provides the possibility to automatically map and merge the heterogeneous data sources into an RDF data model. Also, the integrated ontology is evaluated by a SPARQL query using an Apache Jena Fuseki server. This experiment was conducted in all the Qur'an chapters and verses, containing all the words and segments of the entire Qur'an corpus.

Keywords: Classical Islamic Text · Heterogeneous Data · Ontology Mapping · Ontology Integration · RML · Cellfie plugin.

1 Introduction

Islamic knowledge has naturally always been an interest of Muslims and has been a growing interest of non-Muslims, especially knowledge of the Holy Qur'an. This is because the Qur'an is the primary sacred text in Islam and the Muslim belief that it is an essential source of information, wisdom, and law. Indeed, due to its unique style and metaphorical nature, the Holy Qur'an requires special consideration when it comes to search and information retrieval concerns. The Holy Qur'an text is written in Classical Arabic text and it is divided into 30 divisions

(أجزاء), 114 chapters (سُور), and subdivided into 6236 verses (آيات).

Many studies have been conducted to accomplish keyword searches with the Holy Qur'an, based on developing Quranic resources and ontologies. The fundamental issues with these works are that the ontologies are incomplete, cover different scopes, and represented in various formats, such as CSV, JSON, and XML files.

The benefit of building a Qur'an ontology-based knowledge base lies in the power of ontologies to enable exploration of semantic relations among concepts. Subsequently, Our hypothesis was that integrating the available resources would enrich an ontology that covers as many of the Quranic concepts as possible. Thus, in this paper, we describe an experiment that combines two ontology mapping methods: the resource description framework (RDF) mapping language and Cellfie plugin to extract, map and integrate the selected resources.

The following section provides insight into the Hakkoum ontology and QacSegment dataset. In section 3, the related work is discussed. Section 4 illustrates how the framework can be applied for ontology mapping and integration. Section 5 presents the results obtained from the experiment and visualises the first chapter of the Qur'an, as an example. In the final section, we conclude by summarising our work and describing the outlook for further work.

2 Existing Qur'an Resources

A number of studies aimed to enrich Quranic resources and ontologies. Some ontologies covered most of the Quranic topics as [8], while others focused on specific concepts such as [13] covered the prayer topic (الصلاة), and [15] covered Umrah pilgrims concept. In this section, Hakkoum ontology and QacSegment corpus are presented in order to be mapped and merged using the proposed framework in Fig. 1

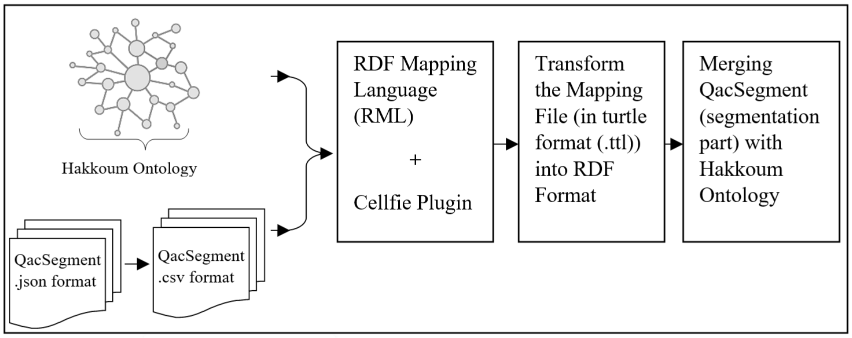

Hakkoum Ontology
QacSegment .json format
QacSegment .csv format
RDF Mapping Language (RML) + Cellfie Plugin
Transform the Mapping File (in turtle format (.ttl)) into RDF Format
Merging QacSegment (segmentation part) with Hakkoum Ontology

Fig. 1. The Proposed Framework

2.1 Hakkoum Ontology

Hakkoum ontology, Qur'an Ontology[1], was constructed by Aimad Hakkoum [8]. Hakkoum ontology has been chosen for extension because it includes key annotation datasets of the Holy Qur'an. This ontology was created using Protégé editor with an aim to represent the Quranic concepts and their relationships. Hakkoum resource has more than one million triples, its size is 123 939 KB.

The Hakkoum ontology links the Quranic text from Tanzil project. Tanzil project provides Qur'an metadata, Qur'an plain Arabic, Uthmani, and English translation texts. It encompasses Qur'an descriptions from the books Tafsir Al-Jalalayn and Al-Muyasser. Furthermore, Hakkoum resource contains the topics discussed in the Holy Qur'an from the index of a widely-cited Qur'an commentary, Tafsir Ibn Kathir, captured and encoded in the Qurany project [1]. The Hakkoum ontology also has the most significant Quranic annotations in the QurAna dataset [14]. However, the morphological segmentation tags and syntactic annotation of each word, in the Arabic and Buckwalter[2] forms, are not included in the Hakkoum ontology.

QacSegment Dataset QacSegment.json[3] is a JSON encoding of the Quanic Arabic Corpus annotated data-set [4, 5] developed by Sharaf and Atwell [14]. This corpus covers the prefix features, roots, lemma, and morphological analysis for each word of the Holy Qur'an. This resource is including the morphological segmentation tags of each Quranic word, which is our interest. Moreover, the syntactic annotation is included that focuses on the dependency grammar to showcase the functional relations between Quranic words. For instance, ["GEN"] is attributed to the last part of a preposition phrase.

3 Related Work

3.1 Ontology

According to Gruber, "*An ontology is an explicit specification of a conceptualization*" [7, p. 199]. Ontology is defined as structuring and representing knowledge explicitly in a machine-readable format that may be integrated into computer-based applications and systems. As a result, the number of researchers studying ontologies for developing Qur'an knowledge bases has increased significantly. The importance of the morphological annotation is demonstrated in [2]. Morphology analysis and a dependency treebank are two popular techniques that can contribute to various natural language processing (NLP) applications such as a knowledge base.

[1] Qur'an ontology Data can be downloaded via: http://Quranontology.com/

[2] This is a computer-readable orthographic transliteration technique that uses ASCII characters to represent Arabic text for non-Arabic academics.

[3] It is can be accessed via: http://textminingthequran.com/

Ontology Mapping. It is called ontology integration or ontology alignment and is the consideration of the semantic correspondences between similar concepts from various ontologies. Ontology mapping plays a crucial role in data interoperability in the semantic web [10]. Ontology mapping aims to unify multiple ontologies within a particular domain [16].

RDF Mapping Language. The Consortium for the World Wide Web created the RDF mapping language (RML) to exhibit particular rules for mapping from various data structures and serialisations, such as tables in the form of comma-separated values (CSV), JavaScript object notation (JSON), and extensible markup language (XML) files to the RDF data model [12]. RML is a suggestion that extends the R2RML recommendation to include diverse data sources [11].

RML is used to map heterogeneous and hierarchical data sources into RDF. [3] provided examples to generate data from two different formats, such as XML and JSON files. To conclude, they evaluated their experiment against various criteria and then stated that RML provided an optimal solution as it was semantically richer and better interlinked than alternatives.

Cellfie is a Protégé plugin that automatically imports and maps spreadsheets to ontologies in OWL (the abbreviation for the web ontology language). It can be used by setting transformation rules to convert spreadsheets into OWL formats. For example, in [6] the Cellfie plugins were used to import data related to COVID-19 from the Indian province of Karnataka. They explained that each row was transformed into a patient class individual, with the values such as patient case and age.

4 Methodology

4.1 Data Preparation

This section describes the vital stage of our work in the initial preparation of the data. The Hakkoum ontology was reviewed and evaluated manually, and we noticed the following limitations:

First, the Hakkoum ontology was built based on Qur'an metadata resource, and the Qur'an metadata has three types of chapter names: Arabic, English, and transliterated, i.e. Arabic words written with English alphabet. However, Hakkoum does not distinguish between transliterated and English names because the ontology has 47 chapter names in English, and 67 chapter names in transliterated words. For example, the chapter "The Opening" (سورة الفاتحة) is represented in the Hakkoum ontology as the English form "The Opening," While the chapter "The Women" (سورة النساء) is expressed in the transliterated word "An-Nisaa." Therefore, we modified the transliterated names manually using Protégé and changed them to English names.

Second, the Hakkoum ontology covers many areas related to Qur'an verses and words, such as displaying the text in simple and Uthmany style, showing the

Ayah numbers for each chapter, dividing each verse into words, indicating the pronoun reference of each term, and showing the word's lemma and root. However, it does not have morphological segmentation tags and syntactic features. Hence, we contribute by linking the QacSegment resource with the morphological segmentation layer and syntactic analyses to clarify the Qur'an text further.

Before conducting our experiment, the QacSegment dataset was downloaded and prepared for the second phase by converting the JSON file to CSV file by importing the CSV and JSON Python libraries. Then, building a conceptual model for the CSV file and extract the keys represented in our ontology such as classes and data properties. We manually created only the class names and data properties to be able to automatically extract their data, the individuals.

RDF Mapping Language. The RDF mapping language (RML) is used to express bespoke mapping rules from the CSV file to the RDF data model. The first step was to map the Hakkoum classes and CSV columns. For instance, in our CSV file, the "SurahId" column was mapped to the "Chapter" class. The same process was applied to the other Hakkoum classes and the CSV columns. The segmentation class was then created and linked with the "Word" concept in the Hakkoum ontology. Although the RML mapped the CSV file instances, it did not connect the Arabic words because the Arabic language is not supported in RML. Finally, the mapped file was saved in the terse RDF triple language (or "turtle") format (.ttl).

Cellfie plugin. The purpose of using the Cellfie tool is to extract the Arabic text by customising specific rules because the RML does not support the Arabic language. The initial step was to convert the CSV file into an Excel file as the Cellfie plugin can handle the Excel format.

Then, SDM-RDFizer is used to transform the turtle file into RDF. SDM-RDFizer is a mapping rule interpreter to transform unstructured data into RDF. The obtained result, N-triple, will be uploaded to the Protégé editor and then converted to OWL format to be uploaded again. Finally, we can import the Hakkoum ontology to finalise our experiment.

5 Results

The integrated ontology is mapped and linked the chosen resources properly. Then, the first chapter "Al-Fateeha" of the integrated ontology is visualised (see Fig. 4). We display the first chapter because the image for all the Qur'an chapters is large. Table 1 shows a comparison of Hakkoum and the integrated ontologies. We can notice that the size of the integrated ontology is increased by 131 88 KB, from 123 939 to 137 127 KB.

5.1 Apache Jena Fuseki

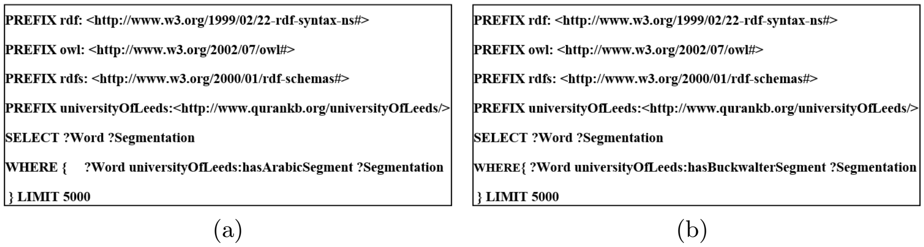
This can be defined as a SPARQL server that can run as an operating system service. This stage of our experiment evaluates the resulting ontology to check the classes and triples. Fig. 2 shows the Arabic and Buckwalter segmentation's SPARQL queries, and Fig. 3 illustrates their results. The purpose of using Apache Jena Fuseki is that Protégé editor cannot create a SPARQL query for a large RDF triples. Protege is widely used for research on small example ontologies. However, it does not scale up to very large data-sets.

Analysis. The Protégé editor did not work very well and crashed during some attempts; we tried to increase the maximum memory allocation pool for a Java virtual machine (JVM) through running a batch file to improve its performance.

Furthermore, the process was time consuming when we linked the Arabic concepts with the Cellfie plugin. Also, during development, another issue had arisen: a SPARQL query cannot work through Protégé; we, therefore, used Apache Jena Fuseki because it had good performance and could process the SPARQL query.

Table 1. Comparing the Integrated Ontology to Hakkoum Ontology

Matrices	Hakkoum Ontology	Integrated Ontology
Axiom	1 282 191	2 471 467
Class count	46	47
Individual count	110 939	239 158
Data property count	23	29



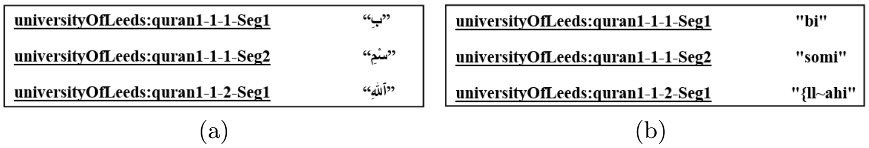
```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schemas#>
PREFIX universityOfLeeds:<http://www.qurankb.org/universityOfLeeds/>
SELECT ?Word ?Segmentation
WHERE {     ?Word universityOfLeeds:hasArabicSegment ?Segmentation
} LIMIT 5000
```

(a)

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schemas#>
PREFIX universityOfLeeds:<http://www.qurankb.org/universityOfLeeds/>
SELECT ?Word ?Segmentation
WHERE{ ?Word universityOfLeeds:hasBuckwalterSegment ?Segmentation
} LIMIT 5000
```

(b)

Fig. 2. SPARQL Query to Generate (a) Arabic and (b) Buckwalter Segment



```
universityOfLeeds:quran1-1-1-Seg1        “بِ”
universityOfLeeds:quran1-1-1-Seg2        “سْمِ”
universityOfLeeds:quran1-1-2-Seg1        “ٱللَّهِ”
```

(a)

```
universityOfLeeds:quran1-1-1-Seg1        "bi"
universityOfLeeds:quran1-1-1-Seg2        "somi"
universityOfLeeds:quran1-1-2-Seg1        "{ll~ahi"
```

(b)

Fig. 3. Segmentation Results in (a) Arabic and (b) Buckwalter of SPARQL Queries

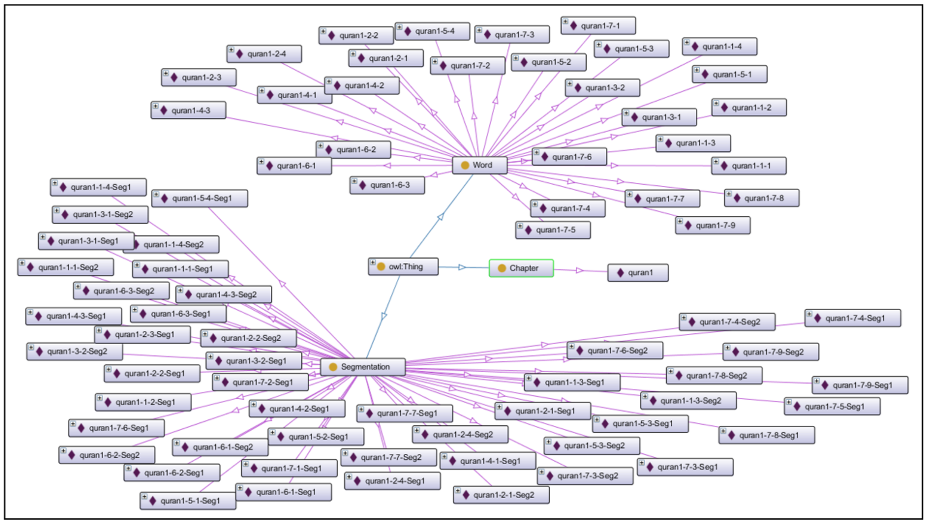


Fig. 4. Integrated ontology Visualisation for the First Chapter

6 Conclusion and Future Work

The contribution of this paper is the development of a framework of ontology mapping techniques applied to the linguistic and religious annotations of the Quran corpus: RML and Cellfie plugin with the tool interpreter, SDM-RDFizer. Although we faced difficulties with the RML method because it does not support the Arabic language, it merged the proposed ontologies. In addition, the Apache Jena Fuseki SPARQL server was used to handle the very large knowldge base. The findings of this study were sufficient because the morphological annotations and syntactic analyses were linked to all the Qur'an chapters, verses, and words in the Hakkoum ontology.

For future work, we intend to continue working on the same process with the remaining Quranic resource content. We can contribute to mapping the transliterated names from the Qur'an metadata resource. Furthermore, we plan to map the Quranic ontology with Hadith in order to build a comprehensive knowledge base of the most important Islamic concepts.

Acknowledgements Ibtisam Alshammari would like to express her deepest gratitude to Saudi Arabia Cultural Bureau (SACB) in London, United Kingdom, as well as University of Leeds, and University of Hafr Al-Batin.

References

1. Abbas, N. H.: Qur'an 'Search for a Concept' tool and website. Unpublished Dissertation. University of Leeds (2009)

2. Ayed, R., Chouigui, A., Elayeb, B.:A new morphological annotation tool for arabic texts. IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), pp. 1–6 (2018)
3. Dimou, A., Sande, M.V., Slepicka, J., Szekely, P., Mannens, E., Knoblock, C., Walle, R.V.d.: Mapping hierarchical sources into RDF using the RML mapping language. IEEE International Conference on Semantic Computing, pp. 151–158. IEEE, New York, NY (2014)
4. Dukes, K., Atwell, E.: LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis. In: Calzolari C et al. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)
5. Dukes, K., Atwell, E., Habash, N.: Supervised collaboration for syntactic annotation of Quranic Arabic. Language Resources and Evaluation Journal. **47**(1), 33–62 (2013)
6. Dutta, B., DeBellis, M.: CODO: an ontology for collection and analysis of COVID-19 data. In: Proceeding of 12th International Conference on Knowledge Engineering and Ontology Development (KEOD), vol.2, pp. 76–85 (2020). DOI: 10.5220/0010112500760085.
7. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition. **5**(2), 199–220 (1993)
8. Hakkoum, A., Raghay, S.: Semantic Q&A system on the Qur'an. Arabian Journal for Science and Engineering (2016). https://doi.org/10.1007/s13369-016-2251-y
9. Ma, C., Molnár, B.: Use of ontology learning in information system integration: a literature survey. In: Sitek, P., Pietranik, M., Krótkiewicz, M., Srinilta, C. (eds) Intelligent Information and Database Systems (ACIIDS), 2020. Communications in Computer and Information Science, vol 1178. Springer. Singapore, (2020). https://doi.org/10.1007/978-981-15-3380-8_30
10. Mao, M.: Ontology mapping: An information retrieval and interactive activation network based approach. In: Aberer K. et al. (eds) The Semantic Web. ISWC 2007, ASWC 2007. Lecture Notes in Computer Science, vol 4825. Springer, Berlin, (2007). https://doi.org/10.1007/978-3-540-76298-0_72
11. Meester, B.D., Heyvaert, P., Verborgh, R., Dimou, A.: Mapping languages analysis of comparative characteristics. In: Knowledge Graph Building and Large Scale RDF Analytics. CEUR Workshop Proceedings, vol. 2489 (2019)
12. RDF Mapping Language (RML), https://rml.io/specs/rml/. Last accessed 13 Dec 2021
13. Saad, S., Salim, N., Zainal, H.: Towards context-sensitive domain of Islamic knowledge ontology extraction. International Journal for Infonomics (IJI),**3**(1), 197–206 (2010)
14. Sharaf, A.-B., Atwell, E.: Qurana: Corpus of the Qur'an annotated with pronominal anaphora. In: Calzolari C et al. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp.130–137. European Language Resources Association (ELRA), Istanbul, Turkey (2012)
15. Sharef, N.M., Murad, M. A. A., 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 (2013)
16. Zaeri, A., Nematbakhsh, M.: A Semantic Search Algorithm for Ontology Matching. Semantic Web Journal Net, 254 (2015)