Semantic Enrichment of Hadith Corpus - Knowledge Graph Generation from Islamic Text

Amna Binte Kamran [a], Nigar Azhar Butt [b] and Amna Basharat [a,*]

[a] *Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan*

[b] *Department of Software Engineering, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan*

E-mails: amna.kamran@nu.edu.pk, nigar.azhar@isb.nu.edu.pk, amna.basharat@nu.edu.pk

Abstract. Knowledge graphs from text have garnered substantial interest across various domains due to their potential to facilitate efficient information retrieval and knowledge exploration. However, knowledge graph generation from textual sources presents unique challenges, particularly in the Islamic domain, where primary sources of knowledge are texts in Arabic, which exhibit complex linguistic and cultural nuances. This paper presents a comprehensive methodology for generating a knowledge graph from the hadith corpus. Hadith, a fundamental resource in the Islamic domain, stands as one of the primary sources of Islamic legislation, encompassing the sayings, actions, and silent approvals of the Prophet Muhammad. Leveraging Natural Language Processing techniques, we systematically extract, annotate, and interlink semantic entities and relationships from the hadith corpus, extend the *SemanticHadith* ontology for entity organisation, and compute textual similarities to establish semantic connections. We generate a comprehensive knowledge graph by applying these methods to six hadith collections, facilitating efficient information retrieval and knowledge exploration in the Islamic domain. This is an essential step towards annotating and linking the hadith corpus to allow semantic search to support scholars or students in creating, evolving, and consulting a digital representation of Islamic knowledge. The *SemanticHadith* knowledge graph is freely accessible at http://www.semantichadith.com.

Keywords: Knowledge Graph, Ontology, Knowledge Modelling, Hadith, Quran, Named Entity Recognition, Knowledge Representation and Reasoning

1. Introduction

In the current landscape of abundant data and information, the structured representation of knowledge has become crucial for efficient information retrieval and utilisation across diverse domains. Such structured representations allow systems to organise, categorise, and interconnect information, enabling faster access and deeper insights. Among the various methods employed for this purpose, knowledge graphs (KGs) have garnered significant attention for their ability to revolutionise information management and retrieval. KGs, with their graph-based organisation of entities and relationships, provide a standardised framework for representing and navigating complex knowledge structures [1–3].

*Corresponding author. E-mail: amna.basharat@nu.edu.pk.

Despite the widespread adoption of knowledge graphs and open data initiatives in various sectors such as biomedicine, education, government, and cultural heritage, the religious domain — particularly Islamic knowledge — has not fully utilised the benefits of linked open data. Generating knowledge graphs from textual sources poses distinct challenges, especially in domains characterised by complex linguistic and cultural nuances such as those found in Islamic texts. Our preliminary work has demonstrated the potential of knowledge graphs for representing Islamic knowledge [4].

The Islamic knowledge domain, rooted in the teachings of the Quran and the sunnah of the Prophet Muhammad, presents a vast repository of textual resources that are yet to be fully integrated into the semantic web ecosystem [5, 6]. Despite the wealth of information contained within the hadith corpus, comprising narrations of the sayings, actions, and approvals of the Prophet Muhammad, systematic efforts to model and represent this knowledge have been limited [7–12]. While significant efforts have been dedicated to morphological annotation and ontology modelling of the Quran, the domain of hadith has received comparatively less attention [7–10]. Existing studies in hadith primarily focus on automating the extraction of the chain of narrators or specific domains within the hadith corpus, such as prophetic medicine [11, 13–16]. Our previous work, *SemanticHadith*, fills a crucial gap by providing a detailed modelling of the structure of hadith and narrator chains [12]. However, the lack of comprehensive semantic modelling for hadith texts hampers their integration into the broader knowledge ecosystem, hindering seamless exploration and retrieval of Islamic knowledge resources. There exists a clear need to bridge the gap between the structured representation of Islamic knowledge and the broader semantic web landscape. By leveraging principles of linked data and knowledge graph construction, the vast potential of Islamic textual resources can be unlocked for efficient information retrieval, exploration, and utilisation.

To address these challenges, this paper presents a comprehensive methodology for generating a knowledge graph from the hadith corpus. This methodology serves as a cornerstone for enhancing the accessibility and interoperability of Islamic knowledge resources. Through meticulous data selection, Natural Language Processing (NLP)-based entity extraction, semantic modelling and knowledge graph construction, our approach aims to facilitate seamless exploration and retrieval of Islamic knowledge resources in the digital age. Furthermore, we introduce the SemanticHadith Ontology Version 2.0.1, which builds upon the foundation laid by the SemanticHadith Ontology version 1.0.1. This updated ontology provides a modelling of entities and topics within the hadith text, thereby enabling a more nuanced understanding and exploration of Islamic teachings.

By formalising this vast knowledge repository and its linkages, we aim to enable new avenues for research, knowledge discovery, and synthesis within the Islamic knowledge domain. This emphasis on comprehensive semantic modelling underscores the importance of integrating various textual resources, including Quranic commentaries, which heavily rely on hadith corpora for interpreting Quranic verses. Such integration promises to enrich the exploration of Islamic knowledge resources and foster a deeper understanding of the interconnectedness between different facets of Islamic teachings. Our research aims to overcome the inherent challenges associated with knowledge acquisition in semantic content creation, particularly for applications relying on semantic technologies. By adopting a semantic perspective, we aim to facilitate the search for concepts and relations within hadith texts.

2. Background Context and Motivation

In Islamic tradition, a hadith serves as a crucial source of knowledge, providing narratives of historical events from the life of Prophet Muhammad, interpretations of Quranic verses, explanations of revelations, contextual backgrounds, and elaborations of essential Islamic concepts. Second only to the Quran, the hadith corpus holds significant importance in shaping Islamic jurisprudence and understanding. The structure of a hadith comprises two primary elements: the matan, representing the narration's content, and the sanad, which is the chain of narrators through whom the hadith was transmitted. The sanad, often presented as a chronological list of narrators, plays a pivotal role in assessing the authenticity of a hadith. Scholars rely on the integrity of the chain of narrators when determining whether to accept or reject a particular hadith.

The vastness of the hadith corpus presents challenges in managing its intricate relationships and concepts. In addition to the primary hadith texts, there exists an extensive body of supporting literature, including commentaries

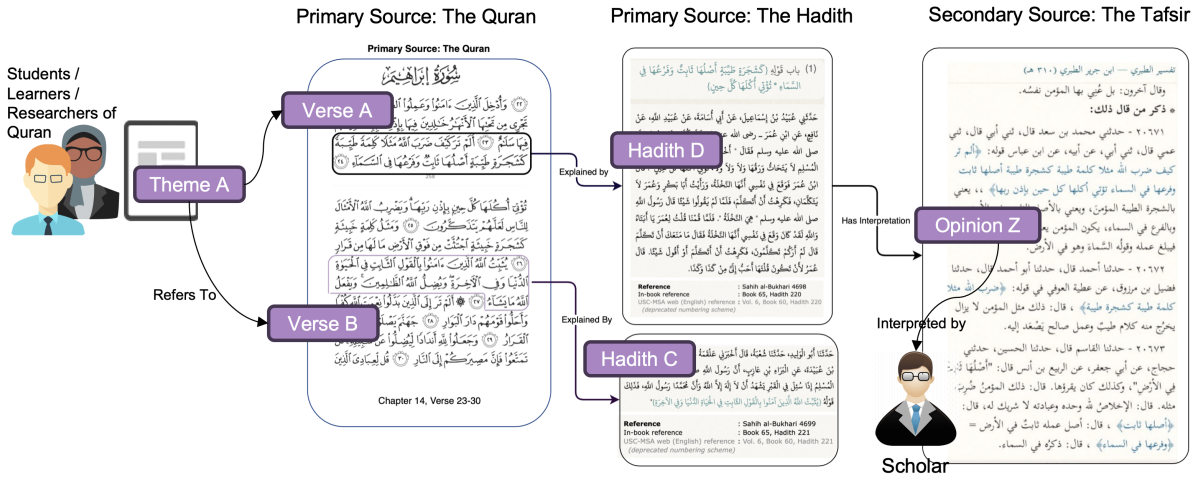


Fig. 1. Motivational Scenario - Connecting Primary and Secondary Sources

and biographical material. Navigating this wealth of information requires a nuanced understanding of the interconnections between different hadith, concepts, and topics. While identifying the narrators of each hadith may be straightforward, delving into questions such as "Which hadith elaborates the meaning of a specific verse from the Quran?" or "Which hadith is similar to another?" requires a deeper comprehension of the relationships across both the hadith and the Quran.

2.1. Importance of Hadith

Understanding the significance of hadith requires a comprehensive approach that incorporates the principles of Quranic understanding and the science of tafseer or exegesis. Quranic verses cannot be fully comprehended in isolation; Hadith provides contextualisation and elucidation. Hadith plays a vital role in elucidating the historical context, reasons for revelation, and elaboration of essential concepts that may not be immediately apparent from the text alone. Throughout history, scholars have authored comprehensive commentaries (tafseer) and explanations to uncover the intended meaning of Quranic verses. Adherence to this principle is essential for producing accurate Tafseer of the Quran [17]. It is noteworthy that reliable and authentic commentaries often draw upon hadith sources.

2.2. Need for Formalised Semantic Modelling

In Islamic studies, navigating between primary Quranic sources, hadith literature and secondary sources, the interpretations in the tafseer, presents a challenge for scholars and students alike. Figure 1 presents the scenario of a student engaged in understanding the Quran, eager to explore connections between verses of Surah Ibrahim. To delve into the profound meanings within these verses, the student instinctively looks to the rich explanations and insights offered by hadith literature, elaborated further in renowned tafseer such as Tafseer Tabari. This scenario highlights the need for effective knowledge modelling and graph generation tailored to Islamic knowledge. Islamic knowledge, with its vast potential for generating links across distributed content available on the web, demands a generic architecture to formalise the entire process. By structuring Quranic verses and related hadith explanations within a knowledge graph framework, scholars and learners can explore the links between primary and secondary sources.

Over the centuries, thousands of Quranic commentaries have been produced in various languages, all of which rely on hadith corpora to provide interpretations of Quranic verses. Formalising this vast knowledge repository and its linkages promises to enable new avenues for research, knowledge discovery, and synthesis. In Figure 2, we provide an example from both the Quranic and hadith texts, highlighting the same entities and topics. This visualisation underscores the interconnected nature of these primary Islamic sources, reaffirming the importance of a unified approach in representing and exploring Islamic knowledge.

2.3. Challenges in Semantic Modelling of Islamic Knowledge Sources

Efforts to publish Islamic knowledge as linked data on the Linked Open Data (LOD) cloud have focused on the Quran, with only a few recent attempts to model hadith. These endeavours have predominantly centred on annotating various constituents of hadith. For instance, Fairouz et al. [18] focused on modelling commentaries for Hadith, while Jaafar and Che Pa [19] concentrated on modelling concepts within the Arabic text of hadith. Khatib et al. [20] proposed the development of a WordNet linguistic resource for hadith. In contrast, Azmi et al. [21] provided a comprehensive review of computational and natural language techniques applied to hadith literature in their survey. Harrag et al. [22] utilised association rule mining to create an ontology for Islamic (Fiqh) jurisprudence from hadith texts, similar to a framework proposed by Al-Arfaj and Al-Salman [23] to create an ontology from Arabic texts. Some studies have focused on indexing and classification of hadith [24, 25], while others have explored the social network aspect of hadith narrators [26]. Tools like E-Narrator [13] and HadithRDF have been developed to parse and analyse hadith texts automatically.

However, realising this vision to encompass many Islamic resources poses a monumental task. Furthermore, efforts in computational techniques applied to Hadith, as comprehensively reviewed by Azmi et al. [21] and Bounhas [27], underscore the dire need for multilingual hadith resources to facilitate NLP, information retrieval, and knowledge extraction tasks specific to hadith literature. Some studies have explored the use of Named Entity Recognition (NER) and Natural Language Processing (NLP) techniques to extract entities and relationships from hadith texts [28, 29], and similarity computations have been explored to identify connections between different hadith texts [30]. Similarly, similarity computations have been employed to identify connections between Quranic Verses [31–33].

Despite these efforts, few publicly available data sources or vocabularies have been published as linked data or knowledge graphs for hadith. Most classical sources do not adhere to a standardised numbering scheme for Hadith, and different levels of authenticity exist among hadith collections. Furthermore, hadith texts vary in length, often making tracing the original cited hadith challenging. Nonetheless, realising the vision of encompassing the plethora of Islamic resources remains monumental.

3. Methods

In this section, we outline our approach to generating a comprehensive knowledge graph from the hadith corpus, encompassing several key stages. Figure 3 presents the overview of this framework. We begin by detailing our process of data selection and acquisition, ensuring the inclusion of relevant hadith collections. Subsequently, we describe our NLP-based custom knowledge extraction methodology, which involves techniques for entity recognition and extraction from textual sources. Following this, we discuss our approach to conceptual knowledge modelling and formalisation, wherein we establish a structured framework for organising and representing the extracted entities. We then explore our methodology for similarity computation and interlinking of Hadith, which involves quantifying textual similarities to establish semantic connections within the knowledge graph. Finally, we generate a knowledge graph and link the graph with external data sources. See Sections 3.1 to 3.6 for the detailed knowledge graph generation process.

3.1. Data Selection and Acquisition

In this section, we describe the initial step in generating a knowledge graph, which involves selecting and acquiring relevant data. Building upon our SemanticHadith ontology and knowledge graph [12], we extend its scope to align with the objectives of this study. For this study, we chose the same six prominent and authentic hadith collections sourced from the Islamic Urdu Books Website[1]. Given that the reference data sources are in Arabic, we take precautionary measures to convert the data into Unicode format to prevent potential presentation issues. These six major hadith collections—Sahih Bukhari, Sahih Muslim, Sunan Abi Daud, Sunan ibn Majah, Sunan An-Nisai, and Jami At-Tirmidhi—constitute a comprehensive corpus totalling 34,458 hadith, each accompanied by full Arabic text as well as translations in Urdu and English.

[1] www.IslamicUrduBooks.com

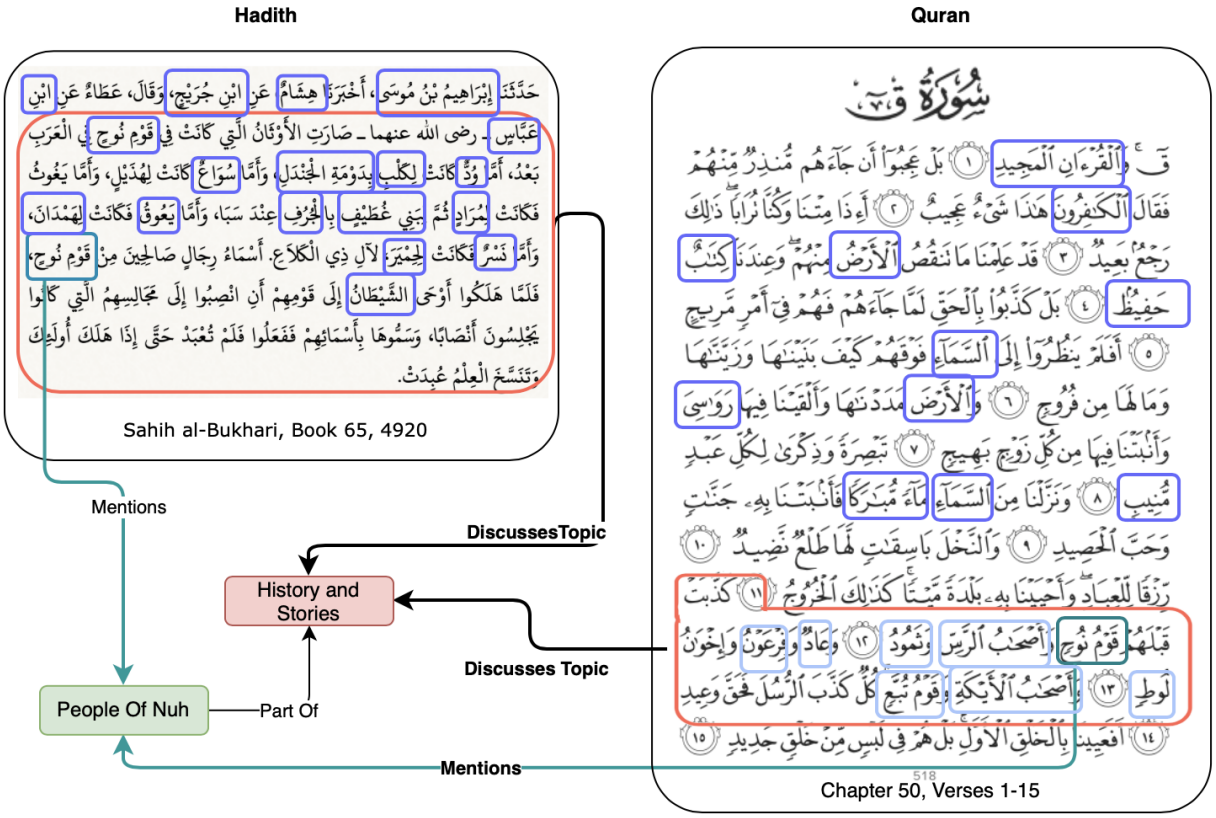


Fig. 2. Motivational Scenario 2 - Connecting Themes, topics, people and places in the Quran and Hadith

3.2. NLP-based Custom Knowledge Extraction

The NLP methodology for entity extraction in developing the SemanticHadith ontology v2.0.1 is pivotal in accurately identifying and extracting relevant entities from the hadith corpus. Our approach ensures precision and comprehensiveness in entity extraction by leveraging a combination of custom-trained Named Entity Recognition (NER) models and expert-validated noun dictionaries.

We begin by preparing the dataset and customising the CANERCORPUS dataset [28] to align with the requirements of our ontology extension. Modifications are introduced with domain experts to incorporate additional concepts relevant to the hadith corpus while removing irrelevant labels. Arabic text processing techniques are then applied to handle text normalisation and tokenisation. Our custom-trained NER model, developed using the state-of-the-art spaCy library, is tailored to the characteristics of Arabic text in the hadith corpus. Transfer learning techniques enhance the model's performance and adaptability to domain-specific language and context. Expert validation and noun dictionaries further enrich the entity extraction process, ensuring accuracy and completeness. Extracted entities are cross-validated against the dictionaries to resolve discrepancies and ambiguities through manual inspection and expert verification. The entity extraction process involves applying the trained NER model to the entire hadith corpus, categorising extracted entities into predefined classes such as angels, prophets, historical figures, and thematic topics. Expert consultation addresses challenges related to person name variations, aiming to reconcile multiple name variants to ontology instances and enhance mapping accuracy.

Overall, our NLP methodology facilitates the systematic extraction of entities from the hadith corpus, laying the foundation for the development of a comprehensive knowledge graph within the SemanticHadith framework. Section 4 provides a detailed description of the NLP methodology for entity extraction.

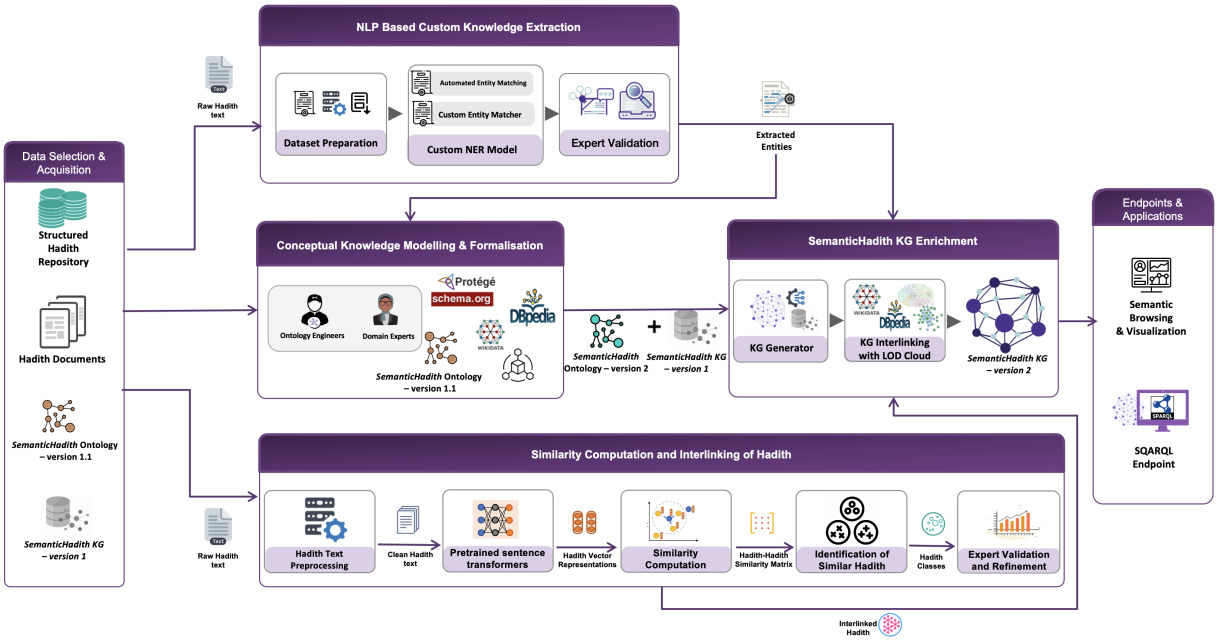


Fig. 3. Overview of the SemanticHadith knowledge graph construction framework. The key stages of the framework include Data Selection and Acquisition, NLP-based Custom Knowledge Extraction, Conceptual Knowledge Modelling and Formalisation, Similarity Computation and Interlinking of Hadith, SemanticHadith Knowledge Graph Enrichment, which encompasses Knowledge Graph Generation and Interlinking with the LOD Cloud, and Endpoints and Applications.

3.3. Conceptual Knowledge Modelling and Formalisation

During this stage, we conceptualise and design a formal ontology structure as described in Section 5. We follow an iterative approach in ontology engineering, where the knowledge model evolves as formalisation progresses. The methodology encompasses seven steps, including determining the scope of the ontology, enumerating important terms, and defining classes, hierarchies, properties, and facets. The scope of the ontology is defined based on competency questions formulated from the findings of the NLP module. Additionally, existing ontologies are reused, and vocabularies such as Schema.org, DBpedia, and Wikidata are leveraged to ensure interoperability and standardisation. The ontology design incorporates key entities and relations relevant to hadith literature, such as Salah, GroupOfPeople, Hadith, and Narrator. Furthermore, strategic design decisions are made to enhance expressiveness and semantic clarity, such as refining class relationships and subclass definitions. The integration and implementation are carried out using Proteg'e version 5.5.0, ensuring compatibility and extensibility. Finally, the ontology is enriched through semantic annotation of hadith texts, facilitating enhanced comprehension and semantic querying capabilities. Figure 4 shows the conceptual model for the *SemanticHadith* ontology.

3.4. Similarity Computation and Interlinking of Hadith

Our study aims to identify similar hadith within the Sahih al-Bukhari and other selected collections, including Sahih Muslim, Ibn Maja, Sunan Abi Dawood, and Nisai. To achieve this, we employ pre-trained Arabic sentence transformers to encode Arabic hadith texts into numerical representations, facilitating the computation of cosine similarity scores between pairs of hadith. The process begins with preprocessing the hadith texts to remove diacritics, punctuation marks, and stop words, ensuring uniformity in representation. These cleaned texts are then encoded using pre-trained sentence transformers, generating embedding vectors for each hadith. Subsequently, a cosine similarity matrix is computed to quantify the similarity between all pairs of hadith. Domain experts review pairs falling within specific similarity bins to validate the identified similar hadith pairs. The experts verify textual similarity and provide feedback on relevance, leading to iterative methodology refinement. Challenges encountered include

discrepancies between isnaad (narrator chains) and textual content and instances where one hadith encompasses a subset of another. Insights gained from the expert validation process highlight the importance of considering contextual relevance beyond textual similarity. Certain hadith pairs, while not textually similar, are deemed relevant due to thematic or historical connections, underscoring the multifaceted nature of similarity in hadith literature. Section 6 describes the identification and linking of similar hadith in detail.

3.5. SemanticHadith Knowledge Graph Enrichment

This section describes the processes involved in the enrichment of the *SemanticHadith* Knowledge Graph. It covers the knowledge graph generation itself and its interlinking with external Linked Open Data (LOD) resources.

3.5.1. Knowledge Graph Generation

The KG-Generation module of our methodology involves aligning the domain and concepts in the data with the ontology classes, automatically translating the hadith records and the entities recognised by our NLP module into Ontology Web Language (OWL) individuals, incorporating data as data properties, and establishing semantic relationships based on object properties and mapping rules. The resulting RDF data is then loaded into the *SemanticHadith* knowledge graph for storage and subsequent reasoning. To transform entities extracted by our NLP pipeline into an RDF-based knowledge graph, we utilise the OntoRefine tool along with our SemanticHadith ontology [34]. Both the *SemanticHadith* ontology version 2.0.1 and the knowledge graph are publicly available on a GitHub repository[2]. GitHub's issue-tracking system will serve as a platform for communication regarding the maintenance and future development of the ontology.

3.5.2. Knowledge Graph Interlinking with LOD Cloud

For linking with external knowledge graphs such as DBpedia [35] and Wikidata [36], we utilised automated interlinking tools like LIMES and OntoRefine [37, 38]. Expert validation was employed to ensure the accuracy of the discovered links and resolve any conflicts or ambiguities. As a result, substantial interlinking was achieved with external knowledge graphs, including DBPedia,[3] Wikidata[4] and QuranOntology [39] vocabularies, thereby enhancing the interconnectedness of the *SemanticHadith* knowledge graph.

However, the reconciliation of links posed challenges with some tools due to their limited compatibility with Arabic data. Nonetheless, we successfully linked significant entities such as prophets, places, and tribes with at least three external knowledge graphs: DBpedia, Wikidata, and QuranOntology. For other entities like animals, topics, plants, and events, we devised an automated approach to establish similarity between entities found in the QuranOntology Knowledge Graph by querying the graph and obtaining all instances of each category.

Expert validation was crucial to verify the discovered links and identify any potential conflicts, duplicates, or ambiguities. We found `owl:sameAs` links between DBpedia and Wikidata for some popular entities and extracted them by querying the DBpedia graph against the Wikidata entities aligned with the narrators in our dataset. Overall, these efforts resulted in substantial interlinking, as reported in Table 3 in Sections 7.1, with the establishment of links using `owl:sameAs` for entities in both DBpedia and Wikidata graphs.

3.6. Endpoints and Applications

A persistent triple store holds the graph data and interacts with other components through a SPARQL endpoint. The *SemanticHadith* KG was uploaded to the triple store by Virtuoso, enabling query capability through the SPARQL endpoint. The SPARQL endpoint service is available at http://www.semantichadith.com/sparql/.

[2]https://github.com/A-Kamran/SemanticHadith-V2
[3]https://dbpedia.org/
[4]https://www.wikidata/wiki/

4. NLP Methodology for Entity Extraction

The extraction of entities from the Sahih al-Bukhari corpus and other selected collections, such as Sahih Muslim, Ibn Maja, Sunan Abi Dawood, and Nisai, is a crucial step in the development of the SemanticHadith ontology extension. This section describes our NLP methodology, which combines custom-trained Named Entity Recognition (NER) model with expert-validated noun dictionaries to identify and extract entities relevant to our ontology accurately. There are many off-the-shelf NER models, such as CAMeLBERT-CA [40]. However, they are not very accurate for Arabic in hadith text. Hence, we need for custom NER model specifically trained on the target domain. Custom NER models can be trained on a specific corpus of text to improve the accuracy and performance of the model [29]. Hence, we used modified CANERCORPUS to train our NER model. Our implementation along with modified corpus is available at https://github.com/nigar-azhar/SemanticHadithNLP.git.

4.1. Dataset Preparation

In our study, we leveraged the CANERCORPUS dataset [28] as the foundation for entity extraction. Given its extensive coverage of hadith texts and specialised domain entities, we further customised the dataset to better align with the requirements of our ontology extension for Sahih al-Bukhari.

In collaboration with domain experts, we modified the dataset to incorporate additional concepts relevant to the hadith corpus, such as holy books, angels, significant crimes, and after-life concepts. Additionally, specific labels deemed irrelevant to our study, such as organisation names (Org), monetary values (Money), and numerical expressions (NUM), were removed or replaced to streamline the training process of our NER model.

Given the unique characteristics of Arabic script and language morphology, specific steps are employed to handle text normalisation, tokenisation, and diacritic stripping. These preprocessing techniques ensure consistency and accuracy in entity extraction from Arabic text.

4.2. Training NER Model

We utilised the state-of-the-art NLP library, spaCy, to train a custom NER model tailored to the characteristics of Arabic text in the hadith corpus. Transfer learning techniques enhance the model's performance and adaptability to domain-specific language and context. The training data consisted of annotated examples of entities such as angels, prophets, historical figures, geographical locations, and thematic topics extracted from Sahih al-Bukhari passages.

We employed transfer learning techniques to enhance the model's performance and adaptability to the domain-specific language and context of hadith texts. We fine-tuned pre-trained word embeddings and language models on our custom dataset to capture relevant linguistic patterns and semantic information.

4.3. Expert Validation and Noun Dictionaries

In collaboration with knowledge experts familiar with Islamic studies and hadith literature, we developed noun dictionaries containing lexicons of entities relevant to our ontology. These dictionaries serve as a supplementary resource for entity extraction, enabling cross-validation of extracted entities with domain-specific terminology. The extracted entities from the NER model are subsequently double-checked against the entries in the noun dictionaries to ensure accuracy and completeness. Any discrepancies or ambiguous cases are resolved through manual inspection and expert verification.

4.4. Entity Extraction Process

The entity extraction process involves applying the trained NER model to the entire corpus for all six selected collections. Each passage is analysed to identify spans of text corresponding to entities such as persons, locations, events, and thematic topics mentioned in the hadith. The extracted entities are then categorised into predefined classes, including angels, prophets, historical persons, geographical locations, events, quranic verses, and thematic topics such as crimes, pillars of Islam, articles of faith, and divine events. This categorisation aligns with the conceptual framework of the SemanticHadith ontology extension.

4.5. Handling Ambiguities and Variations in Person Names

While our NLP methodology successfully extracted specific prophets, pious caliphs, and wives of Muhammad as distinct entities with identifiable names, we encountered challenges in accurately mapping general person names to corresponding ontology instances. In particular, the variability and ambiguity inherent in Arabic naming conventions posed difficulties in systematically linking person names to specific individuals within the ontology.

Individuals may be referred to within hadith passages using various naming conventions, including first names, titles, or familial relations. This diversity in referencing extends beyond individual names and can include honorifics, epithets, or familial attributions, reflecting cultural norms and linguistic nuances. We turned to expert consultation as a vital methodology component to address this issue. Recognising the complexity of reconciling multiple name variants to a single ontology instance, we are developing a crowd-sourcing framework for expert validation. This framework aims to leverage the collective expertise of domain specialists to verify and reconcile named entities extracted from hadith texts with predefined ontology instances.

In this framework, the hadith passage, the extracted named persons and a list of already identified ontology instances, will be presented to experts for validation. Experts will assess the correspondence between named persons and ontology instances, resolving ambiguities and ensuring accurate mapping based on their contextual knowledge and expertise. By incorporating expert consultation into our methodology, we aim to enhance the accuracy and reliability of entity mapping within the SemanticHadith ontology extension, particularly for complex cases involving variations in person names and titles.

5. Design and Development of the Extended SemanticHadith Ontology

In the following subsections, we provide a comprehensive account of the design and development process of the *SemanticHadith* ontology.

5.1. Conceptual Knowledge Modelling

We follow an iterative approach in ontology engineering, where the knowledge model evolves as we formalise our representation. To model the results from Section 4, we follow the Ontology Development 101 methodology [41] to design the *SemanticHadith* ontology consisting of seven steps: (1) *Determine the scope of the ontology*, (2) *Enumerate important Terms*, (3) *Reuse existing ontologies*, (4) *Define classes and their hierarchies*, (5) *Define the class-slot properties*, (6) *Define the facets of the slots*, and (7) *Create instances*. See Sections 5.2 to 5.6 for the detailed ontology design and development process. It is worth noting that the SemanticHadith Ontology presented in this paper builds upon the foundation laid in our previous work [12], extending and refining the ontology to encompass a broader range of concepts, entities, and relationships within hadith texts.

5.2. Scope of the Ontology - Competency Questions

Based on the findings from our NLP module, we define the scope of the *SemanticHadith* ontology by formulating a set of competency questions (CQs) that outline the requirements specified in Table 1. Ren et al. propose a framework for categorising CQs into 12 patterns or archetypes [42]. Here, we present the competency questions relevant to our study and their corresponding patterns.

5.3. Reused Ontologies

In addition to reusing concepts from established ontologies, we extend our SemanticHadith ontology, building upon the foundation laid in our previous work [12]. This extension involves refining and expanding the ontology to encompass a broader range of concepts, entities, and relationships within hadith texts. To ensure maximum interoperability and leverage existing standards, we reuse concepts from established ontologies while designing the ontology for the hadith source. This approach involves obtaining a list of important terms from Hadith, which is

Competency questions	Patterns
What is the source URL for Hadith X?	What is the [DP] for a particular [CE]?
Does every hadith 'discussesTopic' a Topic?	Does every [CE1] [CE2]?
What hadith isSimilar to hadith X?	What is the [CE1] of a given [CE2]?
How many hadith narrations are 'partOf' a Hadith Collection Y?	How many [CE] are there in [PE]?
What are the types of hadith?	What are the types of [CE]?
Which hadith 'containsMentionOf' Event X?	Which [CE1] [OPE] [CE2]?
Find hadith 'discussesTopic' Topic X.	Find [CE1] with [CE2].
How many hadith 'containsMentionOf' Location X?	How many [CE1][OPE] [CE2]?
Does Hadith X 'containsMentionOf' Person Y	Does [CE1] [OPE] [CE2]?
Is there a hadith that 'containsMentionOf' of Prophet A?	Be there [CE1] with [CE2]?
Which individuals are 'mentionedIn' Event A described in a hadith?	Who [OPE] [CE]?
Are there specific entities 'mentionedIn' hadith narrated by certain individuals?	Be there [CE1] [OPE]ing [CE2]?
Which narrators have not narrated any sacred hadith?	Which [CE1] [OPE] no [CE2]?
How many Companions are mentioned in Hadith X?	How much does [CE] [DP]?
What type of hadith is Hadith X?	What type of [CE] is [I]?
Which hadith narrations have more than x number of narrators?	What [CE] has the [NM] [DP]?
What is the most narrated Topic by Narrator A?	What is the [NM] [CE1] to [OPE][CE2]?
Which topics are 'discussedIn' by at least three hadith narrations?	Which [CE1] [OPE] [QM] [CE2]?

Table 1

Competency Questions Mapped to CQ Archetypes/Patterns as identified by [42] (CE = class expression, OPE = object property expression, DP = data type property, I = individual, PE = property expression, NM = numeric modifier, QM = quantity modifier).

informed by a high-level analysis of data from the six prominent hadith collections as elaborated in Section 3.1 as well as the entities and relations from the results of our NLP pipeline. We then design an ontology to model these terms as concepts and relations, providing axioms for formally expressing their meaning. Our design process includes a review of scientific literature and existing standards, particularly focusing on ontologies in the Islamic domain, such as those [10, 13–15, 43–46].

We draw inspiration from existing vocabularies such as the Semantic Quran vocabulary and Quran ontology to model certain concepts within our ontology, including the Quranic verses, geographical and divine locations, divine events, historical groups and people cited, topics in hadith texts [39, 47]. These vocabularies provide comprehensive coverage of numerous concepts mentioned in the Quran and can be leveraged in the future for extracting additional entities from hadith.

In our ontology, we reuse classes and properties from the following established vocabularies:

1. DCMI Metadata Terms (Dublin Core) [48]: This standard ontology is utilised for representing metadata, with terms from this vocabulary employed to describe the metadata of the *SemanticHadith* ontology.
2. Schema.org [49]: Curated primarily by search engine operators, the Schema.org vocabulary enhances search engine results, making it a valuable resource. It includes concepts such as schema:Event, schema:Place, and schema:Person.
3. DBpedia [35]: DBpedia provides structured information extracted from Wikipedia projects as a central component of open knowledge graphs. It includes entities related to various events and places.
4. Wikidata [36]: A collaborative project, Wikidata serves as a free and open knowledge base that can be queried and edited by humans and machines alike. It includes information about events and places.

To reuse these vocabularies, we created sub-classes and sub-properties to some of the existing concepts from http://schema.org. For instance, we have integrated the `schema:Person` class as a super-class for the `HistoricPerson` and `Believer`, `schema:Event` as a super-class for the `DivineEvent` and `YearlyEvent` and `schema:Place` as a super-class for the `DivineLocation` and `GeographicalLocation`. Furthermore, we used `schema:partOf` and `schema: hasPart` as super properties for the object properties, including `hasChain` and `isPartOfHadith`. We link some of our classes and properties with DBpedia, wikidata and QuranOntology via the Web Ontology Language (OWL) properties of `owl:equivalentClass` and

`owl:equivalentProperty`. This decision preserves the domain-specific terminology, in addition to reusing vocabularies.

5.4. Ontology Design

Figure 4 shows the conceptual model for the *SemanticHadith* ontology. Here, we summarise the key entities and relations we chose to include in the conceptual design model of the *SemanticHadith* ontology version 2.0.1. The ontology design can easily be extended further by adding more concepts as the knowledge model matures.

5.5. Key Entities and Relations in the Extended SemanticHadith Ontology

In this section, we provide an overview of the key entities and relations incorporated into the *SemanticHadith* Ontology version 2.0.1, expanding upon the foundational concepts outlined in the original *SemanticHadith* Ontology.

- **Salah:** This class represents the Islamic ritual prayer, encompassing various forms and practices observed by Muslims.
- **GroupOfPeople:** An entity representing a collection of individuals, categorised further into historic groups of people, such as nations and tribes.
  - **HistoricGroupOfPeople:** Subclass of `GroupOfPeople` representing ancient societies or civilisations, including nations and tribes.
    - **Nation:** A group of people sharing common historical, cultural, or linguistic characteristics, forming a distinct political entity.
    - **Tribe:** A social group consisting of families or communities united by common ancestry, traditions, or territory.
- **Hadith:** Central entity in the ontology, encapsulating textual narratives attributed to the Prophet Muhammad or his companions, categorised into different types based on their origin and transmission.
- **HadithBook, HadithChapter, HadithCollection:** Entities for organising and structuring the hadith literature into books, chapters, and collections.
- **HadithText:** Represents the textual content of a hadith narrative, excluding the chain of narrators or Sanad.
- **HadithType:** Classifies hadith narrations into distinct types based on the nature of their transmission chain, including sacred, elevated, severed, and stopped hadith.
- **LivingCreation:** Represents living beings within the ontology, including angels, animals, and jinn, along with believers and historic personalities.
  - **Angel, Animal, Jinn:** Subclasses of `LivingCreation` representing different categories of living beings.
  - **schema:Person:** Subclass of `LivingCreation` representing individual human beings, further categorised into believers and historic figures.
    - **Believer:** Represents individuals who adhere to the Islamic faith, including companions of the Prophet Muhammad and other prominent believers.
    - **HistoricPerson:** Represents significant historical figures, including prophets, companions, and other notable personalities.
- **Narrator, NarratorChain, NarratorChainSegment, RootNarratorChainSegment:** Entities representing individuals involved in transmitting hadith narrations and the chains of narrators through which the narrations are transmitted.
- **schema:Event:** Represents events within the ontology, categorised into divine and yearly events.
- **schema:Place:** Represents locations within the ontology, categorised into divine and geographical locations.
- **Topic:** Represents thematic categories or subjects discussed within the hadith literature, including articles of faith, crimes, and pillars of Islam.

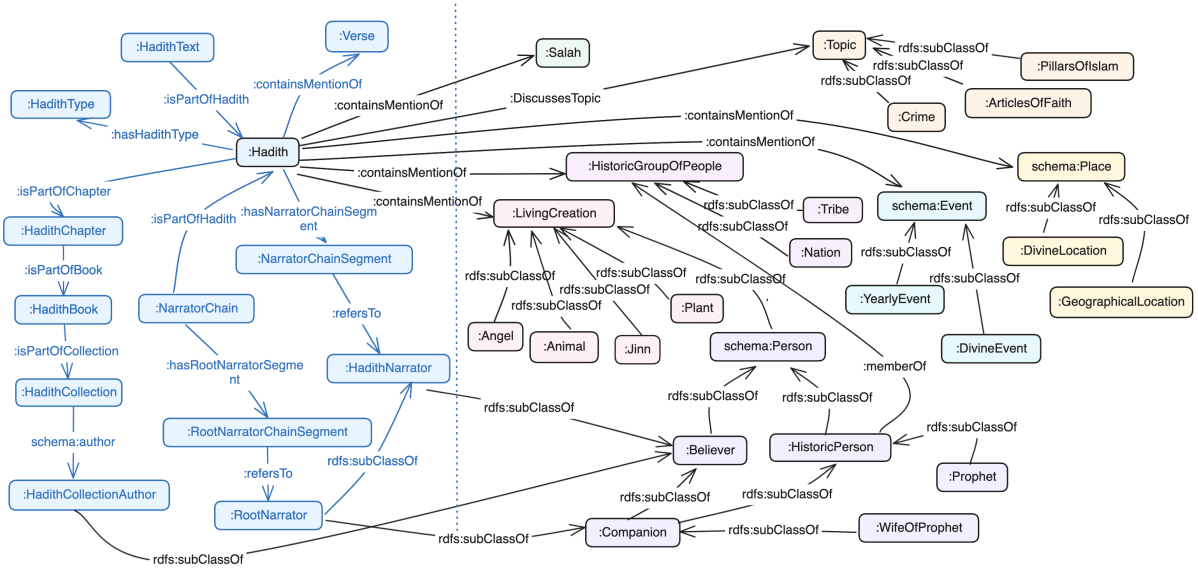


Fig. 4. Conceptual model of the *SemanticHadith* ontology version 2.0.1

- **`Verse`:** Represents verses of the Quran mentioned in hadith narrations, facilitating the linkage between Quranic text and hadith literature.

Additionally, the ontology incorporates the following relations:

- **`:isPartOf`:** Indicates the hierarchical relationship between entities, such as a hadith being part of a chapter or a chapter being part of a book.
- **`hadith discussesTopic Topic`:** Specifies the thematic topics discussed within a hadith narrative, linking the hadith entity to relevant topic entities.
- **`hadith containsMentionOf`**: Indicates the presence of mentions or references to other entities within a hadith text, facilitating semantic analysis and contextual understanding.
- **`isSimilarHadithTo`:** This relation denotes similarity between two hadith based on content, theme, or transmission chain. It establishes connections between related hadith and aids in analysing hadith literature.

5.6. *Classes, Hierarchies, Properties, and Facets*

Based on the list of significant terms identified through the analysis of hadith structure and data examination, we develop classes to represent objects with independent existence in the *SemanticHadith* ontology. Table 1 in the Supplementary Information presents the terms designated as classes in the ontology. Following a top-down approach, we initially define classes such as `Salah`, `GroupOfPeople`, `LivingCreation`, `Hadith`, etc., and then expand the ontology by defining classes and subclasses stemming from these foundational entities. These include `HadithType`, `HistoricGroupOfPeople`, `Narrator`, and `Topic`, among others.

Object properties denoting relationships or links between instances are defined for each class based on the available data. Table 2 in the Supplementary Information describes the object properties in detail. These properties establish connections between various entities in the ontology, facilitating the representation of complex relationships between different elements. Additionally, Table 3 in the Supplementary Information outlines the data properties defined in the *SemanticHadith* ontology version 2.0.1. Data properties provide detailed information about instances, such as attributes and characteristics. Moreover, links to well-known similar or related hadith are established using the property of the `:isSimilar` and `rdfs:seeAlso`. This enables the representation of connections between

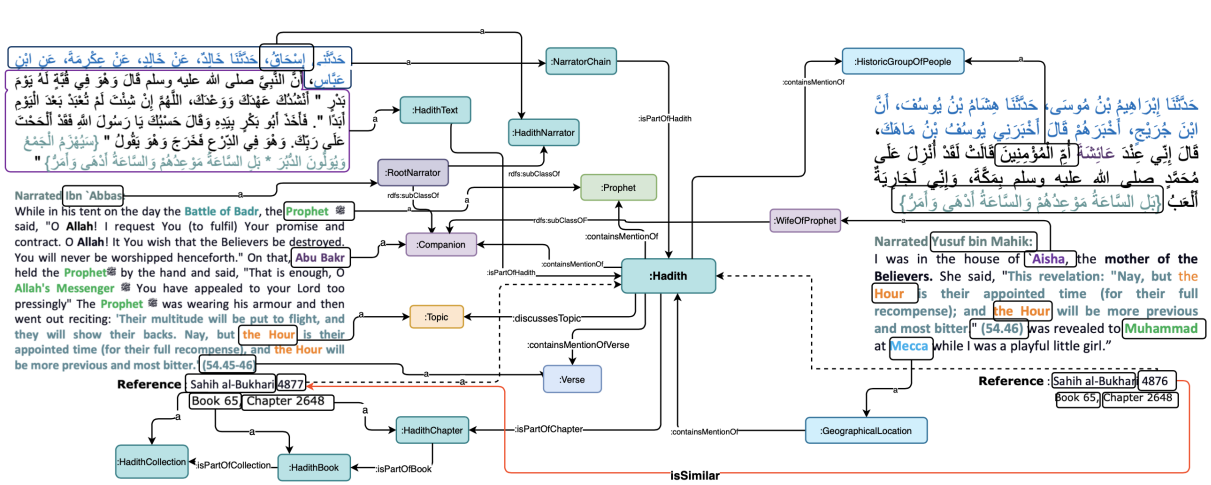


Fig. 5. Semantic Annotation of hadith text with SemanticHadith Ontology

related hadith or those repeated under different chapters within hadith collections. This systematic approach to ontology design ensures the creation of a structured and interconnected knowledge representation of Islamic literature, facilitating comprehensive exploration and analysis of hadith texts.

5.7. *Modelling Decisions*

In refining the *SemanticHadith* ontology, we made strategic design decisions to enhance its expressiveness and semantic clarity:

- **`RootNarrator`:** Recognizing that all root narrators are companions who directly reported from the Prophet Muhammad, we establish a `rdfs:subClassOf` relationship between `RootNarrator` and both `HadithNarrator` and `Companion` classes. This decision reflects the inherent relationship between root narrators, companions, and the broader category of narrators within the ontology.
- **`Companion`:** Given that companions of the Prophet Muhammad are believers and also historic figures, we refine the ontology makink `Companion` an `rdfs:subClassOf` of both `Believer` and `HistoricPerson`. This subclass relationship ensures that companions inherit the properties and characteristics associated with both believers and historic figures, providing a more nuanced representation within the ontology.

These design decisions aim to capture the intricacies of the domain while maintaining semantic coherence and consistency within the ontology structure. By refining class relationships based on domain knowledge and logical inference, the *SemanticHadith* ontology evolves to better represent the complex relationships and attributes inherent in hadith literature and Islamic history.

The last step in the Ontology101 methodology is to create instances for the classes of the ontology [41]. Our KG-Generator automatically generates these individuals for our data, assigns the corresponding values to each data property for the individuals, and then establishes the relations or links between the individuals. As depicted in Figure 5, The hadith text undergoes semantic annotation with ontology classes, facilitating enhanced comprehension and semantic querying capabilities.

5.8. *Integration and Implementation*

There are multiple ontology editors available. Amongst these widely used ones have been comprehensively compared in [50]. For this research, we selected Protégé version 5.5.0 [51] because it is extensible due to plugin support

and has interoperability with other tools and languages such as Jena, XML, HTML, etc. Protégé has built-in support for UTF-8 encoding which is ideal for our data.

We choose the hadith: prefix for the *SemanticHadith* vocabulary. We also ensure the reuse of well-established linked data vocabularies such as Schema[5] [49], and DublinCore[6] [52]. We provide equivalence relations where applicable. Some of the most relevant equivalence relations are with the DBpedia [35], Wikidata[7] [36].

According to the classification by Partridge et al. [53], we decided to opt for top-level ontologies such as `Schema` and `DC-Terms` that are more generic in their usage, and offer a low level of ontological commitment. The expressivity offered by the chosen ontologies was deemed sufficient for the *SemanticHadith* ontology.

6. Identification of Similar Hadith

One of the pivotal objectives of our study is to identify similar hadith within the Sahih al-Bukhari corpus and other selected collections, such as Sahih Muslim, Ibn Maja, Sunan Abi Dawood, and Nisai. To achieve this, we employed pre-trained Arabic sentence transformers to encode Arabic hadith texts into numerical representations, facilitating the computation of cosine similarity scores between pairs of hadith. This process has been used successfully for quranic verses [31, 33] as well as proposed for hadith [30].

6.1. Encoding and Similarity Calculation

The hadith texts were preprocessed to remove diacritics, punctuation marks, and stop words, ensuring uniformity in representation. These cleaned texts were then encoded using pre-trained sentence transformers, generating embedding vectors for each hadith. We used the asafaya bert base Arabic model, a Natural Language Processing (NLP) Model implemented in the Transformer library, using the Python programming language [54]. Subsequently, a cosine similarity matrix of dimensions 7563x7563 (where 7563 is the total number of hadith in Sahih al-Bukhari) was computed to quantify the similarity between all pairs of hadith. We received an initial dataset with similar hadith pairs identified for Bukhari from domain experts. We found the cosine similarity for each identified pair, as we planned to use it as a threshold for determining similar pairs in other collections.

6.2. Discrepancy in Similarity Bins

However, upon encoding the complete Bukhari corpus and computing the similarity bins for our expert annotated dataset of all unique pairs in 7563x7563 similarity matrix, we identified a significant discrepancy in the expected number of similar hadith to the actual number of similar hadith that fell under the threshold. The resulting similarity bins for the complete corpus (as shown in Table 2b and Figure 6b) differed substantially from our initial expectations in comparison to the ones generated using the expert dataset (as shown in Table 2a and Figure 6a), prompting a reevaluation of our methodology. The similarity values were segmented into bins based on cosine similarity scores, each representing a range of similarity values. Bins ranged from 0.0 to 1.0, with higher values indicating greater similarity between hadith pairs.

6.3. Expert Validation and Refinement

We engaged domain experts to review 100 randomly selected pairs falling within the top similarity bins (0.7-0.8, 0.8-0.9, and 0.9-1.0) to validate the identified similar hadith pairs and understand the reasons behind the observed discrepancy. The experts verified the textual similarity between these pairs and provided feedback on their relevance. During the validation process, it was observed that specific pairs exhibited high similarity scores primarily due to shared elements in the isnaad (narrator chains) rather than textual content (matan). This discrepancy led to the

[5] http://schema.org/
[6] http://dublincore.org/
[7] http://wikidata.org

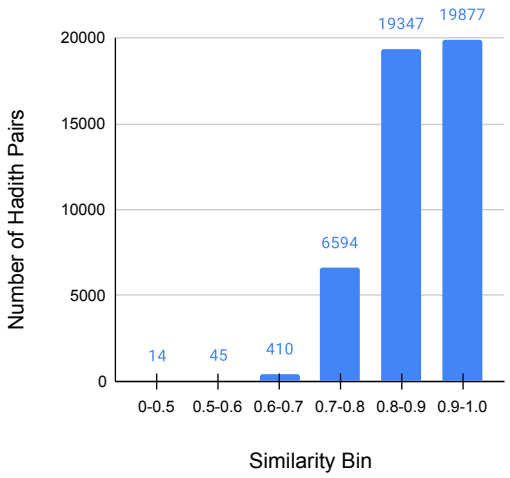
Table 2
Distribution of Hadith Pairs Across Similarity Bins in Sahih al-Bukhari.

(a) Hadith Pairs Shared by Experts

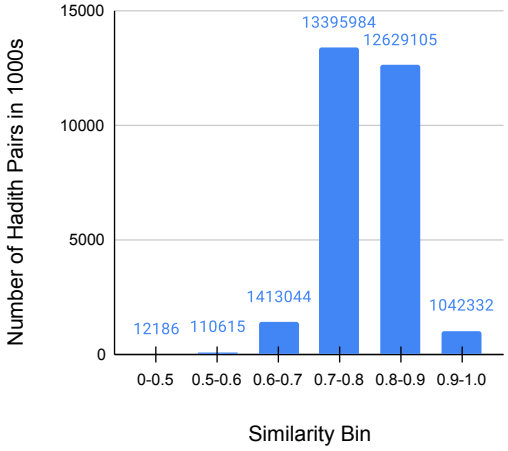
Similarity Bin	Count
0.3 - 0.4	2
0.4 - 0.5	12
0.5 - 0.6	45
0.6 - 0.7	410
0.7 - 0.8	6594
0.8 - 0.9	19347
0.9 - 1.0	19877

(b) Distribution of all unique Hadith Pairs

Similarity Bin	Count
0.2 - 0.3	42
0.3 - 0.4	1107
0.4 - 0.5	11037
0.5 - 0.6	110615
0.6 - 0.7	1413044
0.7 - 0.8	13395984
0.8 - 0.9	12629105
0.9 - 1.0	1042332



(a) Expert-Shared Hadith Pairs Distribution.



(b) Distribution of Hadith Pairs Across Similarity Bins.

Fig. 6. Distribution of Hadith Pairs Across Similarity Bins in Sahih al-Bukhari for expert-shared pairs and complete corpus.

refinement of our methodology, where future analyses will focus solely on the textual content (matan) of hadith to mitigate this issue.

Additionally, experts identified and shared similar pairs not within the top three similarity bins. We calculated there similarities and found they were below 0.7, as shown in Table 2a. We observed a number of reasons for this. For instance, Bukhari 52 and Bukhari 2051 were found to have a similarity score of 0.67656. Further analysis revealed that while the textual content (matan) of both hadith is similar, the dissimilarity arises from the differences in their respective chains of narrators. Despite sharing common elements in the matan, including distinct narrator chains led to a moderate similarity score. Another example is found in the pair Bukhari 3398 and Bukhari 3415. In this case, Bukhari 3398 represents a shorter text segment encompassed within the longer hadith, Bukhari 3415. While the textual content of Bukhari 3398 is a subset of Bukhari 3415, the latter provides additional context or elaboration beyond the content covered in Bukhari 3398. This relationship results in a variable similarity score depending on the specific segment of Bukhari 3415 being compared to Bukhari 3398.

Furthermore, experts highlighted relevant pairs such as Bukhari 69 and Bukhari 4341, which despite exhibiting dissimilar textual content, are considered relevant to each other due to their thematic or historical connections. While the matan of Bukhari 69 differs significantly from that of Bukhari 4341, the content of both hadith addresses related topics or events within Islamic tradition, leading experts to identify them as relevant to each other despite the

lack of textual similarity. These examples underscore the complexities in determining the similarity between hadith pairs, considering factors such as textual content, narrator chain composition, and thematic relevance within Islamic tradition.

6.4. Integration into Knowledge Graph

Ultimately, we chose to map only the hadith pairs from the expert dataset into our knowledge graph. However, through consultation with experts, we augmented these mappings by adding a "strongly similar" property for pairs falling into the top similarity bin (0.9-1.0 cosine similarity). This additional property enhances the representation of highly similar hadith pairs within the knowledge graph, providing a more nuanced understanding of their relationships. Moving forward, our efforts will focus on improving and identifying similar hadith pairs for all collections considered in our study. By extending our analysis to encompass additional collections such as Sahih Muslim, Ibn Maja, Sunan Abi Dawood, and Nisai, we aim to enrich the knowledge graph with a comprehensive representation of textual similarities across diverse sources of hadith literature.

6.5. Challenges and Insights

Several challenges were encountered while identifying similar hadith, including the inclusion of sanad alongside matan in the encoding process. This led to inflated similarity scores for pairs with similar sanad but distinct matan. Additionally, instances where one hadith encompassed a subset of another posed challenges in accurately determining textual similarity. Insights gained from the expert validation process highlighted the importance of considering contextual relevance beyond textual similarity. While not textually similar, certain hadith pairs were deemed relevant due to thematic or historical connections, underscoring the multifaceted nature of similarity in hadith literature.

Based on the findings and insights from the validation process, future efforts will focus on refining the methodology to prioritise textual similarity while accounting for contextual relevance. Developing a crowdsourcing framework for expert consultation also aims to enhance the accuracy and comprehensiveness of similar hadith identification across diverse collections.

7. Results and Discussion

In this section, we present the evaluation of our ontology where we perform checks for logical consistency and common design pitfalls. We also present the metrics of both the *SemanticHadith* ontology and the knowledge graph, the formal ontology design requirements, and answers to competency questions in addition to the intended applications for this endeavour.

7.1. Evaluation of SemanticHadith

The *SemanticHadith* ontology version 2.0.1 underwent a thorough evaluation to ensure its accuracy, consistency, and adherence to best practices in ontology design. Key evaluation steps and outcomes are summarised below:

- **Ontology Editing and Verification:** The classes and properties of the ontology were meticulously described in both English and Urdu. To verify correctness and consistency, the ontology was inspected using Prot'eg'e [51] and the Visual Notation for OWL Ontologies (WebVowl) tool [55].
- **Logical Consistency Checking:** The logical consistency of the ontology was validated using three reasoners: HermiT, Pellet, and FaCT++. No inconsistencies were detected during this process.
- **Pitfall Detection and Resolution:** The ontology underwent evaluation using the OntOlogy Pitfall Scanner! (OOPS!) online service [56] to identify common pitfalls in ontology design. While no major pitfalls were found, minor issues such as missing labels, inverse relationships, disjoint axioms, and naming conventions were addressed through ontology revisions.

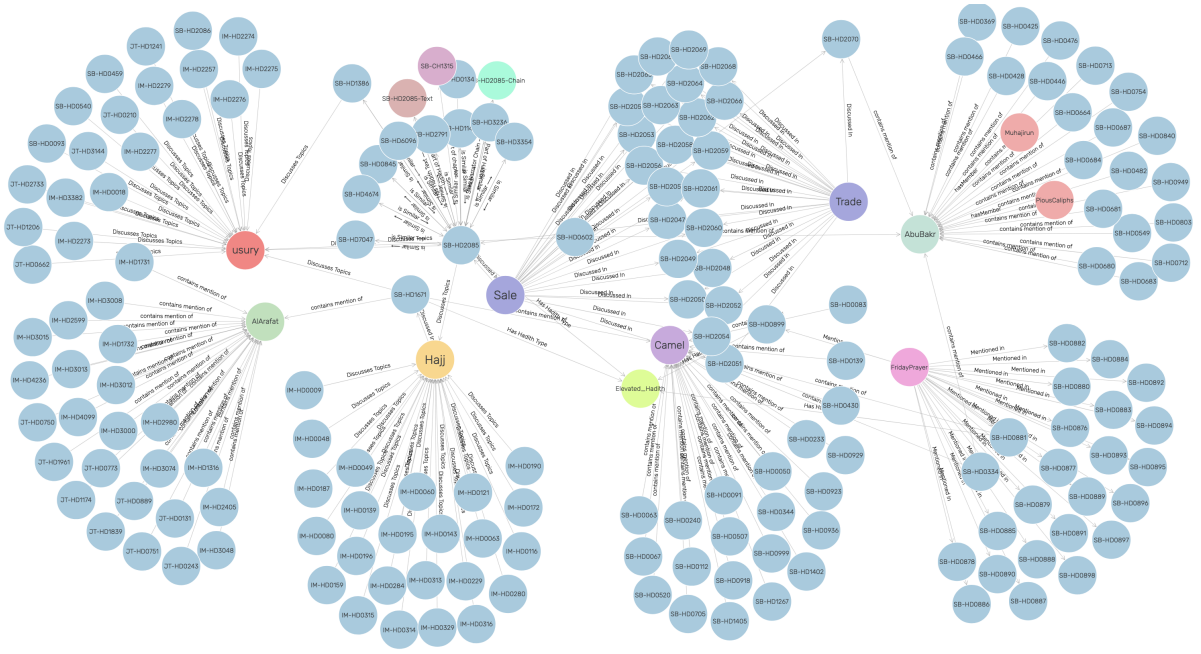


Fig. 7. Visualisation from the *SemanticHadith* Knowledge Graph.

- **MIRO Evaluation:** The Minimum Information for the Reporting of an Ontology (MIRO) guidelines [57] were applied to assess the completeness and reporting standards of the *SemanticHadith* ontology. A detailed MIRO report is available in the GitHub repository[8].
- **Knowledge Graph Correctness:** The correctness of the *SemanticHadith* knowledge graph was evaluated by answering a set of competency questions. SPARQL queries for these questions, along with their results, are provided in the GitHub repository[9]. Notably, the competency questions identified in 5.2 were successfully addressed.
- **Knowledge Graph Generation and Summary:** A knowledge graph was generated for the six prominent hadith collections. The graph's statistics, including ontology elements, axioms, and triple counts, are summarised in Table 3. Additionally, a visual representation from the *SemanticHadith* knowledge graph is presented in Figure 7.

7.2. *Intended Usage*

This section outlines the intended usage and potential applications of the *SemanticHadith* ontology and knowledge graph.

7.2.1. *Annotation of Additional Hadith Collections*

The existing implementation of the *SemanticHadith* ontology has successfully utilised natural language processing (NLP) techniques to annotate hadith texts. While the current focus has been on annotating specific hadith collections, the methodology and infrastructure established can be extended to annotate additional hadith collections. By leveraging NLP technologies, the annotation process can be automated to a significant extent, enabling the efficient annotation of large-scale hadith corpora. This expansion would result in a more comprehensive and interconnected repository of annotated hadith texts, facilitating advanced research and analysis in Islamic studies.

[8] https://github.com/A-Kamran/SemanticHadith-V2/blob/main/MIRO.md
[9] https://github.com/A-Kamran/SemanticHadith-V2/blob/main/CompetencyQuestionsAndSPARQLQueries.md

Table 3
Statistics of the *SemanticHadith* ontology and the *SemanticHadith* knowledge graph.

	Variables	Number
Structure & Ontology	Ontology Classes	43
	Object Properties	34
	Datatype Properties	45
	Annotations	134
Knowledge Graph	Number of Axioms	4,385,110
	Total Entities	303869
	Hadith	34,458
	Person	6822
Internal Links for Hadith	`:discussesTopic`, Topic	20000
	`:containsMentionOf`, Verse	4000
	`:containsMentionOf`, LivingCreation	6733
	`:isSimilar`, Hadith	47496
External Links to Wikidata &/or DBPedia	`owl:sameAs`, Places	34
	`owl:sameAs`, Topics	20
	`owl:sameAs`, Person	634
	`owl:sameAs`, Prophet	23
External Links to Quran Ontology	`rdfs:seeAlso`	62
	`owl:sameAs`	200

7.2.2. Enhanced Knowledge Exploration

Integrating NLP annotations into the *SemanticHadith* ontology opens up new possibilities for knowledge exploration and discovery. Researchers can leverage the annotated data to gain insights into various aspects of Islamic knowledge, including the relationships between entities, events, and concepts mentioned in hadith texts. By applying semantic querying techniques, users can explore the annotated corpus in depth, uncovering hidden connections and patterns within the data. Furthermore, the annotated ontology provides a foundation for the development of advanced semantic search and recommendation systems tailored to the needs of scholars and researchers in the Islamic domain.

7.2.3. Cross-Domain Integration

Beyond the realm of Islamic studies, the annotated *SemanticHadith* ontology holds potential for cross-domain integration with other knowledge domains. The ontology facilitates interdisciplinary research and knowledge discovery by linking annotated hadith texts to relevant entities and concepts in external knowledge graphs, such as DBpedia and Wikidata. This cross-domain integration opens up opportunities for exploring connections between Islamic knowledge and other fields, including history, philosophy, linguistics, and cultural studies. Researchers across various disciplines can benefit from the enriched semantic annotations provided by the *SemanticHadith* ontology, enabling them to leverage Islamic knowledge in novel and interdisciplinary research endeavours.

7.2.4. Educational Applications

The annotated *SemanticHadith* ontology serves as a valuable resource for educational and pedagogical applications in Islamic studies. By providing a structured and semantically enriched representation of hadith texts, the ontology supports interactive learning experiences, digital scholarship, and curriculum development in academic institutions and educational settings. Educators can utilise the ontology to create customised learning materials, interactive quizzes, and educational tools that engage students with authentic hadith texts in a meaningful and contextually rich manner. Furthermore, the availability of annotated hadith data in linked data format enables learner-sourcing initiatives, where students and scholars contribute to the annotation and enrichment of the ontology through collaborative efforts, thereby fostering a culture of knowledge sharing and co-creation within the academic community.

7.2.5. Future Directions

Moving forward, the *SemanticHadith* project aims to enhance the annotation pipeline further and expand the scope of annotated data. By incorporating state-of-the-art NLP techniques and machine learning algorithms, the project seeks to improve the accuracy and efficiency of the annotation process, enabling the annotation of diverse hadith collections and genres. Additionally, efforts will be directed towards enriching the ontology with additional metadata, such as provenance information, temporal data, and linguistic annotations, to provide a more comprehensive and contextually rich representation of annotated hadith texts.

Furthermore, recognising the challenge posed by variations in naming conventions within hadith passages and the need for accurate entity mapping, we are developing a crowdsourcing framework for expert validation. This framework aims to leverage the collective expertise of domain specialists to verify and reconcile named entities extracted from hadith texts with predefined ontology instances. In this framework, experts will assess the correspondence between named persons and ontology instances, resolving ambiguities and ensuring accurate mapping based on their contextual knowledge and expertise. By incorporating expert consultation into our methodology, we aim to enhance the accuracy and reliability of entity mapping within the SemanticHadith ontology extension, particularly for complex cases involving variations in person names and titles.

Collaboration with domain experts, scholars, and stakeholders will guide the evolution and refinement of the annotated *SemanticHadith* ontology, ensuring its relevance and usability in diverse research and educational contexts.

8. Conclusion

In conclusion, our paper presents a comprehensive methodology for generating a knowledge graph from the hadith corpus, addressing key challenges in entity extraction, similarity computation, and knowledge graph construction. By leveraging NLP techniques, expert validation, and ontology engineering, we have successfully extracted entities, identified similar hadith, and enriched the SemanticHadith knowledge graph. We ensured accuracy in entity extraction through meticulous data selection, preprocessing, and custom NER model training, laying the foundation for a robust knowledge graph. Identifying similar hadith, facilitated by cosine similarity computation and expert validation, provided insights into textual similarities and thematic connections within hadith literature. Furthermore, our methodology includes conceptual knowledge modelling and formalisation, ensuring interoperability and interpretability of the knowledge graph. By interlinking with the LOD Cloud and providing an endpoint for SPARQL queries, we enhance accessibility and usability, fostering further research and applications in Islamic studies and related fields. Overall, our study contributes to advancing knowledge graph generation from textual sources, particularly in the domain of Islamic knowledge. Our framework facilitates efficient information retrieval and exploration and opens avenues for interdisciplinary research and the development of intelligent applications in religious studies and beyond.

Additional Information

Supplementary Information accompanies this paper.

Data Availability

Ontology, Knowledge Graph, ontology documentation, SPARQL Queries corresponding to Competency Questions, MIRO report https://github.com/A-Kamran/SemanticHadith-V2. The implementation of the Entity recognition framework along with the modified corpus is available at https://github.com/nigar-azhar/SemanticHadithNLP.git.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Q. Wang, Z. Mao, B. Wang and L. Guo, Knowledge graph embedding: A survey of approaches and applications, *IEEE Transactions on Knowledge and Data Engineering* **29**(12) (2017), 2724–2743.
[2] X. Chen, S. Jia and Y. Xiang, A review: Knowledge reasoning over knowledge graph, *Expert Systems with Applications* **141** (2020), 112948.
[3] Y. Shang, Y. Tian, M. Zhou, T. Zhou, K. Lyu, Z. Wang, R. Xin, T. Liang, S. Zhu and J. Li, EHR-Oriented Knowledge Graph System: Toward Efficient Utilization of Non-Used Information Buried in Routine Clinical Practice, *IEEE Journal of Biomedical and Health Informatics* **25**(7) (2021), 2463–2475.
[4] A. Basharat, B. Abro, I.B. Arpinar and K. Rasheed, Semantic Hadith: Leveraging Linked Data Opportunities for Islamic Knowledge., in: *LDOW@ WWW*, 2016.
[5] S. Hasan, *An introduction to the science of Hadith*, Al-Quran Society London, 1994.
[6] J. Brown, How We Know Early Hadīth Critics Did Matn Criticism and Why It's So Hard to Find, *Islamic Law and Society* **15**(2) (2008), 143–184.
[7] K. Dukes, E. Atwell and A.-B.M. Sharaf, Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank., in: *LREC*, 2010.
[8] 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.
[9] A. Farghaly and K. Shaalan, Arabic natural language processing: Challenges and solutions, *ACM Transactions on Asian Language Information Processing (TALIP)* **8**(4) (2009), 1–22.
[10] A. Hakkoum and S. Raghay, Ontological approach for semantic modeling and querying the Qur'an, in: *Proceedings of the International Conference on Islamic Applications in Computer Science And Technology*, 2015.
[11] S. Altammami, E. Atwell and A. Alsalka, Towards a Joint Ontology of Quran and Hadith, *International Journal on Islamic Applications in Computer Science And Technology* (2020).
[12] A.B. Kamran, B. Abro and A. Basharat, SemanticHadith: An ontology-driven knowledge graph for the hadith corpus, *Journal of Web Semantics* **78** (2023), 100797.
[13] A. Azmi and N.B. Badia, iTree-Automating the construction of the narration tree of Hadiths (Prophetic Traditions), in: *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)*, IEEE, 2010, pp. 1–7.
[14] R.S. Baraka and Y. Dalloul, Building Hadith ontology to support the authenticity of Isnad, *Building Hadith ontology to support the authenticity of Isnad* **2**(1) (2014).
[15] A. Al-Rumkhani, M. Al-Razgan and A. Al-Faris, TibbOnto: Knowledge Representation of Prophet Medicine (Tibb Al-Nabawi), *Procedia Computer Science* **82** (2016), 138–142.
[16] F. Harrag and A. Hamdi-Cherif, UML modeling of text mining in Arabic language and application to the prophetic traditions "Hadiths", *The 1st international sysmposium on computers and Arabic language and exhibition, KACST & SCS* (2007), 11–20.
[17] A.A.B. Philips, *Usool At-Tafseer: the Methodology of Qur'anic Interpretation*, AS Noordeen, 2002.
[18] B. Fairouz, T. Nora and A.A. Nouha, An Ontological Model of Hadith Texts, in: *International Journal of Advanced Computer Science and Applications, Vol. 11, No. 4, 2020*, 2020.
[19] A.H. Jaafar and N. Che Pa, Hadith commentary repository: An ontological approach, in: *Proceedings of the 6th International Conference on Computing and Informatics, ICOCI 2017*, 2016.
[20] M. Alkhatib, A.A. Monem and K. Shaalan, A Rich Arabic WordNet Resource for Al-Hadith Al-Shareef, *Procedia Computer Science* **117** (2017), 101–110.
[21] A.M. Azmi, A.O. Al-Qabbany and A. Hussain, Computational and natural language processing based studies of hadith literature: a survey, *Artificial Intelligence Review* **52**(2) (2019), 1369–1414.
[22] F. Harrag, Text mining approach for knowledge extraction in Sahîh Al-Bukhari, *Computers in Human Behavior* **30** (2014), 558–566.
[23] A. Al-Arfaj and A. Al-Salman, Towards ontology construction from Arabic texts-a proposed framework, in: *2014 IEEE International Conference on Computer and Information Technology*, IEEE, 2014, pp. 737–742.

[24] K.A. Aldhlan, A.M. Zeki and A.M. Zeki, Datamining and Islamic knowledge extraction: alhadith as a knowledge resource, in: *Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World (ICT4M) 2010*, IEEE, 2010, p. H–21.
[25] M. Naji Al-Kabi, G. Kanaan, R. Al-Shalabi, S.I. Al-Sinjilawi and R.S. Al-Mustafa, Al-Hadith text classifier, *Journal of Applied Sciences* **5**(3) (2005), 584–587.
[26] S. Saeed, S. Yousuf, F. Khan and Q. Rajput, Social network analysis of Hadith narrators, *Journal of King Saud University - Computer and Information Sciences* (2021). doi:https://doi.org/10.1016/j.jksuci.2021.01.019.
[27] I. Bounhas, On the usage of a classical Arabic corpus as a language resource: related research and key challenges, *ACM Transactions on Asian and low-resource language information processing (TALLIP)* **18**(3) (2019), 1–45.
[28] R.E. Salah and L.Q.B. Zakaria, Building the classical Arabic named entity recognition corpus (CANERCorpus), in: *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, IEEE, 2018, pp. 1–8.
[29] I.K. Alshammari, E. Atwell and M.A. Alsalka, Evaluation of Arabic Named Entity Recognition Models on Sahih Al-Bukhari Text, Technical Report, EasyChair, 2023.
[30] P. Huang, A. Basharat, U. Nisar and K. Rasheed, Interlinking Hadith Based on Multilingual Text Similarity Analysis, in: *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, The Steering Committee of The World Congress in Computer Science, Computer . . . , 2018, pp. 377–383.
[31] A. Basharat, D. Yasdansepas and K. Rasheed, Comparative study of verse similarity for multi-lingual representations of the qur'an, in: *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, The Steering Committee of The World Congress in Computer Science, Computer . . . , 2015, pp. 336–343.
[32] M. Alshammeri, E. Atwell and M. ammar Alsalka, Detecting semantic-based similarity between verses of the Quran with Doc2vec, *Procedia Computer Science* **189** (2021), 351–358.
[33] M. Alshammeri, E. Atwell and M.A. Alsalka, A Siamese Transformer-based Architecture for Detecting Semantic Similarity in the Quran, in: *The International Journal on Islamic Applications in Computer Science And Technology-IJASAT*, Vol. 9, Design For Scientific Renaissance, 2021.
[34] Ontotext AD, OntoRefine (Version 1.2), 2022. https://ontotext.com/products/graphdb/graphdb-free/.
[35] C. Becker and C. Bizer, DBpedia mobile-a location-aware semantic web client, *Proceedings of the Semantic Web Challenge* (2008), 13–16.
[36] D. Vrandečić and M. Krötzsch, Wikidata: A Free Collaborative Knowledgebase, in: *Proceedings of the 2014 ACM conference on Web science*, ACM, 2014, pp. 106–107.
[37] A.-C. Ngonga Ngomo, M.A. Sherif, K. Georgala, M.M. Hassan, K. Dreßler, K. Lyko, D. Obraczka and T. Soru, LIMES: a framework for link discovery on the semantic web, *KI-Künstliche Intelligenz* (2021), 1–11.
[38] K. Ham, OpenRefine (version 2.5). http://openrefine. org. Free, open-source tool for cleaning and transforming data, *Journal of the Medical Library Association: JMLA* **101**(3) (2013), 233.
[39] A. Hakkoum, The Quran Schema vocabulary.
[40] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor and N. Habash, The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models, in: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Association for Computational Linguistics, Kyiv, Ukraine (Online), 2021.
[41] N.F. Noy, D.L. McGuinness et al., Ontology development 101: A guide to creating your first ontology, Stanford knowledge systems laboratory technical report KSL-01-05 and . . . , 2001.
[42] Y. Ren, A. Parvizi, C. Mellish, J.Z. Pan, K. Van Deemter and R. Stevens, Towards competency question-driven ontology authoring, in: *European Semantic Web Conference*, Springer, 2014, pp. 752–767.
[43] H.S. Al-Khalifa, M.M. Al-Yahya, A. Bahanshal and I. Al-Odah, SemQ: A proposed framework for representing semantic opposition in the Holy Quran using Semantic Web technologies, in: *2009 International Conference on the Current Trends in Information Technology (CTIT)*, IEEE, 2009, pp. 1–4.
[44] S. Saad, N. Salim and H. Zainal, Islamic knowledge ontology creation, in: *2009 International Conference for Internet Technology and Secured Transactions,(ICITST)*, IEEE, 2009, pp. 1–6.
[45] S. Saad, N. Salim, H. Zainal and Z. Muda, A process for building domain ontology: An experience in developing Solat ontology, in: *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics. Bandung, Indonesia*, 2011, pp. 1–5.
[46] A.R. Yauri, R.A. Kadir, A. Azman and M.A.A. Murad, Ontology semantic approach to extraction of knowledge from Holy Quran, in: *2013 5th International Conference on Computer Science and Information Technology*, IEEE, 2013, pp. 1–5.
[47] M. Sherif, The Quran Schema vocabulary.
[48] D.U. Board, DCMI Metadata Terms.
[49] R.V. Guha, D. Brickley and S. Macbeth, Schema. org: evolution of structured data on the web, *Communications of the ACM* **59**(2) (2016), 44–51.
[50] E. Alatrish, Comparison Some of Ontology, *Journal of Management Information Systems* **8**(2) (2013), 018–024.
[51] S. University, PROTÉGÉ.
[52] D.C.M. Initiative et al., Dublin core metadata element set, version 1.1, Dublin Core Metadata Initiative, 2012, [Online; accessed 20. Aug. 2022].
[53] C. Partridge, A. Mitchell, A. Cook, J. Sullivan and M. West, A Survey of Top-Level Ontologies-to inform the ontological choices for a Foundation Data Model (2020).

[54] A. Safaya, M. Abdullatif and D. Yuret, KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media, in: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 2054–2059. https://www.aclweb.org/anthology/2020.semeval-1.271.
[55] S. Lohmann, V. Link, E. Marbach and S. Negru, WebVOWL: Web-based visualization of ontologies, in: *International Conference on Knowledge Engineering and Knowledge Management*, Springer, 2014, pp. 154–158.
[56] M. Poveda-Villalón, M.C. Suárez-Figueroa and A. Gómez-Pérez, Validating ontologies with oops!, in: *International conference on knowledge engineering and knowledge management*, Springer, 2012, pp. 267–281.
[57] N. Matentzoglu, J. Malone, C. Mungall and R. Stevens, MIRO: guidelines for minimum information for the reporting of an ontology, *Journal of biomedical semantics* **9**(1) (2018), 1–13.