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, structured representation of knowledge has become crucial for efficient information retrieval and utilisation across diverse domains. Knowledge graphs (KGs), with their graph-based organisation of entities and relationships, have gained significant attention for their ability to revolutionise information management and retrieval [1, 2]. Although knowledge graphs have been widely adopted in fields such as biomedicine [3], education [4], and cultural heritage [5], their application to religious domains, particularly Islamic knowledge, remains under-explored. Generating

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

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 to represent Islamic knowledge [6].

The Islamic knowledge domain, rooted in the Quran and the sunnah of the Prophet Muhammad, presents a vast repository of textual resources that remain under-represented in the semantic web ecosystem [7]. Among these resources, the hadith—narrations of the Prophet's sayings, actions, and approvals—constitute a vital source for understanding Islamic teachings and jurisprudence [8]. While prior studies have addressed specific subdomains, such as prophetic medicine [9], chain-of-narrators representation [10], and thematic categorisation [11], these efforts often operate in isolation, lacking a unified semantic framework. Similarly, ontological models [12, 13] and WordNet-like linguistic resources [14] frequently suffer from limited interoperability and interdisciplinary applicability. Our previous work, *SemanticHadith*, addresses part of this challenge by systematically modelling the structure of hadith and narrator chains [15]. However, the absence of comprehensive semantic representation for hadith texts continues to hinder their seamless integration into the broader semantic web landscape, limiting effective exploration and retrieval of Islamic knowledge. Overcoming these limitations remain critical to advancing the accessibility and interconnectedness of Islamic knowledge resources.

This study addresses the broader challenges of hadith modelling by introducing a comprehensive methodology for generating a knowledge graph from the hadith corpus. By leveraging advanced Natural Language Processing (NLP) techniques, such as Named Entity Recognition (NER) and relationship extraction, our methodology enables a granular understanding of entities, narrators, and their semantic relationships. Tools like spaCy and transfer learning models tailored to Classical Arabic are employed, alongside preprocessing techniques such as diacritic normalisation, tokenisation, and morphological analysis, to address the linguistic diversity inherent in Arabic texts. However, our methodology operates under several assumptions, including a reliance on standardised textual sources in Classical Arabic and expert validation to resolve ambiguities. Limitations such as linguistic diversity, regional dialects, and varying levels of authenticity across hadith collections remain challenges for accurate semantic modelling. Despite these limitations, the proposed approach significantly advances the representation of Islamic knowledge by integrating principles of linked data and external knowledge graphs, such as Wikidata and DBpedia, enabling connections between Islamic studies and fields like history, linguistics, and cultural studies. Such integration is critical for fostering interdisciplinary research and making Islamic knowledge resources accessible to a wider audience. Unlike traditional approaches, which rely on keyword-based searches and static indexing, this framework provides dynamic interlinking of entities and thematic contexts. For instance, a thematic query like 'charity' retrieves not only narrations that explicitly mention giving or donations but also those addressing related topics such as the virtues of helping neighbours, assisting the poor, and the prohibition of hoarding wealth. These connections, enabled by semantic relationships within the knowledge graph, go beyond traditional keyword searches by uncovering indirect but meaningful links between narrations. This granularity and interconnectedness enable researchers, students, and educators to explore Islamic texts systematically and intuitively. The inclusion of external knowledge sources further enhances interdisciplinary research, bridging Islamic studies with fields such as linguistics, cultural studies, and history.

The *SemanticHadith* ontology and knowledge graph provide practical benefits that cater to diverse audiences. Scholars and researchers can use the framework to perform advanced semantic queries, uncovering patterns and relationships within the corpus that were previously difficult to identify. Educators and students benefit from interactive tools that facilitate thematic exploration or the creation of tailored pedagogical resources. The ontology's scalability and expert-validated mappings enhance its reliability, making it a valuable resource for both rigorous academic study and public engagement. By integrating state-of-the-art NLP technologies with domain expertise, this framework enables an accessible, interactive, and context-rich exploration of Islamic knowledge.

We present the *SemanticHadith* ontology version 2.0.1, an enhanced iteration that builds upon the foundation of version 1.0.1. This updated ontology expands the modelling of entities and topics within hadith

texts, enabling a more nuanced understanding of Islamic teachings. By formalising this extensive knowledge repository and interlinking its components, the ontology supports new avenues for research, discovery, and synthesis within the Islamic knowledge domain. This emphasis on comprehensive semantic modelling highlights the importance of integrating various textual resources, including Quranic commentaries that heavily rely on hadith corpora for interpreting Quranic verses. Such integration enriches the exploration of Islamic knowledge, fostering a deeper understanding of the interconnectedness across different facets of Islamic teachings. Through this work, we aim to address challenges in knowledge acquisition and semantic content creation, particularly for applications dependent on advanced semantic technologies. By adopting a semantic perspective, the framework facilitates more effective search and discovery of concepts and relationships within hadith texts.

2. Background Context and Motivation

In Islamic tradition, hadith serves as a vital source of knowledge, offering narratives of historical events from the life of Prophet Muhammad, interpretations of Quranic verses, and elaborations of essential Islamic concepts. Second only to the Quran, the hadith corpus significantly influences Islamic jurisprudence and understanding. Each hadith consists of two primary components: the matan (narration content) and the sanad (chain of narrators). The sanad, presented as a chronological list of narrators, is instrumental in assessing the authenticity of a hadith, as scholars evaluate the integrity of the chain to determine its reliability. The expansive nature of the hadith corpus presents challenges in managing its intricate relationships and concepts. Beyond the primary texts, a substantial body of supplementary literature, including commentaries and biographical records, adds layers of complexity. Navigating these resources requires not only identifying narrators and texts but also a deeper comprehension of the relationships across both the hadith and the Quran, such as which hadith elaborate specific Quranic verses or share thematic similarities.

Historically, the study of Islamic knowledge has relied on unstructured textual resources, making systematic exploration and analysis arduous. Despite its critical role in shaping Islamic thought, the hadith corpus remains underutilised within the semantic web ecosystem. Advances in semantic modelling and knowledge graph construction present transformative opportunities for organising and linking Islamic texts. Expanding beyond canonical hadith collections to include Quranic commentaries (tafsir), biographies of narrators, Islamic jurisprudence (fiqh), and classical scholarly works could foster a more comprehensive understanding of Islamic tradition. Moreover, multilingual knowledge graphs would make these resources accessible to diverse audiences worldwide, addressing linguistic and cultural nuances to enhance inclusivity. Cross-domain integration with datasets from fields such as history, sociology, and science offers the potential for interdisciplinary research, uncovering novel insights into the influence of Islamic knowledge on global history and culture. Future research should also prioritise the development of intuitive tools, such as AI-driven reasoning systems and natural language search interfaces, to empower scholars, students, and the general public in engaging seamlessly with Islamic knowledge graphs. These directions highlight the potential for semantic modelling to advance Islamic studies by combining traditional knowledge with modern computational methods, offering unprecedented opportunities for exploration, education, and cross-disciplinary collaboration.

2.1. Importance of Hadith

Understanding the significance of hadith requires integrating principles of Quranic understanding and the science of exegesis (tafsir), as Quranic verses often depend on hadith for contextualisation and elaboration. Tafsir relies heavily on authentic hadith to clarify historical contexts, reasons for revelation, and essential concepts not immediately apparent from the Quranic text. For example, comprehensive commentaries such as Tafsir al-Tabari frequently reference hadith to elucidate meanings. Adhering to these principles is critical for producing accurate interpretations of the Quran [16].

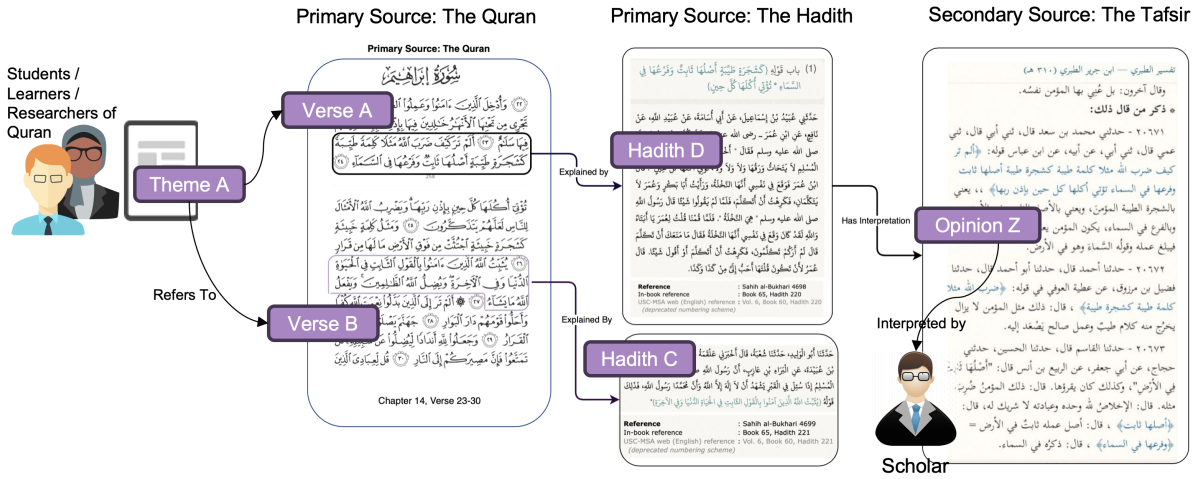


Figure 1. Motivational Scenario - Connecting primary and secondary Islamic knowledge sources

2.2. Need for Formalised Semantic Modelling

Navigating Islamic knowledge—spanning Quranic texts, hadith literature, and tafsir—presents significant challenges. For instance, a student seeking to explore the connections between verses in Surah Ibrahim may instinctively turn to hadith literature for explanations, supplemented by renowned tafsir such as Tafsir al-Tabari. Figure 1 illustrates this scenario and highlights the necessity for structured knowledge modelling tailored to Islamic texts. By organising Quranic verses and related hadith explanations within a knowledge graph framework, scholars and learners can systematically explore the intricate links between primary and secondary sources.

Hadith complements the Quran by elaborating on its verses and providing practical guidance. Over centuries, thousands of commentaries have been produced in various languages, all of which rely on hadith to interpret Quranic teachings. Formalising this extensive repository through semantic modelling and KG generation promises to unlock new avenues for research and synthesis. In Figure 2, we provide an example from both the Quranic and hadith texts, highlighting the same entities and topics. This visualisation of linking of shared entities and topics across Quranic and hadith texts, emphasise the interconnected nature of Islamic knowledge and the importance of a unified framework. Other examples of hadith explicitly explaining Quranic principles, demonstrating the interconnectedness of the two sources of Islamic knowledge include:

Elaboration on Quranic Verses

- **Wudu (Ablution)**
  Quranic Verse: "O you who have believed, when you rise to [perform] prayer, wash your faces and your forearms to the elbows and wipe over your heads and wash your feet to the ankles..." (Surah Al-Ma'idah, 5:6)
  Hadith: The Prophet provided a practical demonstration of wudu: "Whoever performs ablution as I do, and then prays two rak'ahs without allowing his thoughts to wander, will have all his past sins forgiven." (Sahih Bukhari, 159)
  This expands on the Quranic verse by illustrating how to perform wudu properly and its spiritual benefits.

Explanation of Islamic Concepts

- **Definition of Faith (Iman)**
  Quranic Verse: "The believers are only those who, when Allah is mentioned, their hearts become

fearful, and when His verses are recited to them, it increases them in faith." (Surah Al-Anfal, 8:2)
Hadith: The Prophet defined Iman succinctly: "Iman is to believe in Allah, His angels, His books, His messengers, the Last Day, and to believe in divine decree, both its good and its bad." (Sahih Muslim, 8)

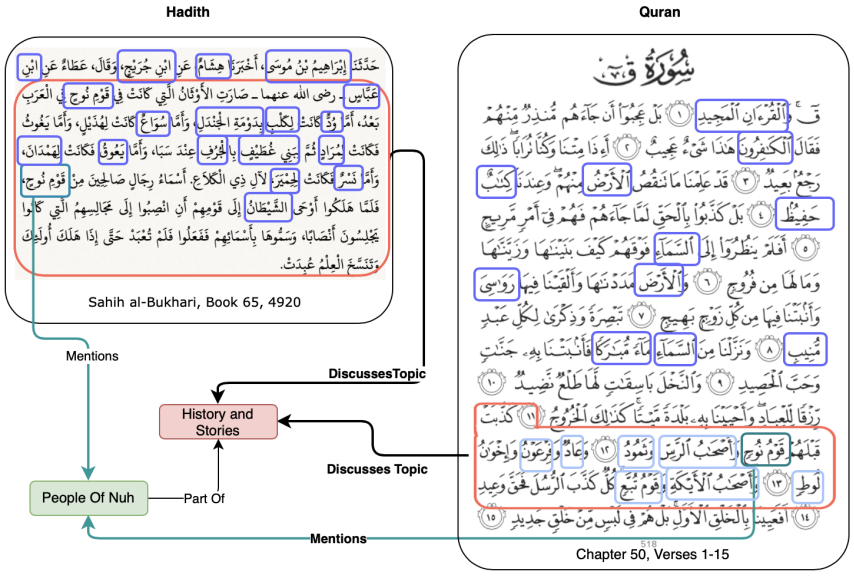


Figure 2. Motivational Scenario 2 - Connecting Themes, topics, people and places in the Quran and hadith

2.3. Advances and Challenges in Semantic Modelling of Islamic Knowledge

Efforts to model and publish Islamic knowledge as linked data have predominantly focused on the Quran, while the hadith corpus remains relatively under-explored. Existing research spans various aspects of hadith, including thematic categorisation [11], jurisprudential classification [12, 17], and linguistic analysis [18, 19].For instance, Fairouz et al. [20] modelled hadith commentaries, while Jaafar and Che Pa [13] concentrated on concepts within Arabic hadith texts. Similarly, Khatib et al. [14] proposed a WordNet-like linguistic resource, and Azmi et al. [21] provided a comprehensive review of computational techniques applied to hadith literature.

Despite this diversity, the field remains fragmented. While works such as Dukes et al. [19] explored syntactic and morphological annotations for Arabic texts, and Al-Khalifa et al. [22] proposed semi-automated ontology construction for Quranic verses, these efforts lack integration into a unified semantic framework. Farghaly and Shaalan [18] highlighted the under-representation of Islamic textual resources in Arabic NLP tools, whereas Hakkoum and Raghay [17] focused narrowly on ontologies for Islamic jurisprudence. More recent contributions include Altammami et al.'s annotated Arabic-English corpus for hadith [23] and their work on text segmentation [24] and ontological modelling [11]. Additionally, association rule mining techniques have been employed for jurisprudential ontology [12], similar to the framework proposed by Al-Arfaj and Al-Salman [25] to create an ontology from Arabic texts. Automated parsing and classification tools have further contributed to the field. E-Narrator [10] focused on parsing hadith text and generating graphical representations of narrator chains. Building on E-Narrator, HadithRDF [26] extended this approach by developing an ontology-based system for assessing isnad authenticity based on scholarly rules. Data mining techniques for indexing hadith collections by Aldhlan et al. [27] and Naji et al. [28] and social network analyses of narrators by Saeed et al. [29] further illustrate the diversity of computational approaches.

However, these advancements primarily address isolated challenges, such as narrator chains or thematic categorisation, and fall short of providing comprehensive, multilingual, and publicly accessible knowledge graphs for hadith. While they contribute important foundational work, they lack integration into broader frameworks that can dynamically interlink entities and themes across multiple domains. The lack of standardised numbering schemes, inconsistent citation practices, and varying levels of authenticity hinder dataset unification. Moreover, the linguistic complexity of Arabic—encompassing regional dialects, intricate morphology, and the layered structure of sanad (chain of narrators) and matan (narration content)—adds to the difficulties of semantic modelling. Nevertheless, emerging computational techniques, such as NLP, NER and similarity computations, hold promise for addressing these challenges. By formalising and interlinking the vast hadith corpus, researchers can pave the way for a more structured, dynamic, and accessible repository of Islamic knowledge.

2.4. NLP Techniques for Processing Hadith Text

The complex structure and linguistic richness of hadith texts demand advanced NLP techniques for effective semantic modelling. Existing methods provide a solid foundation for extracting entities and relationships, which we adapt and fine-tune to meet the specific challenges posed by the hadith literature. Preprocessing steps such as text normalisation (e.g., standardising variations in Arabic script), diacritical mark removal, and sentence tokenisation ensure uniformity across the corpus. Domain-specific tools, such as Arabic tokenisers, are employed to accommodate the intricate morphology of the language. Segmenting hadith into sanad (chain of narrators) and matan (narrative content) further supports targeted semantic analysis and enhances downstream tasks, such as entity extraction and relationship mapping. Previous work, such as Azmi et al. [21] and Bounhas [30], have emphasised the importance of creating multilingual hadith resources for tasks like NLP, information retrieval, and knowledge extraction. Building on this, we adapt and extend preprocessing workflows to ensure high-quality input for our entity extraction and modelling pipelines.

NER plays a central role in extracting key entities such as narrators, locations, and thematic concepts within hadith. Prior studies, such as Salah et al. [31], have used NER to extract narrators and associated events, demonstrating the technique's ability to support semantic representations in hadith. However, the complexities of Arabic—including its variable word order, rich morphology, and diacritical marks—necessitate the use of domain-adapted NER systems. For instance, custom models trained on datasets like CANERCORPUS enable more precise entity extraction in this domain. These approaches ensure consistent linkage of entities, such as narrators like "Abu Huraira," to their corresponding nodes in the knowledge graph, improving semantic accuracy. In our work, we fine-tune existing NER models to better align with the specific challenges of hadith literature, incorporating domain-specific annotations for narrators, prophets, crimes, and holy books. This adaptation is critical to overcoming the limitations of general-purpose NER systems when applied to religious texts.

Similarity computations have been extensively employed to identify semantically or contextually related narrations in hadith and Quranic texts. Huang et al.[32] used cosine similarity in combination with embedding-based techniques to link related hadith, particularly paraphrased or overlapping narrations, demonstrating the utility of vector-based approaches in clustering narrations with shared meanings. Their work highlighted the potential of cosine similarity for semantic alignment but also noted challenges in capturing more nuanced relationships, such as contextually related but textually dissimilar narrations. Similarly, Basharat et al. [33] explored similarity computations in the context of Quranic verses, comparing different similarity metrics, including cosine and Jaccard similarity. Their study emphasised the importance of embeddings in semantic similarity, showcasing their effectiveness in clustering narrations by themes or topics. However, they also observed that metrics such as cosine similarity alone may struggle to capture deeper thematic relevance in complex texts. In a more recent study, Alshammeri et al. [34] applied embedding techniques using Siamese networks to detect relationships between Quranic verses and hadith. Their work advanced the use of pre-trained language models to generate embeddings for Arabic religious texts, illustrating the effectiveness of modern NLP tools for capturing semantic overlap. These methods

enabled the automatic clustering of narrations by shared topics, further enriching the interconnectedness within Islamic texts.

Collectively, these studies underscore the utility of similarity computations in Arabic text analysis while highlighting challenges that arise due to the linguistic and contextual complexity of religious texts. Techniques like cosine similarity are foundational in these efforts but often require adaptations or complementary methods to fully capture thematic and contextual relationships within large collections of narrations.

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 provides the overview of this framework. The process begins with data selection and acquisition, ensuring the inclusion of relevant hadith collections. This is followed by a description of our custom knowledge extraction methodology, which involves Natural Language Processing (NLP) techniques for entity recognition and extraction from textual sources. Subsequently, we discuss conceptual knowledge modelling and formalisation, wherein we establish a structured framework to organise and represent the extracted entities systematically. Next, we describe our methodology for similarity computation and the interlinking of hadith narrations. This involves quantifying textual similarities to identify and establish semantic connections, further enriching the structure and utility of the knowledge graph. Finally, we outline the integration of external data sources and the final generation of a linked knowledge graph. Detailed explanations of each stage are provided in Sections 3.1 to 3.7.

This study operates under a set of clearly defined assumptions that guide its methodology and scope. First, the textual representations of the hadith corpus are assumed to be standardised and faithful to their original compilations, minimising concerns about variations across different print editions or translations. Second, the corpus is treated as linguistically uniform, adhering to Classical Arabic for NLP tasks. While this approach facilitates processing, it is acknowledged that certain classical forms and regional dialects inherent to historical narrations may not be fully captured. Third, the authenticity of the selected hadith collections is assumed based on the authority of their compilers; individual evaluations of narrations' authenticity (e.g., weak or fabricated narrations) fall outside the scope of this study. Finally, expert validation serves as the gold standard for resolving ambiguities in named entities and relationships, ensuring the semantic accuracy and reliability of the resulting knowledge graph.

3.1. Data Selection and Acquisition

Developing an enhanced knowledge graph begins with the careful selection and acquisition of relevant data. Building on the *SemanticHadith* ontology and knowledge graph from prior work [15], this study broadens its scope while maintaining continuity. The same six authoritative hadith collections — Sahih Bukhari, Sahih Muslim, Sunan Abi Dawood, Sunan Ibn Majah, Sunan An-Nisai, and Jami At-Tirmidhi — sourced from the Islamic Urdu Books Website[1], were utilised. Collectively known as the Kutub al-Sittah, these collections, comprising 34,458 hadith, are widely regarded as the most authentic compilations in Islamic tradition. Each narration includes the original Arabic text alongside Urdu and English translations. To ensure platform compatibility, all textual data were standardised into Unicode format using Python's `unicodedata` library and the `normalize()` function with the NFKC (Normalisation Form KC) option. This process ensures consistent representation of Arabic script, including diacritical marks, while maintaining data integrity. Validation steps included preservation of Arabic script, detection and rectification of encoding issues, and manual review of random samples to confirm data accuracy. Reproducibility is supported through code provided as supplementary materials. The dataset encompasses all narrations from the selected collections to ensure comprehensive coverage without applying detailed inclusion or

[1] www.IslamicUrduBooks.com

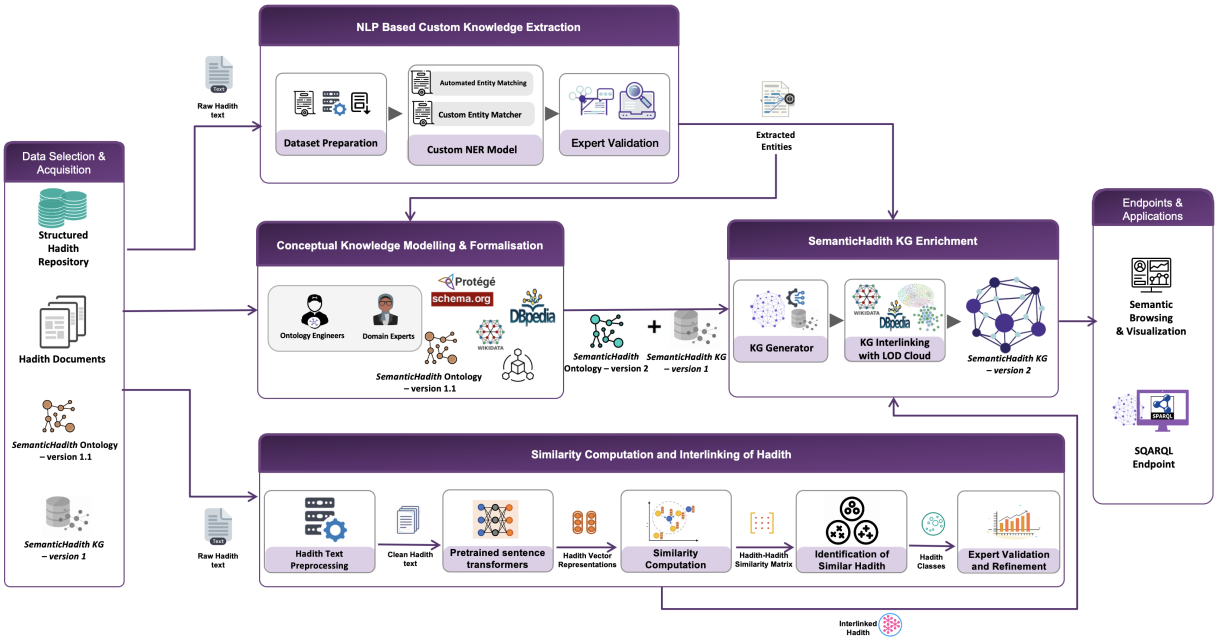


Figure 3. Overview of the *SemanticHadith* knowledge graph construction framework. The key stages of the framework include Data Selection and Acquisition, NLP-based Custom Knowledge Extraction, Conceptual Knowledge Modelling and Formalisation, Similarity Computation and Interlinking of hadith, *SemanticHadith*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

exclusion criteria. This approach aligns with the foundational knowledge graph's scope established in earlier work. Standardisation ensures structural and semantic alignment with the pre-existing knowledge graph, even without filtering narrations based on attributes like completeness or classification. Dataset integrity was ensured through automated encoding checks, manual spot-checking of random samples, and cross-referencing with the pre-existing knowledge graph to validate structural and semantic consistency.

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.

Our approach employs a customised NER pipeline to extract entities from the hadith corpus, enabling the extension of the *SemanticHadith* ontology. The process begins with a modified version of the CANER-CORPUS dataset [31], enriched with domain-specific entities such as prophets, angels, holy books, and crimes, ensuring alignment with the semantic characteristics of hadith texts. The customised dataset includes 14 entity classes relevant to the hadith corpus, with a total of 57,763 labelled entities. Linguistic preprocessing steps, including diacritic removal, orthographic normalisation, and tokenisation, address the structural challenges of Arabic text, ensuring consistent and accurate entity extraction.

The NER model is trained using the spaCy NLP library and pre-trained embeddings from the CAMeLBERT-CA NER model. We fine-tuned it for the hadith domain. The dataset is split in an 80-20 ratio for training and validation. Training stabilises at 10 epochs, achieving high precision, recall, and F1-scores for frequently occurring entity classes such as Persons, Prophets, and Allah, while rare entity classes (e.g., Crimes and Natural Objects) show lower recall due to limited representation in the dataset.

To ensure alignment with the ontology and to standardise entity mappings, noun dictionaries are employed, accounting for variations in Arabic morphology and transliterations. This methodology includes

expert validation to review annotations and ensure that extracted entities are accurate and contextually relevant. Further details of the methodology, including dataset customisation, hyper-parameter optimisation, and model evaluation metrics, are discussed in Section 4.

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 HadithNarrator`. 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 initially explored the use of cosine similarity and pre-trained Arabic sentence transformers to identify similar hadith within Sahih Bukhari and other selected collections, including Sahih Muslim, Ibn Maja, Sunan Abi Dawood, and Nisai. This approach was informed by prior successful applications of semantic similarity computations for Quranic verses [34] and hadith literature [32]. The asafaya BERT base Arabic model, implemented in the Transformer library, was employed to encode Arabic hadith texts into numerical representations, facilitating the computation of cosine similarity scores between pairs of hadith. The process involved preprocessing hadith texts to remove diacritics, punctuation marks, and stop words, ensuring uniformity in representation. Cleaned texts were then encoded into embedding vectors, and a cosine similarity matrix was computed to quantify similarity between all pairs of hadith. Initial thresholds for similarity were derived using a small dataset of expert-identified similar pairs, which served as a benchmark for evaluating the suitability of this approach for larger datasets.

However, significant challenges were encountered in applying this methodology to the complete Sahih Bukhari corpus, including issues with inflated similarity scores due to shared narrator chains (sanad) and difficulty capturing thematic or contextual relevance. As a result, we concluded that the approach was not sufficiently refined to reliably identify similar hadith across the corpus. Consequently, we relied exclusively on the expert-provided similar hadith pairs for interlinking within the *SemanticHadith* knowledge graph. Further details on the similarity computation process, challenges encountered, and expert validation are provided in Section 6.

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 Web Ontology Language (OWL) individuals, incorporating data as data properties, and

establishing semantic relationships based on object properties and mapping rules. To transform entities extracted by our NLP pipeline into a Resource Description Framework (RDF)-based knowledge graph, we utilise the OntoRefine tool [35] along with our *SemanticHadith* ontology [15]. 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 [36] and Wikidata [37], we utilised automated interlinking tools like LIMES [38] and OntoRefine [35]. 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 KGs, including DBPedia,[3] Wikidata[4] and QuranOntology [39] vocabularies, thereby enhancing the interconnectedness of the *SemanticHadith* KG.

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 KG 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 people in our dataset. Overall, these efforts resulted in substantial interlinking, as reported in Table 5 in Sections 7.1, with the establishment of links using `owl:sameAs` for entities in both DBpedia and Wikidata graphs.

3.6. Expert Validation

The dataset preparation, annotation, and ontology development processes were supported by domain experts selected based on specific criteria to ensure accuracy and relevance throughout the evaluation process. They possessed graduate-level qualifications and a foundational understanding of relevant disciplines. Additionally, the experts demonstrated a working knowledge of Islamic studies, particularly regarding the Quran and hadith, to provide domain-specific insights. Proficiency in both Arabic and English was essential to facilitate accurate linguistic evaluation and ensure consistency during text processing. They also had experience with semantic annotation and a strong understanding of entity categorisation, particularly in the context of Islamic knowledge representation.

Three domain experts were engaged to ensure the accuracy and domain relevance of the research, contributing to various aspects of the study. They guided the development of the extended *SemanticHadith* ontology by advising on entity definitions, relationships, and thematic categorisations to capture the complexity of the corpus. The experts annotated key entities such as prophets, angels, crimes, and holy books, refining these annotations to align with semantic and linguistic goals while excluding irrelevant labels. They also created dictionaries for named entities like people and locations, addressing variations in Arabic naming conventions to support entity extraction. To validate extracted entities, the experts cross-referenced identified entities with corresponding hadith passages, reviewing a random sample of 100 instances from each collection. Additionally, they provided insights into challenges encountered during similarity computations, such as distinguishing contextually similar but semantically distinct narrations and addressing paraphrased hadith versions. This collaborative effort ensured the dataset and ontology were accurate, contextually grounded, and semantically robust.

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

3.7. 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/.

4. NLP Methodology for Entity Extraction

The extraction of entities from the six canonical hadith collections is a crucial step in the development of the extended *SemanticHadith* ontology. This section describes our NLP methodology, which combines custom-trained 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 [41]. Hence, we used modified CANERCORPUS [31] 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

To build a domain-specific NER model for extracting meaningful entities from the hadith corpus, we utilised the CANERCORPUS dataset [31] as the foundational resource. This dataset, known for its extensive coverage of Arabic texts and entity annotations, served as a robust starting point for our research. However, to ensure alignment with the unique semantic and structural characteristics of the hadith domain, we applied significant modifications in collaboration with domain experts.

4.1.1. Customisation and Annotation

The original CANERCORPUS dataset was adapted to include additional entities specific to the hadith corpus. These customisations aimed to enhance the alignment of the dataset with the extended *SemanticHadith* ontology, which includes entities relevant to the selected six collections. These modifications include:

- **Addition of Domain-Specific Labels**: Entities such as holy books, angels, significant crimes, and after-life concepts were annotated to expand the scope of domain-specific knowledge. For example, holy books such as *Quran* (القرآن), *Bible* (التوراة), *Injeel* (الإنجيل), and *Zabur* (الزبور) were not previously identified as named entities. Similarly, in the angels category, entities such as *Jibreel* (جبريل) and *Mikail* (ميكائيل) were annotated to reflect their frequent mentions in the hadith corpus.
- **Removal of Irrelevant Labels**: Annotations for categories not pertinent to the hadith corpus—such as monetary values (Money e.g. Dinar / دِينار), and numerical expressions (NUM e.g., Seventy / سبعون)—were removed or replaced to streamline the model's focus.

4.1.2. Arabic Text Preprocessing

The linguistic structure of Arabic posed unique challenges for text processing. Preprocessing steps were applied to normalise text and ensure consistency across the training data:

- **Diacritic Stripping**: Diacritics were removed to standardise tokenisation and reduce data sparsity.
- **Orthographic Normalisation**: Variants of Arabic letters were unified to ensure consistency in tokenisation and entity extraction.
- **Tokenisation**: Text was segmented into tokens using a domain-adapted Arabic tokeniser, accommodating the complex morphology of Arabic script.

Table 1
Statistical overview of the customised dataset

Entity Class	Frequency	Entity Class	Frequency
Person	39201	Paradise	294
Allah	7814	Hell	245
Prophet	6366	Crime	209
Location	1252	Religion	184
Clan	678	HolyBook	130
Natural Object	669	Month	66
Date	602	Angel	53

The customised dataset includes 14 entity classes relevant to the hadith corpus, such as `Prophets, Persons, Holy Books, and Crimes`, with a total of 57,763 labelled entities. Table 1 provides an overview of entity class distribution. The dataset provides a rich, domain-specific resource tailored for the semantic modelling of hadith texts, enabling effective training of the NER model.

4.2. Training the NER Model

We trained a custom NER model using the spaCy NLP library to extract entities from the Arabic hadith corpus. This model was fine-tuned using transfer learning techniques on the customised dataset prepared for this study. The training process included a series of steps to optimise the model's performance and evaluate its effectiveness across domain-specific entities.

4.2.1. Training Setup

The training data was split into an 80-20 ratio for training and validation. The NER model was initialised using pre-trained embeddings from the CAMeLBERT-CA NER model[5]. The model is trained on classical Arabic texts. We fine-tuned it further for our domain-specific tasks using the following parameters during training:

- **Number of Epochs**: 20 epochs were completed, with performance stabilising by epoch **10**. Hence, epoch 10 was selected as the final model state to avoid over-fitting in subsequent epochs.
- **Batch Size**: A moderate batch size of **32** samples was used to optimise training speed while ensuring adequate memory utilisation. This size also allows for stable updates to the model's parameters in each batch.
- **Learning Rate**: A small learning rate of **3e-5** was employed to ensure gradual and precise updates to the model weights during fine-tuning, which is crucial when using pre-trained embeddings.
- **Dropout Rate**: 0.2 was used as a regularisation technique to prevent over-fitting, especially given the imbalanced dataset where certain entity classes (e.g., Persons, Prophets) were over-represented.
- **Loss Function**: Categorical cross-entropy loss was chosen as it is a commonly used loss function for multi-class classification tasks. It minimises the divergence between the predicted and actual probability distributions over entity classes, ensuring effective learning.

4.2.2. Metrics

The model's performance was evaluated using **precision**, **recall**, and **F1-score** for each entity class. To summarise overall performance, three types of averages—**micro**, **macro**, and **weighted**—were calculated:

- **Micro Average**: Combines all true positives, false positives, and false negatives across all classes before computing precision, recall, and F1-score. This provides a global measure of the model's overall performance, heavily influenced by high-frequency classes.
- **Macro Average**: Computes the unweighted mean of precision, recall, and F1-score across all classes, treating each class equally. This highlights performance on rare classes, such as **Crimes**, regardless of their support.

[5] `bert-base-arabic-camelbert-ca-ner`: https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-ca-ner

Table 2
The summary of the precision, recall, and F1-score for each entity class at epoch 10

Entity Class	Precision (%)	Recall (%)	F1-Score (%)
Allah	99.0	98.0	98.0
Angel	100.0	97.0	98.0
Clan	94.0	94.0	94.0
Crime	88.0	41.0	56.0
Date	76.0	74.0	75.0
Holy book	89.0	100.0	94.0
Hell	95.0	91.0	93.0
Location	86.0	88.0	87.0
Month	100.0	100.0	100.0
Natural Object	88.0	44.0	58.0
Paradise	100.0	96.0	98.0
Person	97.0	98.0	98.0
Prophet	99.0	97.0	98.0
Religion	100.0	100.0	100.0
Micro Average	**97.0**	**96.0**	**97.0**
Macro Average	**89.0**	**85.0**	**85.0**
Weighted Avg	**97.0**	**96.0**	**96.0**

– **Weighted Average**: Computes the mean of precision, recall, and F1-score across all classes, weighted by the number of true instances for each class. This provides a balanced view that accounts for class imbalance, making it a reliable metric for the dataset as a whole.

4.2.3. Observations and Results

The model's performance stabilised at epoch 10, with minimal improvements observed in subsequent epochs. Some rare entity classes, such as `Crimes`, suffered from lower precision and recall due to limited representation in the dataset. Expanding the annotated dataset for these classes could improve their performance in future iterations. Table 2 summarises the precision, recall, and F1-score for each entity class at epoch 10. The model achieved an F1-score of 96.6% (micro average) across all entity classes, indicating robust overall performance. The macro average F1-score was slightly lower at 85%, reflecting challenges with rare entity classes such as `Crimes`. High-frequency classes, including `Persons, Prophets, and Allah`, consistently achieved precision and recall above 97%, resulting in F1-scores exceeding 98%. These classes benefited from sufficient representation in the training dataset. Rare entity classes, such as `Time and Crimes`, exhibited lower F1-scores due to limited training data. For instance, the Crime class achieved a precision of 88% but a recall of only 41%, leading to an F1-score of 56%.

The training pipeline, implemented using spaCy and the seqeval library, is fully reproducible. Scripts, annotated datasets, and performance logs are included in the supplementary materials to support further research and model adaptations.

4.3. Entity Extraction Process

The entity extraction process utilised the trained NER model to analyse the hadith corpus across all six collections. For each passage, the model identified spans of text corresponding to entities such as *persons, locations, events*, and thematic topics, categorising them into predefined classes like *angels, prophets, clans, crimes*, and others, based on the extended *SemanticHadith* ontology.

To address variations in entity representation and ensure consistent mapping, *noun dictionaries* were developed in collaboration with domain experts. These dictionaries included:

– Instance ID: A unique identifier for each entity aligned with the ontology,
– English Variations: Common transliterations and spellings in English, and
– Arabic Variations: Variants in Arabic script accounting for morphology and diacritics.

For example, the *Clans* class in the ontology includes an instance for "Quraysh," represented as:

- Instance ID: `Quraysh`
- English Variations: Quraysh, Quraish
- Arabic Variations: قُرَيْش, قُرَيْشٍ, قُرَشِيٌّ, قُرَشِيَّان

When the NER model identified an entity (e.g., *Quraish*), the noun dictionary was used to map it to the corresponding ontology instance (`Quraysh`). If the identified entity did not match any instance in the dictionary, it was discarded. This ensured that only valid entities aligned with the ontology were retained.

For each hadith collection, the extraction process generated a file for each entity class. These files listed the identified instances along with their corresponding *Hadith IDs*. For example, in the *Clans* class within Sahih Bukhari, **837 hadith passages** were identified as containing entities related to clans. A snippet of the identified instances is shown below:

```
Hadith Number    Clan Entities
7                ['Romans', 'Jews', 'Quraysh']
118              ['Muhajirun', 'Ansar']
603              ['Jews', 'Christians']
...
4028             ['Jews', 'BaniNadir', 'BanuQurayzah', 'BanuQainuqa']
...
```

An expert reviewed these results by cross-referencing the identified entities with the corresponding hadith passages. For instance, in hadith number seven from Sahih Bukhari, the expert verified that the entities *Romans, Jews, Quraysh* were correctly identified and appropriately mapped to their ontology instances. This process was repeated for a random sample of 100 instances from each collection. No discrepancies were reported during this validation phase, confirming the reliability of the mapping process.

4.4. Handling Ambiguities and Variations in Person Names

Our NLP methodology effectively extracted specific entities, such as prophets, pious caliphs, and the wives of Prophet Muhammad, by leveraging dictionaries and predefined mappings. However, accurately mapping general person names to corresponding ontology instances remains a significant challenge. In particular, shared names among multiple individuals in Islamic tradition, coupled with the variability in Arabic naming conventions—such as the use of familial relations, titles, and honorifics—complicate systematic linking of names to ontology instances.

For example, the name *Zainab* (زينب) refers to multiple prominent figures in hadith literature:

- **Zainab bint Jahsh** (زينب بنت جحش): A wife of Prophet Muhammad, also known as *Umm al-Masakin*.
- **Zainab bint Muhammad** (زينب بنت محمد): The Prophet's eldest daughter.
- **Zainab bint Abi Salamah** (زينب بنت أبي سلمة): A stepdaughter of the Prophet.

Such diversity confuses automated tools, especially when references lack explicit details, like whether *Zainab* refers to *Zainab bint Muhammad* or another individual. Currently, the system uses predefined dictionaries to map explicit names to ontology instances accurately. However, ambiguous references or uncommon terms, like *Umm al-Masakin*, are flagged for manual review or excluded to avoid errors.

To address this, we suggest a crowdsourcing framework for expert validation as a future enhancement. In this framework, ambiguous names extracted from hadith texts would be presented to domain experts alongside relevant contextual information, such as the full hadith passage and a list of potential ontology matches. For example, when encountering *Zainab* in a hadith, experts could determine whether it refers to *Zainab bint Muhammad* or another individual such as *Zainab bint Jahsh* based on the passage's context and thematic content. This iterative validation process would systematically reconcile ambiguities, expanding the dictionaries and improving the system's ability to handle complex naming conventions. Incorporation of expert validation would significantly enhance the reliability of entity mapping in the *SemanticHadith* ontology, particularly for shared names and cases involving contextual or cultural nuances.

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 3

Competency Questions Mapped to CQ Archetypes/Patterns as identified by [43] (CE = class expression, OPE = object property expression, DP = data type property, I = individual, PE = property expression, NM = numeric modifier, QM = quantity modifier).

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 [42] 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 [15], 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 3. Ren et al.[43] propose a framework for categorising CQs into 12 patterns or archetypes. 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 [15]. 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 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 based on hadith [10, 26].

We draw inspiration from existing vocabularies such as the Semantic Quran vocabulary [44] and Quran ontology [17] 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. 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) [45]: 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 [46]: 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 [36]: 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 [37]: 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 OWL properties `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.
- **HadithNarrator, 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.
- **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.
- **discussesTopic:** Specifies the thematic topics discussed within a hadith narrative, linking the hadith entity to relevant topic entities.
- **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.

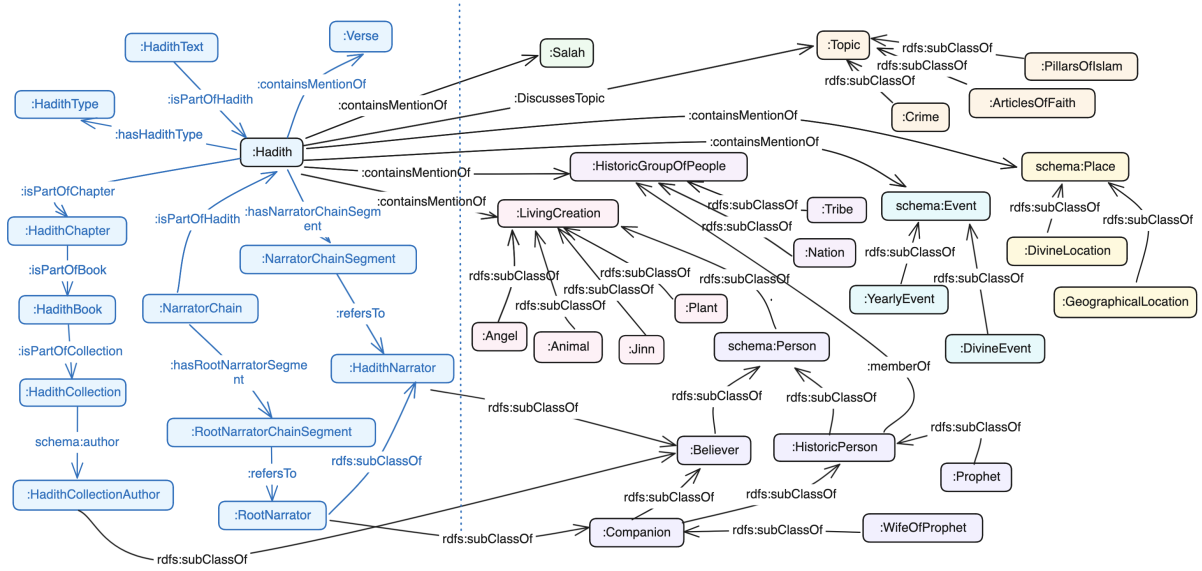


Figure 4. Conceptual model of the *SemanticHadith* ontology version 2.0.1

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`, `HadithNarrator`, 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 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`:** Recognising 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.

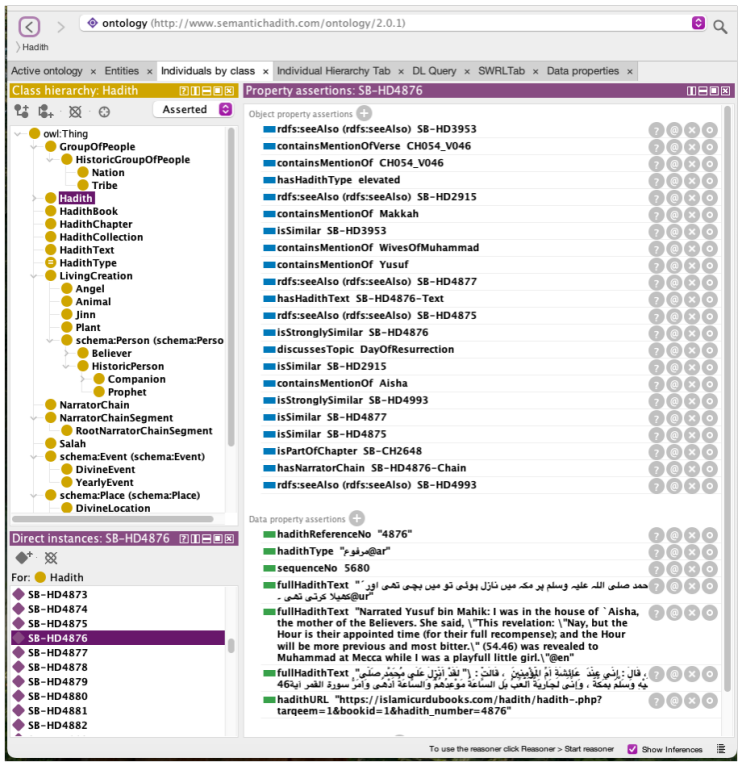


Figure 5. Ontology class :Hadith with property assertions for hadith instance SB-HD4876 in Protégé.

– **`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 [42]. 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. To provide clarity on how entities and relationships are represented, Figure 5 shows the Hadith class in Protégé, detailing its connections to narrators, locations, and thematic topics. This visualization demonstrates the structured nature of the ontology and its ability to capture semantic relationships within hadith texts. As depicted in Figure 6, the hadith text undergoes semantic annotation with ontology classes, facilitating enhanced comprehension and semantic querying capabilities.

5.8. Integration and Challenges in Ontology Implementation

There are multiple ontology editors available. Amongst these widely used ones have been comprehensively compared in [47]. For this research, we selected Protégé version 5.5.0 [48] for its extensibility, UTF-8 support for Arabic data, and compatibility with tools like Jena and XML. The hadith: prefix was adopted for the vocabulary, while established vocabularies such as Schema.org [46] and DublinCore [49] were reused

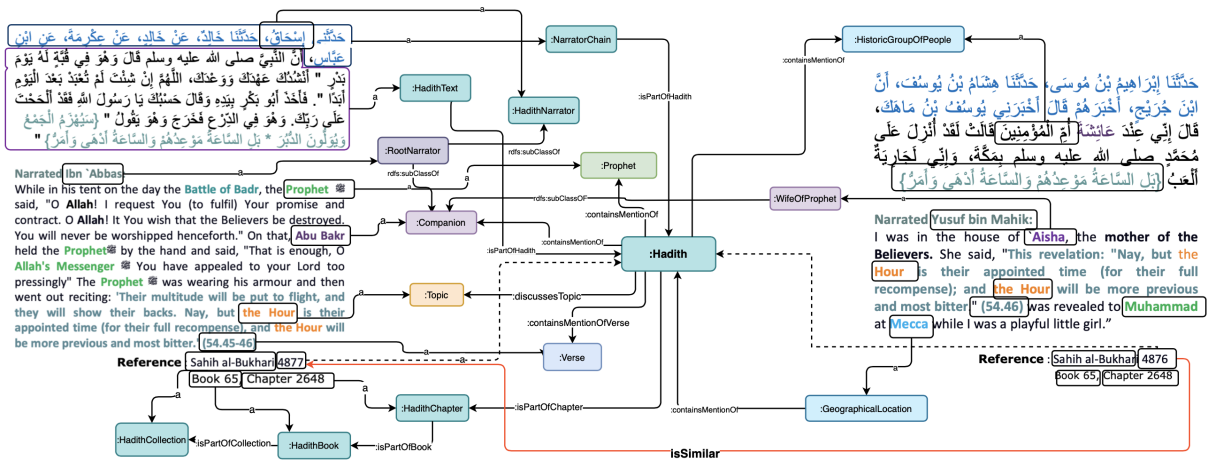


Figure 6. Semantic Annotation of hadith text with *SemanticHadith* Ontology

to ensure interoperability. Equivalence relations with DBpedia [36] and Wikidata [37] enhanced linkage with external datasets. According to the classification by Partridge et al. [50], 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.

Integration challenges included reconciling structural and semantic differences between *SemanticHadith* and external ontologies. Structural alignment issues were resolved by reusing external classes where possible and introducing custom subclasses linked through `owl:equivalentClass` and `rdfs:subClassOf`. Translation challenges, particularly for Arabic concepts with multiple meanings, were addressed through expert validation to ensure accurate semantic alignment. Automated tools like OntoRefine [35] supported the initial mappings, which were refined manually to guarantee contextual accuracy. Limited Arabic-specific support in external tools was mitigated through SPARQL-based validations and standardised RDF/OWL mappings. These efforts resulted in a robust, interoperable ontology that links the hadith corpus to global knowledge graphs, facilitating cross-disciplinary research and enabling seamless exploration of Islamic knowledge.

6. Identification of Similar Hadith

One of the pivotal objectives of our study is to identify similar hadith within the six canonical hadith collections. 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 [33, 34] as well as proposed for hadith [32].

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, an NLP Model implemented in the Transformer library, using the Python programming language [51]. Subsequently, a cosine similarity matrix of dimensions 7563x7563 (where 7563 is the total number of hadith

in Sahih Bukhari) was computed to quantify the similarity between all pairs of hadith. We received an initial dataset with similar hadith pairs identified for Sahih 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

During the computation of cosine similarity scores for the complete Sahih Bukhari corpus, a significant discrepancy was observed between the similarity distributions of the expert-annotated dataset and the full corpus. The similarity matrix, comprising all unique pairs from the 7563×7563 corpus, was divided into bins based on cosine similarity scores ranging from 0.0 to 1.0. Higher bins indicated greater similarity between hadith pairs.

The similarity bins for the expert dataset (shown in Table 4a and Figure 7a) revealed 19,877 hadith pairs in the highest similarity bin (0.9–1.0). However, the same bin in the complete corpus (as shown in Table 4b and Figure 7b) contained 1,042,332 pairs, which is an order of magnitude larger. Similar discrepancies were observed in other bins, with the complete corpus yielding significantly higher counts of pairs with moderate to high similarity scores, compared to the expert dataset.

Table 4
Distribution of Hadith Pairs Across Similarity Bins in Sahih 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

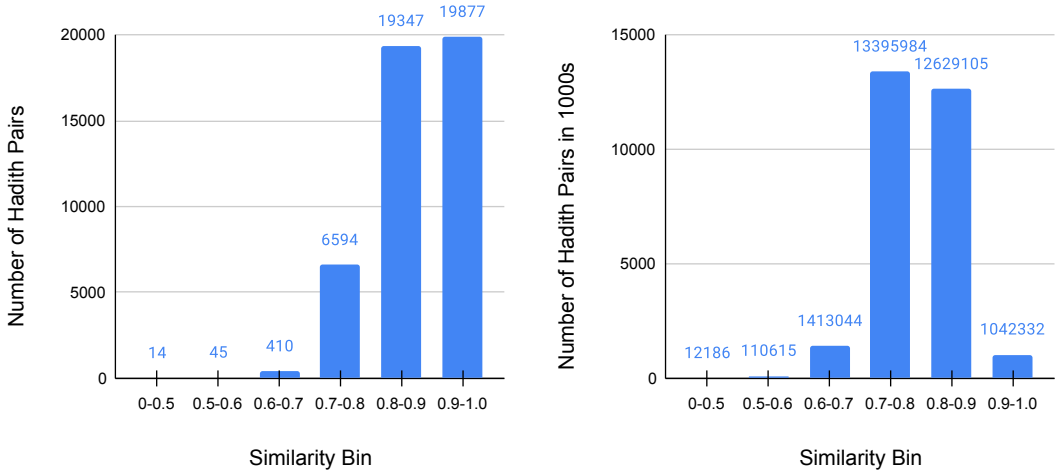
(b) Distribution of all unique Hadith Pairs

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

6.3. Expert Validation and Insights

We engaged domain experts to validate 100 randomly selected hadith pairs from the top similarity bins (0.7–0.8, 0.8–0.9, and 0.9–1.0) to assess the accuracy of the similarity computation and understand discrepancies in the results. The experts used two primary criteria to evaluate the pairs. First, they assessed textual similarity (matan) by determining whether the similarity score reflected genuine overlap in wording, meaning, or structure. A high degree of similarity in matan was required for pairs to be classified as similar. Cases where similarity scores were inflated due to shared narrator chains (sanad) rather than genuine textual overlap, were considered as false positives. Second, experts considered contextual and thematic relevance, going beyond textual similarity to identify hadith pairs addressing related topics or events. Even when the matan differed significantly, pairs were classified as relevant if they shared meaningful thematic or historical connections.

The experts helped us identify reasons for the discrepancy in actual and expected results with concrete examples. For example, pairs like Sahih Bukhari 3398 and Sahih Bukhari 3415 involved subset relationships, where one hadith encompassed a shorter segment of another. In such cases, the shared matan warranted classification as similar despite differences in length or additional context in the longer narration. This difference in textual length resulted in a lower similarity score. Another issue arose with pairs



(a) Expert-Shared Hadith Pairs Distribution. (b) Distribution Across Similarity Bins.

Figure 7. Distribution of Hadith Pairs Across Similarity Bins in Sahih Bukhari for expert-shared pairs and complete corpus.

like Sahih Bukhari 52 and Sahih Bukhari 2051, which shared overlapping matan but scored moderately (0.67656) due to differences in sanad. Additionally, thematic connections were observed in pairs like Sahih Bukhari 69 and Sahih Bukhari 4341, where little to no overlap in matan existed hence scoring low in similarity score, yet the hadith were deemed similar by the experts because they addressed related topics within Islamic tradition.

The validation process highlighted significant challenges with the automated similarity computation. Many high-scoring pairs were identified as false positives, primarily due to inflated scores caused by overlapping sanad rather than genuine matan similarity. This discrepancy highlights a limitation in the current methodology, and future work will aim to address this issue by focusing solely on the textual content (matan) of hadith to mitigate the impact of overlapping narrator chains (sanad). Additionally, several valid pairs fell outside the top similarity bins, demonstrating that cosine similarity metrics alone were insufficient to capture the multifaceted nature of hadith similarity. For example, thematic connections were observed in pairs where the matan showed little to no overlap, further emphasising the limitations of a purely automated approach.

Furthermore, the sheer scale of the similarity matrix, encompassing over 2 million unique hadith pairs in Sahih Bukhari alone, made manual validation impractical. Given these challenges, we decided to rely solely on the expert-annotated dataset for this phase of the study. The expert dataset, which includes carefully validated hadith pairs, was integrated into the *SemanticHadith* knowledge graph to ensure the accuracy and reliability of the relationships represented.

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 other 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. To mitigate inflated results, future approaches could isolate matan by preprocessing the text to exclude sanad before embedding. Alternative distance measures, such as Euclidean distance (measuring absolute spatial distance) or Manhattan distance (sum of absolute differences in dimensions), could complement cosine similarity by capturing variations in embedding magnitude. These measures could provide additional insights into contextual relationships that cosine similarity alone may overlook. Using large language models (LLMs) fine-tuned for Islamic texts can provide embeddings that better encode semantic and contextual nuances. Furthermore, integrating hybrid metrics, such as a weighted combination of cosine similarity and contextual relevance derived from LLMs, could better capture the multifaceted nature of hadith relationships. Supervised learning methods trained on expert-labelled data could further refine the thresholds for similarity classification. Additionally, as previously outlined, future work will involve developing a crowdsourcing framework for expert consultation, where hadith pairs meeting a predefined similarity threshold will be reviewed by domain experts. To address potential disagreements or conflicts, mechanisms such as majority voting or weighted input from experienced experts could be explored.

7. Results and Discussion

This section evaluates the design and implementation of the *SemanticHadith* ontology and knowledge graph, focusing on its logical consistency, scalability, practical applications, and future potential. We also present the metrics of both the *SemanticHadith* ontology and knowledge graph, the formal ontology design requirements, and answers to competency questions in addition to the intended applications for this endeavour. A thorough analysis of the ontology's structure, evaluation outcomes, and scalability adaptations highlights the project's achievements and identifies avenues for further research.

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 [48].
- **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. Figure 5 presents the ontology class `:Hadith` in Protégé, with detailed property assertions for an instance `SB-HD4876` referring to (Sahih Bukhari, 4876). This demonstrates the relationships captured by the ontology, including links to narrators, topics, and referenced Quranic verses, as well as semantic similarity connections to other hadith.

Table 5
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

– **Pitfall Detection and Resolution:** The ontology underwent evaluation using the OntOlogy Pitfall Scanner! (OOPS!) online service [52] 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.
– **MIRO Evaluation:** The Minimum Information for the Reporting of an Ontology (MIRO) guidelines [53] were applied to assess the completeness and reporting standards of the *SemanticHadith* ontology. A detailed MIRO report is available in the GitHub repository[6].
– **Knowledge Graph Correctness:** The correctness of the *SemanticHadith* knowledge graph was evaluated by answering a set of competency questions. Competency questions, as outlined in Section 5.2, were successfully addressed through SPARQL queries, demonstrating the ontology's robustness and usability in answering semantic queries. SPARQL queries for these questions, along with their results, are provided in the GitHub repository[7].
– **Knowledge Graph Generation and Summary:** The generated knowledge graph, encompassing six canonical hadith collections, achieved significant milestones, including the annotation of over 34,000 hadith, integration of nearly 7,000 narrators, and establishment of 47,000 internal semantic links (`:isSimilar`) between narrations. External links to resources like Wikidata, DBpedia, and QuranOntology further enriched the graph, providing broader context and enabling cross-domain integration. The statistics in Table 5 and the visual representation in Figure 8 illustrate the ontology's structural and semantic achievements. Figure 9 visualises the relationships between hadith `SB-HD4877 and SB-HD4876` - referring to (Sahih Bukhari, 4876 & 4877) - using the *SemanticHadith* knowledge graph. The `:isSimilar`. property highlights semantic similarity between the two narrations, while edges such as `:containsMentionOf` and `:discussesTopic` link the hadith to shared entities, people, and topics. This visualisation highlights the interconnected nature of the knowledge graph, facilitating advanced exploration and discovery of related hadith.

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

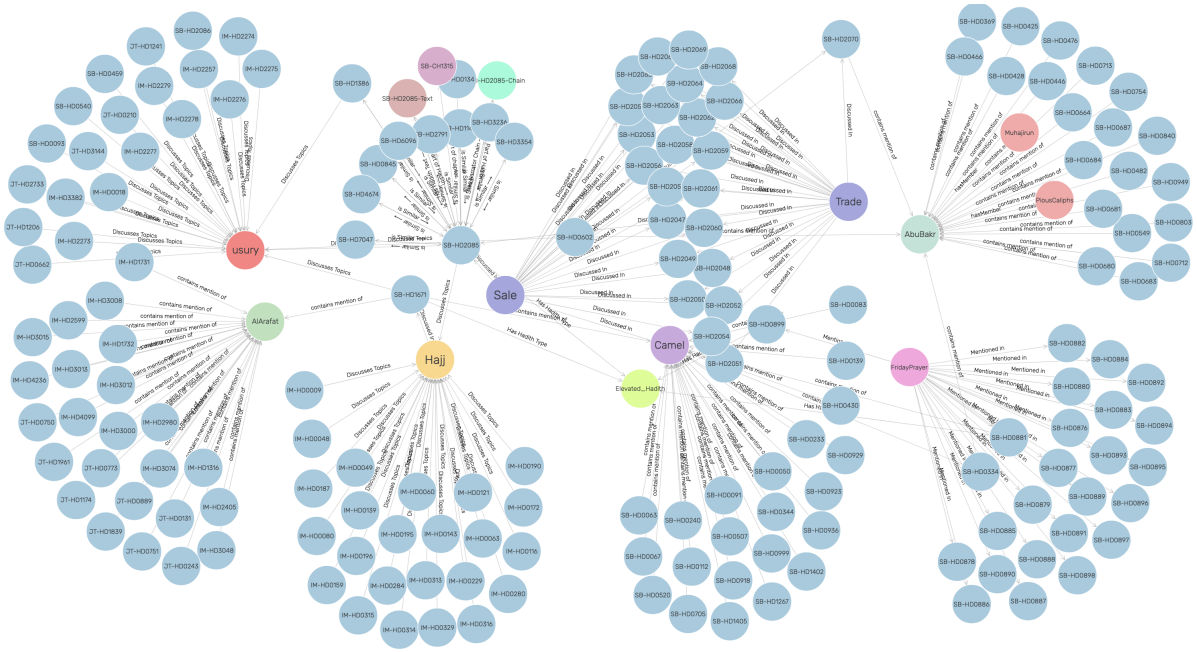


Figure 8. Visualisation from the *SemanticHadith* Knowledge Graph.

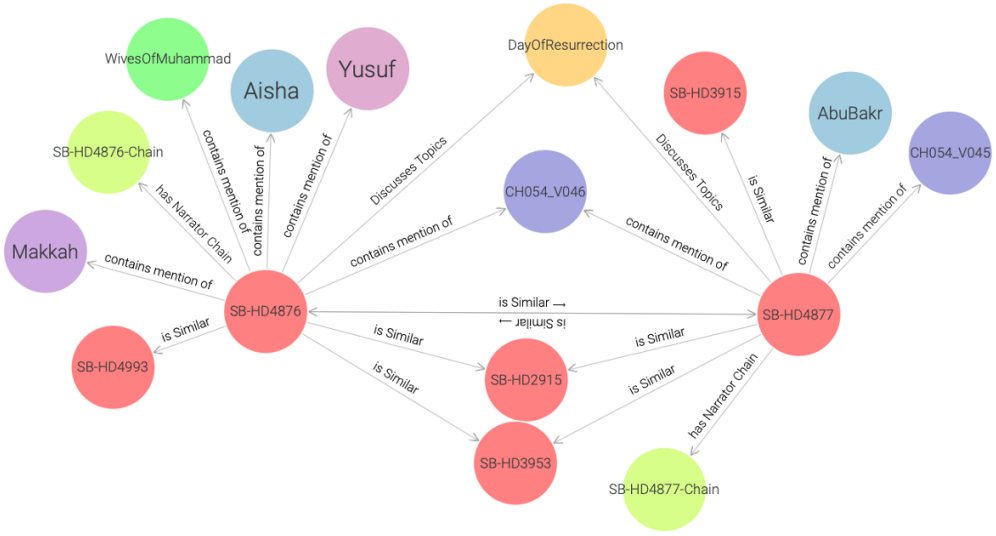


Figure 9. Knowledge graph visualisation showing relationships between hadith (`SB-HD4877` and `SB-HD4876`) and their linked entities, topics, and Quranic references.

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 NLP techniques to annotate hadith texts. While the current focus has been on specific hadith collections, the methodology and infrastructure established can be extended to 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.

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 knowledge graph 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 KG 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. For example, Figure 10, a GraphDB screenshot, illustrates a query and its results, which capture shared entities (`Verse 54:46`) and topics (`DayOfResurrection`) linked to these hadith. The structured representation within the knowledge graph also offers potential for automated reasoning, enabling systematic exploration of Islamic rules and principles by analysing the semantic relationships embedded in the corpus.

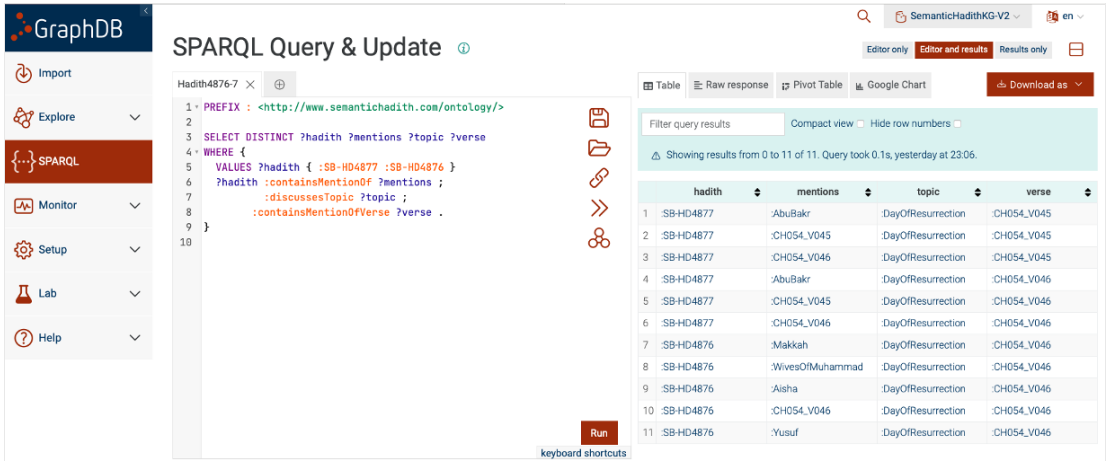


Figure 10. SPARQL query execution and results showing relationships between hadith, entities, and topics in GraphDB.

7.2.3. Cross-Domain Integration

Beyond the realm of Islamic studies, the *SemanticHadith* KG holds potential for cross-domain integration with other knowledge domains. The KG 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 *SemanticHadith* KG 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

KG 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 KG through collaborative efforts, thereby fostering a culture of knowledge sharing and co-creation within the academic community.

7.3. Addressing Scalability Challenges and Existing Limitations

Expanding the *SemanticHadith* knowledge graph to encompass a wide range of Islamic resources, including the extensive hadith corpus and related literature, presents several scalability challenges. These challenges include managing large volumes of data, handling the linguistic and thematic diversity of Islamic texts, and ensuring computational efficiency. Effectively addressing these issues requires managing heterogeneous datasets, maintaining coherence across domains (e.g., tafsir, fiqh, and historical biographies), and verifying the authenticity and contextual relevance of entities and relationships.

One of the primary limitations is linguistic diversity. Historical and regional variations in classical Arabic can result in processing inaccuracies, even with customised NLP tools. Similarly, implicit meanings, contextual nuances, and cultural idioms may not be fully captured, limiting the granularity of the ontology. The inclusion of additional hadith collections or related Islamic texts can further complicate the ontological structure and computational performance. Despite employing advanced NLP techniques, such as transfer learning and pre-trained models, challenges like entity misclassification and relationship extraction errors persist in linguistically complex passages.

To address these challenges, a modular approach to ontology expansion has been adopted, enabling dynamic integration of additional domains while minimising structural inconsistencies. Validation processes have been enhanced with semi-automated workflows, leveraging advanced machine learning models to assist with initial verification tasks and allowing experts to focus on complex cases. Optimised similarity computations, incorporating clustering techniques and pre-filtering strategies, reduce the computational burden associated with large-scale similarity analysis. For NLP tasks, pre-trained models fine-tuned on domain-specific data (e.g., hadith commentaries and classical Arabic prose) enhance the adaptability of the pipeline, ensuring consistent performance across diverse text genres.

Scalability also necessitates infrastructure improvements. Transitioning to distributed storage systems and adopting scalable architectures, such as cloud-based infrastructures, would support the processing and storage demands of the expanding knowledge graph. Federated SPARQL queries could further enhance interoperability by enabling seamless integration with external linked datasets, broadening the scope of interdisciplinary applications.

Moving forward, the *SemanticHadith* project aims to refine the KG-generation pipeline and expand the scope of annotated data by incorporating state-of-the-art NLP techniques and machine learning algorithms. These advancements will improve annotation accuracy and efficiency, enabling the inclusion of diverse hadith collections and genres. Additional efforts will focus on enriching the ontology with metadata, including provenance information, temporal data, and linguistic annotations, to provide a more comprehensive and contextually rich representation of hadith texts.

To address variations in naming conventions within hadith passages and improve entity mapping accuracy, a crowdsourcing framework for expert validation is under development. This framework will allow domain specialists to reconcile extracted named entities with predefined ontology instances, resolving ambiguities and ensuring reliable mapping based on contextual knowledge and expertise. Collaboration with scholars and domain experts will guide the ongoing evolution of the *SemanticHadith* ontology, ensuring its relevance and usability in research, education, and broader interdisciplinary contexts.

By combining methodological adaptations, expert input, and technological advancements, the *SemanticHadith* framework is positioned to meet the challenges of scalability and adaptability while expanding its utility for the comprehensive exploration of Islamic knowledge.

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 of the KG. 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] S. Dietze, S. Sanchez-Alonso, H. Ebner, H.Q. Yu, D. Giordano, I. Marenzi and B.P. Nunes, Interlinking educational resources and the web of data: A survey of challenges and approaches, *Program* (2013).
[5] J. Marden, C. Li-Madeo, N. Whysel and J. Edelstein, Linked open data for cultural heritage: evolution of an information technology, in: *Proceedings of the 31st ACM international conference on Design of communication*, 2013, pp. 107–112.
[6] A. Basharat, B. Abro, I.B. Arpinar and K. Rasheed, Semantic Hadith: Leveraging Linked Data Opportunities for Islamic Knowledge., in: *LDOW@ WWW*, 2016.
[7] 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.
[8] S. Hasan, *An introduction to the science of Hadith*, Al-Quran Society London, 1994.
[9] 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.
[10] 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.
[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] F. Harrag, Text mining approach for knowledge extraction in Sahîh Al-Bukhari, *Computers in Human Behavior* **30** (2014), 558–566.
[13] 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.
[14] 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.
[15] 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.
[16] A.A.B. Philips, *Usool At-Tafseer: the Methodology of Qur'anic Interpretation*, AS Noordeen, 2002.
[17] 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.
[18] 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.
[19] K. Dukes, E. Atwell and A.-B.M. Sharaf, Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank., in: *LREC*, 2010.
[20] 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.
[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] 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.
[23] S. Altammami, E. Atwell and A. Alsalka, Constructing a Bilingual Hadith Corpus Using a Segmentation Tool, in: *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 3390–3398. ISBN 979-10-95546-34-4.
[24] S. Altammami, E. Atwell and A. Alsalka, Text segmentation using n-grams to annotate Hadith corpus, in: *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, 2019, pp. 31–39.
[25] 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.
[26] 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).
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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.

[32] 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.
[33] 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.
[34] 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.
[35] Ontotext AD, OntoRefine (Version 1.2), 2022. https://ontotext.com/products/graphdb/graphdb-free/.
[36] C. Becker and C. Bizer, DBpedia mobile-a location-aware semantic web client, *Proceedings of the Semantic Web Challenge* (2008), 13–16.
[37] 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.
[38] 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.
[39] A. Hakkoum, The Quran Ontology 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] 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.
[42] 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.
[43] 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.
[44] M.A. Sherif and A.-C. Ngonga Ngomo, Semantic Quran, *Semantic Web* **6**(4) (2015), 339–345.
[45] D.U. Board, DCMI Metadata Terms.
[46] 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.
[47] E. Alatrish, Comparison Some of Ontology, *Journal of Management Information Systems* **8**(2) (2013), 018–024.
[48] S. University, PROTÉGÉ.
[49] D.C.M. Initiative et al., Dublin core metadata element set, version 1.1, Dublin Core Metadata Initiative, 2012, [Online; accessed 20. Aug. 2022].
[50] 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).
[51] 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.
[52] 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.
[53] 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.