Moheet Corpus: A Methodology for Building a Citation-Grade Semantic Corpus of Arabic and Persian Islamic Scholarship

Shady Ali

Sadiqoon Technologies Ltd, London

Registered in England and Wales, no. 17120720 · www.sadiqoon.uk · info@sadiqoon.uk

November 2026

Methodology paper — version 1.0

Moheet Corpus is a service of Sadiqoon Technologies Ltd · moheet-ablibrary.net

Abstract

This paper presents the methodology behind Moheet Corpus, a citation-grade semantic corpus that ingests the Ahl al-Bayt Library in both of its languages, Arabic and Persian, and rests on a single unit of construction: the *cited passage*, a bounded span of text that carries its full bibliographic identity with it (book, author and year of death, volume, page, heading chain, and text type). Books pass through a structure-preserving ingestion that keeps the editors' footnotes in place, are then segmented on semantic boundaries, represented by multilingual 1536-dimensional embeddings alongside a sparse lexical representation, and retrieved by rank fusion followed by reranking. Above the vector space sits a knowledge graph in which no relation between two concepts, persons or books is recorded without its literal textual evidence and page location.

As of November 2026 the corpus holds about 13.1 million cited passages from roughly seven thousand works, out of the eighteen thousand in the library. Three properties are measured: attribution to the printed page (99.99% on a measured sample of ten books), prevention of duplicate books through file fingerprints and library identifiers, and ingestion cost (about three US cents per book for embedding, or one dollar per ten thousand pages). The paper closes with the limits of the method and a commitment to open release: the bibliographic graph of the library under an open licence, and a public benchmark for entity and relation extraction from classical Arabic scholarly prose.

Keywords: digital Islamic heritage, hybrid retrieval, knowledge graph, bibliographic provenance, Model Context Protocol (MCP), Arabic and Persian.

1. Introduction: the problem and the question of this paper

The digitisation of Islamic heritage has passed through two generations: preservation by imaging, then the indexed text with lexical search that the large digital libraries represent today. In the second generation a researcher finds a text when they know its wording and misses it when they know only its meaning; and the relations that bind a question in a jurisprudential work to its root in a hadith collection and its critique in a later commentary are visible only through a lifetime of human effort. Large language models have entered this field with the promise of understanding and the danger of fabrication: texts and attributions that do not exist, stated with full linguistic confidence [1], a danger that cannot be tolerated in sciences where the text itself is the object of devotion.

The question of this paper is purely methodological: how does one build a semantic knowledge structure over an encyclopaedic corpus of many genres and two languages such that it becomes structurally impossible for a text to reach the researcher without attribution, or a relation without evidence? The answer presented here is not an improvement to a model but a design constraint that runs through every layer: provenance is part of the building material, not a layer added at display time.

The paper stands between two other works by the author: the introductory paper "Moheet Corpus: A Shift in the Architecture of Shi'i Islamic Knowledge" (September 2026) [2], which treated the project's place on the path of ijtihad and its effect on scholarly study, and the open extraction benchmark for which this paper prepares the ground. It is therefore confined to method: what was built, how, under which constraints, and what was measured.

It draws on the practice of the academic projects for Arabic heritage digitisation, in particular OpenITI and the KITAB project in author-name normalisation and edition disambiguation [3], and on the peer-reviewed hadith knowledge-graph projects [4][5]; it differs from them in two respects: the breadth of the corpus across fourteen genres in two languages, and the literal-evidence constraint in the graph.

2. The corpus and the unit of construction

The ingested corpus is that of the Ahl al-Bayt Library programme in its critically edited digital editions: about eighteen thousand Arabic and Persian works, some twenty-four gigabytes of text, distributed over fourteen genres: hadith, jurisprudence, legal theory, Qur'anic exegesis, creed and theology, biographical evaluation of narrators, ethics, history, supplication, lexica, scholarly commentaries, philosophy and logic, language, grammar and rhetoric, and literature. Coverage is complete rather than selective, because the value of a semantic structure comes from relations across genres, not within a single genre. The bilingualism has methodological weight: a substantial share of lesson transcripts and commentaries of recent centuries was written in Persian and has remained largely cut off from the Arabic-reading researcher, and semantic representation places both traditions in one space that can be queried in either language.

Books enter from the library's exported files with their original structural markup: printed page breaks, footnote blocks, tables of contents, and the book record

(author and year of death, editor, edition, publisher, volume). The record is read automatically from the file itself, so provenance is never entered by hand and mistyped. A complete inventory of 14,690 files established that the page marker in the library's export matches the printed edition in every file, which makes attribution to the printed page possible from the source rather than by estimation.

The unit of construction is neither the book nor the page but the *cited passage*: a bounded span of text (median length about 350 tokens) that carries the following fields in its metadata and never parts from them in any later pipeline:

Field	What it carries	Source
Book, author, year of death	Bibliographic identity	The book record in the file
Volume and page (start and end) and page-break offsets inside the passage	Attribution to the printed edition	Page markup
Heading chain (book → chapter → section)	Position in the book's structure	Table of contents and injected headings
Hadith number and chapter number where present	The customary internal reference of hadith works	Line beginnings, structurally
Text type	Body, body with footnotes, commentary, editor's note	Footnote and commentary markup
Library file identifier and SHA-256 fingerprint	Edition identity and life cycle	The file itself

Under this design it is structurally impossible for an unattributed text to appear in the system: provenance is a property of the passage from the moment it is created, and everything built on top of it — a vector, a relation in the graph, a tool's answer — inherits the same attribution.

3. Structure-preserving ingestion: from library file to passages that carry their provenance

Ingestion proceeds in four successive stages governed by one constraint: no operation may discard structural information that was present in the original file.

1. **Structural parsing of the file.** The book record is extracted, the full hierarchy is built (book → chapters → sections), headings present in the table of contents but absent from the body are injected at their positions, and export markup that carries no meaning for a reader is removed (link, appendix and raw verse tags). Edition identity is fixed by the library's file identifier rather than by title and

volume alone, because titles recur in the tradition (the Amali of al-Mufid, of al-Saduq and of al-Tusi, for example).

2. **Merging footnotes at their positions.** Every numbered footnote is merged into the body at the exact point of its reference, tagged ‹footnote: …›, so that a narrator's biography stays attached to its place in the chain of transmission wherever the segmentation falls; the editor's asterisked notes are distinguished as ‹editor's note: …› and commentaries as ‹commentary: …›. The aim is that the researcher receives the fabric of the critical edition as printed, not a body text stripped of its editors' work.

3. **Semantic segmentation.** The text is divided into passages on meaning boundaries (completion of paragraph and sentence) with a measured overlap between neighbouring passages so that an idea is not cut at the boundary. Because a page break may fall inside a passage, each passage carries its start and end pages and the offsets of the breaks within it, so a retrieved sentence is attributed to its own page rather than to the passage's page in general ("pp. 62–63" where it spans two). Segmentation does not drop short lines, since these hold hadith numbers, chapter headings and page markers.

4. **Verification and storage.** A SHA-256 fingerprint is computed for every file and governs its life cycle: a duplicate copy is rejected, a newer edition of a book replaces the older one atomically, a critical text edition takes precedence over a scanned one, and no original is deleted from the archive but is flagged and kept as an audit trail. The same book is never ingested twice even in a different edition, and where an Arabic original exists its Persian counterpart is not loaded. Page images of scanned books are kept in an object store, and the vector index stores only the vector and the reference.

The four stages run in parallel lanes that claim books atomically from a single queue, and for every book the ledger records what happened to it: accepted, skipped or rejected, why, and at what cost.

4. Vector representation, hybrid retrieval and reranking

Each cited passage is represented twice, and both representations are stored together in a Qdrant vector database [6]:

• **Dense representation (meaning):** a 1536-dimensional vector produced by a commercial multilingual, multimodal embedding model (Cohere Embed v4) in "search document" mode for passages and "query" mode for questions, under cosine similarity. What is close in meaning is adjacent in this space even without a shared word, as the dense-retrieval literature has shown [7][8]; a question

phrased in today's language reaches a text written a thousand years ago, and an Arabic question reaches a Persian text.

• **Sparse representation (wording):** a sparse BM25 vector that captures the exact seminary term and the phrase as written, which is what is needed in source-tracing and in collating narrations.

At query time both searches run together, the two lists are merged by reciprocal rank fusion (RRF), which relies on a result's position in each list rather than its score and so avoids the mismatch of scales between the two engines [9], and the candidates then pass through a reranking model that compares each candidate with the question directly and more precisely than the first pass. Behind the chain runs the HNSW approximate-neighbour algorithm [10], which makes search across 13 million passages a matter of fractions of a second.

Three methodological decisions deserve to be stated:

1. **The embedding model is fixed for the whole ingestion.** Vectors produced by different models cannot be mixed in one space, so any change of model or dimensionality means re-embedding the whole corpus. The model is therefore fixed from the start of the project to its completion, and the vectors are treated in the backup policy as primary assets rather than cheap derived data.
2. **Bibliographic filtering is part of the query.** Because every passage carries its author's year of death, its genre and its book, the researcher can confine a search to authors who died before or after a given Hijri year, or to one genre, which opens the door to studies of conceptual development across the ages that once required the researcher to build a library ordered by generations.
3. **The retrieved text is the printed text.** The presentation layer neither rewrites nor summarises the passage; it shows the faithful text with its footnotes, hadith number and chapter, attributed to the printed page and to the order of the scanned copy where the two differ.

5. The knowledge graph bound by literal evidence

Above the vector space sits a knowledge graph in a Neo4j database whose nodes are concepts, persons, books, narrators and events, and whose edges are typed relations from a closed, documented relation schema (relies on, opposes, critiques, derives from, comments on, narrates from, and their like), with schema layers specific to each genre (relations in biographical works differ from those in philosophy). Knowledge graphs have become the standard structure for representing interconnected knowledge [11], and they entered Islamic studies

through hadith graphs confined to a particular collection [4][5]; Moheet extends the same logic across an encyclopaedic, bilingual corpus.

The governing design constraint: no relation is recorded in the graph without its literal evidence from the text and its location by book and page. The graph does not "hold opinions"; it holds interrelated texts, and when asked about the relation between concept A and concept B it answers in the attributed words of the authors themselves.

Extraction method. Entities and relations are extracted from each passage by an open-weight language model running in weekly batches on a cloud inference API, under a prompt that restricts it to relations explicitly stated in the passage and requires it to return the evidential span verbatim. The span is then verified against the source passage, and whatever fails to match does not enter the graph. Synonymous entities (one person under his names, agnomens and titles) are unified by a later merge step that keeps all alternative names. Growth of the graph is monitored by one measure that cannot deceive: the number of new verified mentions in the database, not the number of passages processed.

Quality is measured before any change. The project maintains a gold set of 150 passages distributed across the genres, annotated with their entities and relations, and an A/B harness that measures, for any new model or prompt, three things: entity recall, relation recall, and fidelity of the literal evidence to the text. Neither the extraction model nor its prompt is replaced in production until it equals or exceeds the incumbent on this set. The set and the harness are released as an open benchmark (section 8).

Thematic communities. Above the graph, a community-detection algorithm of the Leiden family [12] crystallises islands of tightly connected concepts, and a thematic summary is generated for each community, in the manner of graph-augmented retrieval [13], for the holistic questions that cannot be answered by retrieving scattered passages.

State of the graph in November 2026: extraction is complete for about 5% of passages and proceeds at a pace governed by inference cost; this is where the forthcoming work lies (section 8).

6. Access: the research connector, the tools and the architecture of honesty

The corpus is not offered as a closed website but as a software connector on the Model Context Protocol (MCP) [14], the open standard for connecting language

models to knowledge sources, so that any research assistant or AI agent can address its tools and compose them in a single workflow. The connector exposes eleven read tools covering the scholarly work cycle:

Tool	**Function**	**Layer it relies on**
Corpus search	General semantic search with control parameters (genre, author, year of death, book)	Hybrid retrieval
Graph exploration	Verified relations between concepts, persons and books, with the literal evidence	Graph
Thematic communities	A panoramic survey of the major themes connected to a field	Communities
Book card	Bibliographic data, edition and editor	Book register
Narration source-tracing	Occurrences of a narration, ordered by the author's year of death on request, strictly within the strong-match range	Hybrid retrieval
Hadith text verification	An explicit verdict: present / a similar text occurs / not present in this wording, with a match score and a guard for numeral discrepancies	Hybrid retrieval
Qur'anic exegesis and scholarly explanation	Queries directed at the exegesis and commentary genres	Hybrid retrieval
Sequential reading and page view	Reading a page range (three by default) and the original page image through a temporary signed link	Store and vision

Three features reflect the craft's conventions rather than mere programming skill: chronological restriction by authors' deaths; preference for the earliest sources in source-tracing, with a guard that prevents a weakly matching source from advancing merely because it is older; and an explicit verdict in verification instead of a display of look-alikes.

The architecture of honesty. The corpus's answer to the phenomenon of fabrication [1] is not an improvement to the model but an architectural restriction of it, in the manner of retrieval-augmented generation [15], which it then goes beyond by placing the output of every tool under explicit match thresholds: the

system returns "no relevant results" or "no verified relations" whenever the match falls short of the threshold, and treats the empty answer as an honest result rather than a failure. The machine's role in all of this is to alert, not to adjudicate: it is a factory of indications that presents texts with their attribution, and the weighing belongs to the researcher.

The connector runs under operational controls: authenticated access by account, usage quotas measured per call and call type, an access log, and no bulk export of the corpus.

7. Measured evaluation

The figures below are measured from the live system, not estimated, and the date of measurement is given for each.

Property	**Measured value**	**Sample and date**
Corpus size	13,127,394 cited passages, each with a 1536-dimensional dense vector and a sparse vector	Live collection, 1 November 2026
Works ingested	About 7,000 of about 18,000	Book ledger, November 2026
Attribution to the printed page	99.99% of passages attributed to their correct page	10 books collated against the library's site, September 2026
Page marker matching the printed edition in library files	100% (14,690 of 14,690 files)	Full inventory, September 2026
Structural recovery of hadith and chapter numbers	98.9–100%	Selected hadith works, September 2026
Passage length	Median 356 tokens	3,676 books, September 2026
Embedding cost per book	Median USD 0.031 (mean 0.033, 90th percentile 0.060) ≈ one dollar per ten thousand pages	3,676 books, September 2026
Graph-extraction cost per passage	USD 0.00035–0.00045 (about 1,700 input and 1,100 output tokens)	5,947 books processed, November 2026
Hybrid query time	Fractions of a second before reranking	Live collection

Three methodological remarks deserve to be stated. First, attribution accuracy is measured by collating the retrieved passage with its page on the library's own website, and it is recalibrated weekly on a fresh sample with an automatic alert on drift, in the manner of sampling inspection (ISO 2859-1) rather than a single

check at ingestion. Second, the only paid cost in ingesting a book is embedding, while graph extraction is by far the largest item (about 95% of the cost of completing the library), which governs the ordering of priorities in the next section. Third, the corpus has served researchers and institutions in production since October 2026, and usage reports (calls by type and reranking units) are measured daily and form the material of a later study of patterns of research questioning in the heritage.

8. Limits, forthcoming work and open release

Limits of the method. The corpus retrieves what has been ingested, so an absent book is absent from the answer even if it is the origin of the question; the scope of coverage is therefore published on the site by genre and count. Vector representation approximates meaning and does not grasp the argumentative context of a phrase: it may place a text quoted for refutation beside a text quoted in support, and here lies the value of the graph, which distinguishes "relies on" from "opposes" and is still in its beginnings. Scanned books are read automatically on demand and remain less accurate than critically edited text. The language model that drives the tools on the user's side lies outside the corpus's control, so the corpus's guarantee stops at what its tools return: a faithful text with correct attribution.

Forthcoming work. (1) Completing ingestion of the library: about 11,000 works adding some twenty million passages, with the same embedding model. (2) Completing the graph for the whole corpus (about 32 million passages), which calls for moving extraction to cheaper or sovereign compute infrastructure under one non-negotiable condition: the alternative model must equal or exceed the incumbent on the gold set. (3) Relation-schema layers for the genres not yet completed (biographical evaluation and indices). (4) Matching book and author names at citation time through a register of aliases and editions built from the library's data, because names in the tradition are a highly fluid environment that textual matching cannot pin down.

Open release. The company commits to releasing three things under open licences on Hugging Face and Zenodo:

1. **The bibliographic graph of the Ahl al-Bayt Library:** persons, books, narrators, concepts and the relations explicitly stated between them, with evidence pointers (book, volume, page) and without the texts of the books.

2. **An open benchmark for entity and relation extraction from classical Arabic scholarly prose:** the gold set (150+ annotated passages across the genres), the relation schema, the measurement harness and the results of open models on it.
3. **This paper**, with a DOI, and its English translation.

The citation form for the corpus in scholarly work: Sadiqoon Technologies, "Moheet Corpus", version and date of query, moheet-ablibrary.net — together with attribution to the printed book itself in every retrieved text, because the corpus is a means of access, not a source.

References

1. Ji, Ziwei, et al. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys*, vol. 55, no. 12, 2023, pp. 1–38.
2. Shady Ali. "Moheet Corpus: A Shift in the Architecture of Shi'i Islamic Knowledge" (in Arabic, English and Persian). London: Sadiqoon Technologies, September 2026. https://moheet-ablibrary.net/en/docs/moheet-paper.html
3. Romanov, Maxim, and Masoumeh Seydi. "OpenITI: A Machine-Readable Corpus of Islamicate Texts." Zenodo, 2019–. https://openiti.org ; and the KITAB project (Institute for the Study of Muslim Civilisations, Aga Khan University, London). https://kitab-project.org
4. Kamran, Amna Binte, Bakhtawar Abro, and Amna Basharat. "SemanticHadith: An Ontology-Driven Knowledge Graph for the Hadith Corpus." *Journal of Web Semantics*, vol. 78, 2023, Article 100797.
5. Safe, M., et al. "Constructing a Knowledge Graph for Hadith Narrators." *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 6, 2022, pp. 3385–3396.
6. Qdrant. *Qdrant Vector Database Documentation*. https://qdrant.tech/documentation
7. Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." *Proceedings of EMNLP-IJCNLP 2019*, ACL, 2019, pp. 3982–3992.
8. Karpukhin, Vladimir, et al. "Dense Passage Retrieval for Open-Domain Question Answering." *Proceedings of EMNLP 2020*, ACL, 2020, pp. 6769–6781.
9. Cormack, Gordon V., Charles L. A. Clarke, and Stefan Büttcher. "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods." *Proceedings of SIGIR 2009*, ACM, 2009, pp. 758–759.

10. Malkov, Yu. A., and D. A. Yashunin. "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, 2020, pp. 824–836.
11. Hogan, Aidan, et al. "Knowledge Graphs." *ACM Computing Surveys*, vol. 54, no. 4, 2021, Article 71.
12. Traag, V. A., L. Waltman, and N. J. van Eck. "From Louvain to Leiden: Guaranteeing Well-Connected Communities." *Scientific Reports*, vol. 9, 2019, Article 5233.
13. Edge, Darren, et al. "From Local to Global: A Graph RAG Approach to Query-Focused Summarization." arXiv:2404.16130, 2024.
14. Model Context Protocol. *Specification*. https://modelcontextprotocol.io
15. Lewis, Patrick, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems 33*, 2020.
16. ISO 2859-1:1999, *Sampling procedures for inspection by attributes*; ISO/IEC 25012:2008, *Data quality model*; IFLA, *Library Reference Model (LRM)*, 2017.