Int. J. Sci. R. Tech., 2026 3(7)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

Development Of An AI-Ready Structured Knowledge Base For Classical Homeopathic Materia Medica: A Conceptual Database Architecture And Knowledge Representation Framework

Aafiya Sheikh[1], Nazim Dadan[2*]

[1]*Department of Paediatrics, Dr. G. D. Pol Foundation Homoeopathic Medical College, Navi Mumbai, Maharashtra, India*
[2]*Homeopathic Consultant, SweetPills Homeopathy Clinics, Navi Mumbai, Maharashtra, India*

ABSTRACT

The classical homeopathic materia medica the works of Hahnemann, Kent, Boericke, Allen, and Clarke encodes more than two centuries of clinically verified remedy knowledge, yet that knowledge remains locked in narrative prose that resists computational retrieval, cross-author comparison, and integration with modern artificial-intelligence tools. This short communication sets out the design of a structured, AI-ready knowledge base for the classical materia medica. We propose a conceptual framework rather than a validated software implementation. At its core is an entity–relationship data model that captures remedies, symptoms, modalities, mind and generalities, and inter-remedy relationships (complementary, inimical, and follows-well), with every record tagged to the source text from which it derives; we then weigh relational against graph database architectures for representing these data. After distinguishing the proposal from a conventional repertory, we outline an extraction pipeline that pairs OCR-cleaned digitized source texts with named-entity recognition and expert curation, and we describe three representative applications: structured symptom search, repertorization support, and retrieval-augmented generation (RAG) for grounding large language model responses in verified classical sources. Because the architecture is proposed rather than deployed, the work is offered as a template for digitizing traditional-medicine corpora while preserving fidelity to the primary texts.

Keywords: Homeopathy; Materia Medica; Repertory; Knowledge Base; Database Architecture; Retrieval-Augmented Generation.

INTRODUCTION

The clinical practice of homeopathy rests on a defined canon of nineteenth- and early twentieth-century texts. Hahnemann's *Organon of Medicine*, published across six editions between 1810 and 1842 (the sixth appearing posthumously only in 1921), lays out the methodological framework of case-taking and individualization, while his *Chronic Diseases* (1828) introduces the miasmatic theory that underlies the treatment of recurring conditions [1,2]. Kent's *Lectures on Homoeopathic Materia Medica* (1904) reframed drug study around narrative "drug pictures," giving mental and emotional generals priority over physical particulars [3]. Boericke's *Pocket Manual* compressed verified clinical symptoms into a dense practical reference [4], Allen's *Keynotes* (1898) introduced a graded, comparative method for differentiating similar remedies [5], and Clarke's three-volume *Dictionary of Practical Materia Medica* (1900–1902) cross-references clinical indications, characteristics, and remedy relationships [6]. Despite their enduring clinical importance, these texts remain predominantly narrative in structure, which limits computational analysis, interoperability, and semantic retrieval.

Each of these works organizes knowledge differently, and none was designed for machine querying. Digital tools such as RadarOpus, Hompath Zomeo, and Vithoulkas Compass have improved retrieval speed across large digitized libraries, and Vithoulkas

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Compass in particular applies Bayesian likelihood-ratio statistics to rank symptom–remedy associations drawn from pooled outcome data. These platforms, however, rely on proprietary, closed schemas. A 2023 international survey found that although 71% of practitioners regarded such software as valuable, cost and insufficient training remained significant barriers to adoption [7]; and recent evaluations of general-purpose AI chatbots on homeopathic case data reported only 6% agreement with practitioner recommendations, a reminder that ungrounded language models are not yet reliable decision aids in this domain [8].

HOW THIS DIFFERS FROM A REPERTORY

A repertory is, by design, a one-directional index: it maps a symptom (rubric) to a graded list of remedies, and nothing more. It cannot natively answer the reverse query which symptoms a given remedy produces nor can it represent remedy-to-remedy relationships, potency data, or bibliographic provenance without switching to an entirely separate reference work, whether the materia medica itself or a relationship chart. A repertory entry also carries no structured record of which author, edition, or clinical case contributed a given grading; the grading appears as a fixed numeral with no machine-readable trace back to its source.

The architecture proposed here is not a digitized repertory but the underlying knowledge base from which a repertory-style view and several other views can be generated. Remedies, symptoms, modalities, and inter-remedy relationships are stored as interlinked records, each tagged to its originating text and edition, so that a query can traverse the data in any direction: symptom to remedy, remedy to its full symptom profile, or remedy to its documented relationships with other remedies. This is the basis for the applications described later, and for RAG-based question answering in particular, which depends on being able to cite the specific source record behind an answer something a conventional repertory grading was never built to provide. Accordingly, the proposed knowledge base should be viewed as the underlying data infrastructure from which multiple representations including repertories, materia medica views, relationship charts, and AI retrieval systems can be derived.

PROPOSED DATA MODEL AND DATABASE ARCHITECTURE

We define six core entity types: Remedy (the source substance, with its Kingdom Plant, Mineral, or Animal), Symptom (a free-text description linked to a controlled vocabulary), Modality (an aggravating or ameliorating factor, with polarity and intensity), Mind/Generalities (constitutional and psychological traits, following Kent's hierarchy), Relationship (complementary, inimical, antidote, or follows-well relationships between remedies), and Source (author, edition, and page reference). Every symptom or relationship record is provenance-tagged to its originating author and edition, preserving Allen's verification-grade information and Clarke's clinical cross-references as structured attributes [5,6]. The conceptual entity–relationship model is illustrated in Figure 1.

Entity	Purpose	Example Attributes
Remedy	Canonical medicine	Name, Kingdom
Symptom	Clinical observation	Description, Body region
Modality	Aggravation / amelioration	Factor, Polarity
Mind / Generality	Constitutional features	Category, Description
Source	Provenance	Author, Edition, Page
Remedy Relationship	Inter-remedy links	Complementary, Antidote, Follows-well

Table 1. Core entities of the proposed knowledge base.

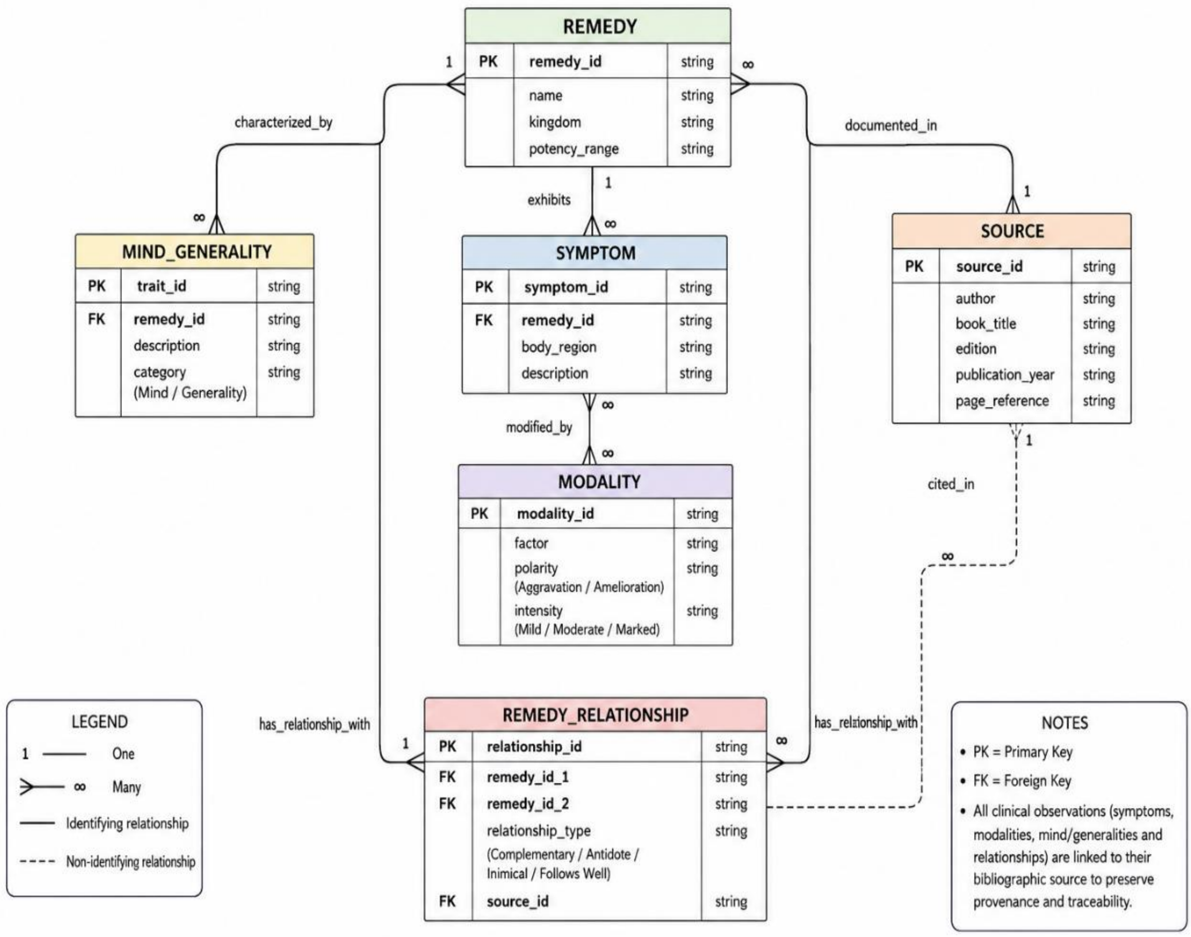

Figure 1. Entity–relationship model of the proposed AI-ready materia medica knowledge base. The schema shows the principal entities, their attributes, and the relationships needed to preserve provenance while supporting bidirectional querying and knowledge-graph generation.

Materia medica data is intrinsically network-shaped: a remedy connects to hundreds of symptoms, those symptoms connect to modalities, the modalities recur across many remedies, and the remedies connect to one another through documented relationships. Comparative benchmarking of relational and graph databases on healthcare data shows graph databases executing multi-hop relationship queries substantially faster than relational joins as query complexity grows, albeit at the cost of a larger storage footprint [9]. Because repertorization-style queries are fundamentally multi-hop, we propose a property-graph architecture (for example, Neo4j) as the primary store, complemented by a relational layer for structured bibliographic metadata (Figure 2).

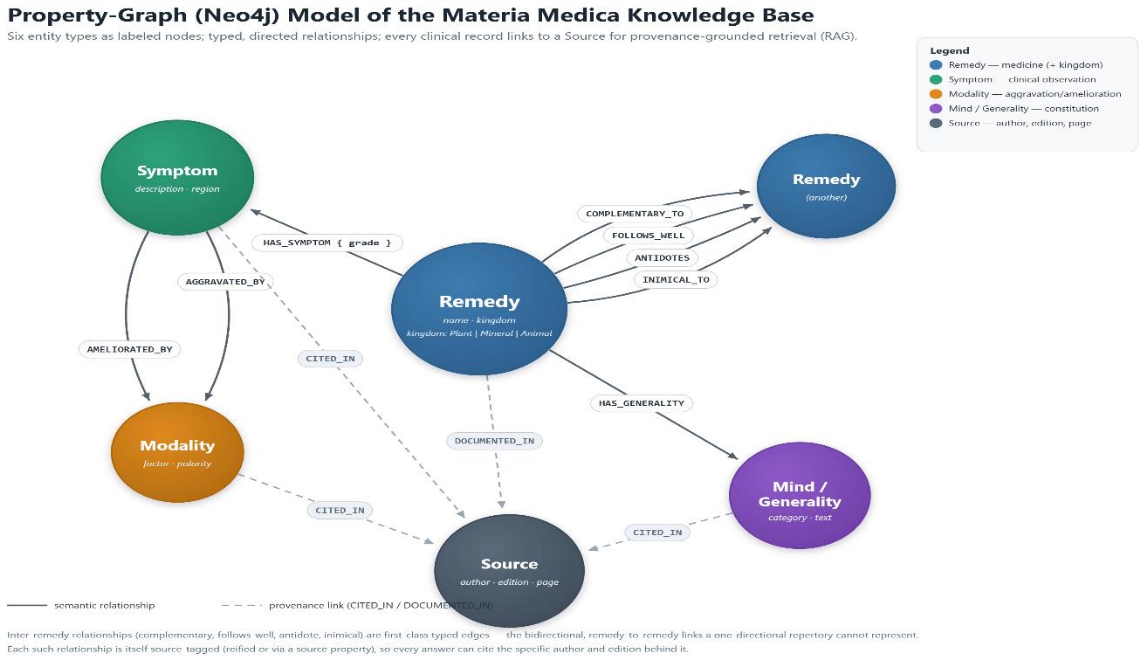

Figure 2. Property-graph (Neo4j) realization of the proposed knowledge base. Each of the six entity types is a labeled node, and every clinical fact is stored as a typed, directed relationship. The four inter-remedy relationship types (complementary, follows-well, antidote, and inimical) are represented as first-class edges, and each symptom, modality, generality, and inter-remedy relationship links to a Source node, so that any retrieved fact can be traced to its originating author and edition for provenance-grounded retrieval-augmented generation.

EXTRACTION PIPELINE AND AI-READINESS

Populating this schema calls for a staged pipeline: (i) acquisition of public-domain digitized editions, available through archive.org and comparable open archives [1,2]; (ii) OCR cleanup, since scanned nineteenth-century typefaces introduce recognition noise that degrades downstream extraction accuracy; (iii) named-entity recognition to identify remedy names, symptom phrases, and modality keywords a task on which domain-adapted transformer models have shown higher precision on clinical text than general-purpose language models [13]; and (iv) mandatory expert curation, because automated extraction cannot reliably resolve period-specific terminology without human verification.

AI-readiness is built into the schema rather than bolted on afterward: every record carries an embedding-compatible text field for semantic search, and the graph structure exports directly to a retrieval corpus for retrieval-augmented generation. This follows evidence from adjacent traditional-medicine digitization work, where the AyuRAG framework reported high accuracy when a language model was grounded in a structured, chunked knowledge base rather than in the model's parametric knowledge alone [11]. RAG is preferable to fine-tuning here because it preserves source traceability: given that even leading language models fabricate plausible but false medical details in a substantial minority of cases [10], a provenance-tagged knowledge base following the logic of ontology-enhanced rare-disease knowledge-graph construction [12] provides a verifiable substrate against which any AI-generated output can be checked.

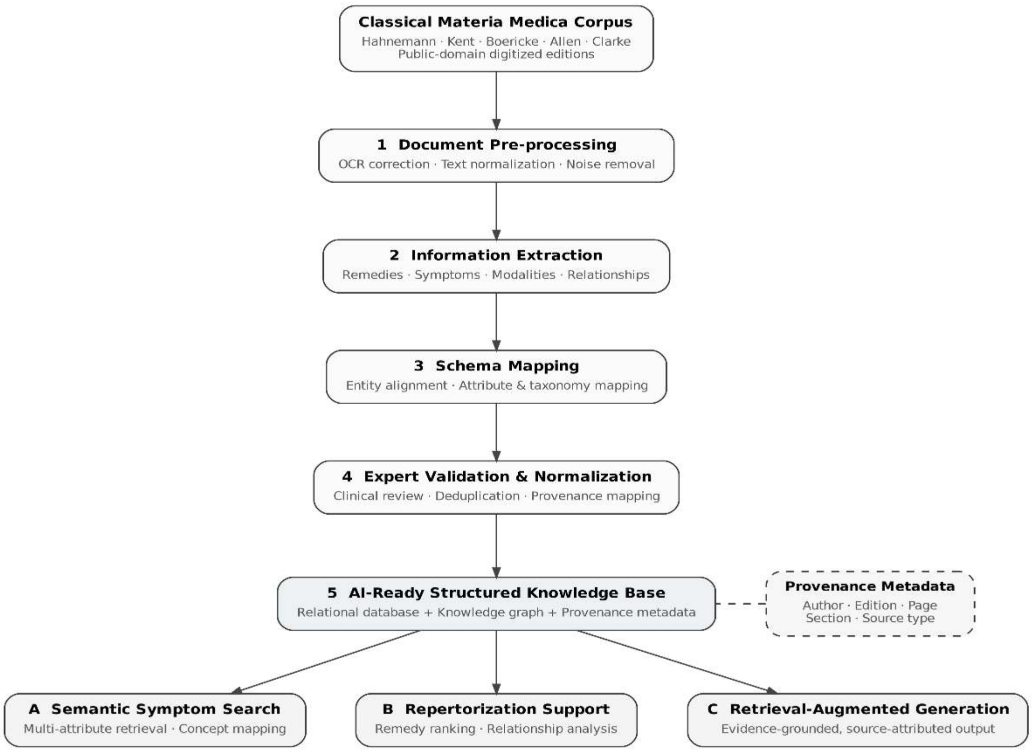

Figure 3. Proposed workflow for constructing an AI-ready structured knowledge base from the classical homeopathic materia medica. Public-domain texts pass through document pre-processing, information extraction, schema mapping, expert validation, and provenance-aware normalization before being integrated into a structured knowledge base that supports semantic retrieval, repertorization, and retrieval-augmented generation.

REPRESENTATIVE APPLICATIONS

Three applications follow from this architecture. Structured symptom search resolves a query such as "remedies with headache aggravated by motion and ameliorated by pressure" into a graph traversal that returns a ranked, source-attributed candidate list. Repertorization support automatically flags documented complementary or inimical relationships between remedies   information easily missed when cross-referencing five separate texts by hand. RAG-grounded question answering lets a natural-language clinical query be answered by retrieving the relevant structured records and citing the specific remedy entry and source edition behind the response, directly addressing the low agreement rates observed when large language models are used for homeopathic case analysis without such grounding [8].

DISCUSSION AND LIMITATIONS

This is a proposed design, not a validated deployment. We report no extraction-accuracy, query-latency, or retrieval-precision figures here, because no working system has yet been built and benchmarked; empirical evaluation against a gold-standard, expert-annotated subset of the corpus is the necessary next step. Classical materia medica also carries an interpretive subjectivity that no schema can fully resolve symptom language varies by author, era, and translation   so expert curation remains a required part of the pipeline rather than a step to be automated away. Finally, the knowledge base is conceived as a means of organizing and retrieving classical textual knowledge; it makes no claim about the clinical efficacy of homeopathy, and any downstream clinical-support application would require independent clinical validation.

Future work should evaluate extraction accuracy against expert-annotated benchmark datasets and compare graph-based retrieval with conventional relational database implementations. Aligning the schema with established biomedical standards and ontologies could further ease integration with broader health-informatics ecosystems while preserving the provenance of the classical homeopathic literature.

CONCLUSION

The classical homeopathic materia medica is at present accessible only as narrative text distributed across five authors, with existing repertories offering a single, one-directional index into that text rather than a query able knowledge structure. We have proposed a property-graph-centered data model Remedy, Symptom, Modality, Mind/Generalities, Relationship, and Source entities together with a staged extraction pipeline and a set of representative AI-grounded applications. The proposed framework establishes a foundation for provenance-aware, interoperable digital representations of classical homeopathic knowledge and provides a platform for future research into semantic retrieval, knowledge graphs, and explainable AI systems grounded in primary-source literature.

REFERENCES

1. Hahnemann S. Organon of Medicine. 6th ed. Boericke W, translator. New Delhi: B. Jain Publishers; 1922.
2. Hahnemann S. The Chronic Diseases: Their Specific Nature and Homoeopathic Treatment. Tafel LH, translator. Philadelphia: Boericke & Tafel; 1896.
3. Kent JT. Lectures on Homoeopathic Materia Medica. Philadelphia: Boericke & Tafel; 1905.
4. Boericke W. Pocket Manual of Homoeopathic Materia Medica. 9th ed. San Francisco: Boericke & Runyon; 1927.
5. Allen HC. Keynotes and Characteristics with Comparisons of Some of the Leading Remedies of the Materia Medica. Philadelphia: Boericke & Tafel; 1898.
6. Clarke JH. A Dictionary of Practical Materia Medica. Vols 1–3. London: Homoeopathic Publishing Co.; 1900–1902.
7. Gray AC, Pracjek P, Straiges D. Attitudes to and uptake of repertory software in homeopathy clinical practice results of an international survey. Homeopathy. 2023;112(2):97–106. doi:10.1055/s-0042-1748841.
8. Doherty R, Pracjek P, Luketic CD, Straiges D, Gray AC. Comparing AI chatbots to live practitioners of homeopathy: a comparative retrospective study. Healthcare (Basel). 2026;14(7):909. doi:10.3390/healthcare14070909.
9. Van der Aa J, Elfrink SH, Smit JW, van Gils SA, de Boer A, Tebeje TH, et al. Can Neo4j replace PostgreSQL in healthcare? Stud Health Technol Inform. 2020;270:1292-1296. doi:10.3233/SHTI200426.
10. MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models. arXiv preprint. 2025. arXiv:2502.14302.
11. AyuRAG: Ayurvedic Knowledge Integration with Retrieval-Augmented Lightweight Large Language Models. ResearchGate preprint. 2025.
12. Liu S, Wang A, Xiu X, Zhong M, Wu S. AutoRD: An Automatic and End-to-End System for Rare Disease Knowledge Graph Construction Based on Ontology-Enhanced Large Language Models. JMIR Med Inform. 2024;12:e60665. doi:10.2196/60665.
13. Yashwanth N, Nikhil S, Shashank R, Sidenur VC, Megha V. Named Entity Recognition for Healthcare: A Comparative Survey. ICT: Applications and Social Interfaces. 2026.

HOW TO CITE: Aafiya Sheikh[1], Nazim Dadan[2]*, Development Of An AI-Ready Structured Knowledge Base For Classical Homeopathic Materia Medica: A Conceptual Database Architecture And Knowledge Representation Framework, Int. J. Sci. R. Tech., 2026, 3 (7), 975-980. https://doi.org/10.5281/zenodo.21641830