Enhancing Semantic Search with Retrieval-Augmented Generation and Agentic AI

Francisco Azeredo
Academic Advisor: Prof. Sérgio Luís Proença Duarte Guerreiro
Industrial Advisor: Eng.º Filipe Mendes Correia (Link Consulting SA)

Instituto Superior Técnico, Universidade de Lisboa

MSc Thesis Defense

Presentation Structure

1. Motivation and Problem
2. Objectives
3. Architecture and Components
4. Related Work (Context)
5. Evaluation and Results
6. Conclusions and Future Work

Enterprise Information Challenges

Critical Business Context:

- Enterprises rely on *accurate, current* information for project approvals, compliance decisions, and operational governance
- **High-stakes work**: Incorrect or outdated information can lead to regulatory violations, project failures, or financial losses

Current Problems:

- Professionals spend significant time searching, reading, and analyzing multiple documents
- Manual verification of document dependencies creates **high cognitive load**
- Decisions risk relying on superseded or nullified information
- **Even modern semantic search systems fail** in critical situations—they cannot guarantee information currency or detect contradictions

Context: Edoclink Enterprise

Document Management Platform:

- Organizes documents with structured workflows and business rules
- Supports evolution from ad-hoc to complex workflow configurations
- Enables end-to-end process automation and digitalization
- Features: document lifecycle management, collaborative work, ERP integration
- Used in public sector and enterprises with rich, complex processes

Thesis Opportunity: Leverage this rich structure to enhance semantic search and retrieval accuracy.

Problem Illustration

put this image https://www.bbc.com/news/war-in-ukraine

Problem: Example Scenario

Example Timeline

- Jan: Conflict escalation reported (doc A)
- Mar: Ceasefire agreement reached (doc B)
- Apr Query: "What is the current status of the War in Ukraine?"

Risk: Retrieves A, ignores B.

Core Challenge: Capture and reason over document dependencies.

Problem

- Current search methods don’t account for information interdependencies
- GraphRAG ideas emerging; enterprise schemas underused

Conclusion: Need interdependency-aware, structure-guided retrieval.

Objectives Overview

- **Represent** enterprise document structure and relations
- **Detect** updates / contradictions across documents
- **Enterprise Ready:**
  - **Standardize** solution in emerging frameworks
  - **Scale** to large information repositories
- **Evaluate**: Evaluate efficiency, and quality of solutions

Objective Details

1. Schema design (workflows, entities, metadata)
2. Graph construction (entity merging, references)
3. Agent traversal (multi-hop, temporal ordering)
4. MCP tools (validated queries, follow references)
5. Evaluation (retrieval accuracy, answer quality)

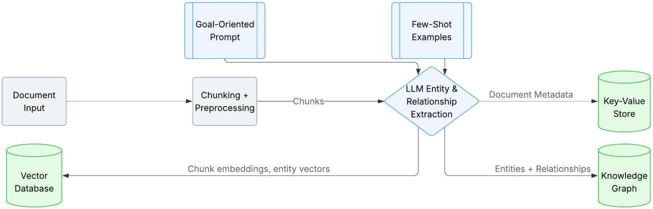
Architecture Overview

We developed 3 independent components for each step in document processing

1. Insertion/Index (Automatic Knowledge Graph construction)
2. Query (Query techniques, some that take better advantage of the graph)
3. Generation (Context given to Agents or LLMs for a readable output)

Knowledge Graph Construction

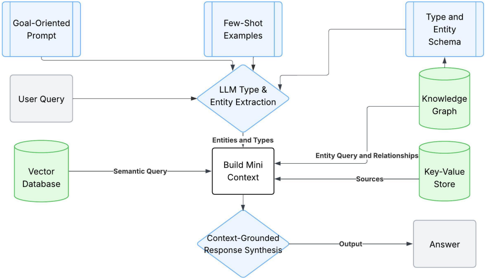
1. Extract entities and relationships
2. Merge nodes with matching entities
3. Store in a graph



Query Techniques

The query techniques go from the one of least latency to highest latency

1. Semantic, BM25, and hybrid queries
2. Query reformulation (extract entities and relations from query) for graph retrieval
3. ReAct (Agent) queries using MCP server



Agent Configuration

1. Instructed to be a document assistant connected to a Weaviate database; answers are concise and grounded with sources.
2. Tools via an MCP server that translate agent requests (natural language) to valid GraphQL over the schema.
3. And descriptions of collections to guide collection selection.

MCP Server (Weaviate Bridge)

Context: Agent–database communication layer

Core Point: Enables schema-aware, validated queries for agentic retrieval

- **Standard tool interface** for agents
- **Schema validation** for queries
- **Key tools:**
  - weaviate-query: hybrid search and direct object retrieval
  - weaviate-origin: return object with appended references context
  - weaviate-follow-ref: follow one-hop references and return referenced objects

Conclusion: Enables agentic graph traversal and reliable, context-grounded answers

Evaluation Questions

- Does MiniRAG successfully capture document interdependency?
- Does the model for retrieval matter?
- How much can Agentic system improve retrieval and answer generation?

MiniRAG Interdependencies

Temporal QA Setting: Dataset encodes time-stamped / evolving facts; correctness depends on capturing updates and supersessions (interdependent information).

Objective: Assess whether MiniRAG's profiling + relational reasoning improves Token Recall (temporal factual accuracy) over naive single-pass RAG.

Token Recall (Temporal QA Benchmark)

System	Benchmark	Thesis (Qwen2.5-3B)
Naive RAG	43%	44%
MiniRAG	49%	38%
MiniRAG (gpt-4o-mini)*	54%	–
MiniRAG Multi-Hop (gpt-4o-mini)*	68.4%	–

1

[1] Benchmarking with larger reasoning models was not performed due to prohibitive computational costs relative to

Does the model matter?

Agentic RAG: 60.4% retrieval, 61% correct answers (∼2–4× over baselines)

Approach	Calls	Cost (300 Q)	Answer	Retrieval
Naive RAG (Qwen2.5)	1	0	19%	14%
Naive RAG (GPT-5)	1	$5.07	12%	14%
Agentic ReAct (GPT-5)	2–20	$18.35	61%	60%

Table: Agentic ReAct used an average of 5 LLM calls per question

Multi-Collection Retrieval

Agentic ReAct across multiple collections shows improvements due to curated collection descriptions that guide more informed semantic queries.

- Correct collection selection
- 2–3× improvement across strategies
- Robust to format variation

Approach	Avg. Retrieval Rate
Agent_OpenAI_MixedLiHua	93.0%
Agent_OpenAI_Mixed	86.0%
Agent_OpenAI_MixedSynthetic	80.5%
Agent_OpenAI_300	60.4%

Table: Share of questions with the correct document retrieved, by result file

Summarization, compressing storage

Optimizations that select the most relevant text improve retrieval by reducing noise stored in the database.

Table: Document retrieval and summarization metrics

Approach	Retrieval	Token Recall	Jaccard
Agentic ReAct (GPT-5)	60.4%	61.1%	55.7%
LexRank	33.6%	29.5%	35.9%
BART	17.8%	3.0%	11.1%
Naive RAG (Qwen2.5)	13.8%	19.1%	22.1%

Using Youden's J thresholds for classification boundaries.

Integration for Edoclink's workflow organization

- Use Weaviate cross-references in a workflow-organized company database.
- Enforce ingestion rules for graph consistency and provenance.
- Agent performs multi-hop queries: files → stages → flows → entities.
- Store information in focused snippets for robust retrieval within context limits.

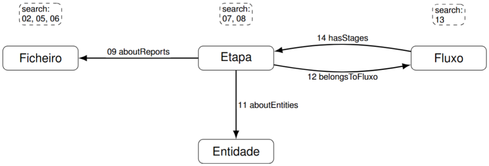


Figure A.2: Q1 walkthrough: base searches at each class and schema-guided hops.

Figure: Agent path over cross-referenced workflow to answer a question. "How is the war on Russia and Ukraine?"

Final Answer

Final answer. It remains an active, violent conflict.
Grounding (evidence).

- A matched stage in the “Russia vs Ukraine War” flow is tied to a report dated 2024-07-12 titled “Drone attack hits military base”.
- Excerpt: “A drone strike has reportedly destroyed parts of a military base near the front. Witnesses describe significant damage and casualties.”

Interpretation. • Continued strikes on military targets near the front and ongoing casualties, involving both Russia and Ukraine as the referenced entities

Related Work:

- Google NotebookLM: personal, source-grounded assistant over curated notes; footnoted syntheses and study guides
- Perplexity AI: web-grounded conversational search with inline citations and multi-document aggregation
- Vector databases : Weaviate, Qdrant, and Milvus.
- Multi-Hop Agentic Retrieval: semi-structured enterprise information in ruled workflow configurations (e.g., Edoclink)

Comparison to Prior Approaches

My solution Multi Hop Agentic Retrieval.

- Integrates schema-driven traversal
- Standardizes graph access (MCP server)
- Focus: enterprise interdependency reliability

Conclusions

- **Graph-aware reasoning**: prevents outdated answers
- **Structure-aware retrieval**: leverages cross-references and metadata for precision
- **Well-instructed agents**: schema-aware tools guide retrieval and sourcing
- **Tool standardization** enables reproducible agent pipelines

Conclusions

- **Graph-aware reasoning**: prevents outdated answers
- **Structure-aware retrieval**: leverages cross-references and metadata for precision
- **Well-instructed agents**: schema-aware tools guide retrieval and sourcing
- **Tool standardization** enables reproducible agent pipelines

Conclusions

- **Graph-aware reasoning**: prevents outdated answers
- **Structure-aware retrieval**: leverages cross-references and metadata for precision
- **Well-instructed agents**: schema-aware tools guide retrieval and sourcing
- **Tool standardization** enables reproducible agent pipelines

Conclusions

- **Graph-aware reasoning**: prevents outdated answers
- **Structure-aware retrieval**: leverages cross-references and metadata for precision
- **Well-instructed agents**: schema-aware tools guide retrieval and sourcing
- **Tool standardization** enables reproducible agent pipelines

Limitations

- Scale constrained (compute / budget)
- No public benchmark for workflow organized information (construction overhead)
- NLP ecosystem is highly dynamic; conclusions are time-bound
- Model and embedding churn can change retrieval behavior
- RAG and agent frameworks evolve; pin versions, datasets, prompts, and eval protocols

Future Work

- Integration with Edoclink platform, including it's access control layers.
- Beta testing multi-hop agentic retrieval in real-world enterprise settings.
- Benchmark release (structure-aware retrieval)
- Cache implementation for latency reduction in multi-hop retrieval
- Agentic Workflows (e.g, better Graph RAG, Relevant/Outdated information management, chatbot memory)
- More MCP for other applications

Thank You

Thank you!

Francisco Azeredo
Instituto Superior Técnico
Questions?