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

TÉCNICO LISBOA

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

Problem Illustration

NEWS War in Ukraine

Four dead in Russian attack as diplomatic efforts to end war continue

As US negotiating teams continue to hold talks with Moscow and Kyiv, Russia's attacks on Ukraine continue.

3 hrs ago | Europe

Ukraine talks 'productive' but more work needed, Rubio says

American and Ukrainian delegations meet in Florida to discuss the outlines of a peace deal with Russia.

16 hrs ago | World

Putin doubles down on demands for Ukrainian territory ahead of talks with US in Moscow

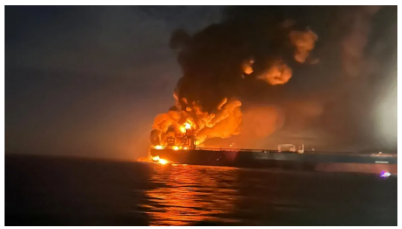
The Russian president accuses Kyiv of wanting to fight "to the last Ukrainian" - which he says Russia is also "in principle" ready to do.

4 days ago | Europe

Trump defends Witkoff after leak appears to show envoy coaching Russia

The US president said he had not heard the audio, but that it sounded like "standard" negotiations.

5 days ago | World



Ukraine hits tankers in Black Sea in escalation against Russia

The two ships struck by drones were thought to be used to bypass Western sanctions on Russia.

1 day ago | Europe

Russian strikes cause power outages for more than 600,000 in Ukraine

The overnight attack kills at least three and injures dozens more.

2 days ago | Europe

Hungary's Orban defies EU partners and meets Putin again in Moscow

Russia's leader praised what he called Orban's "balanced position on the situation in Ukraine".

3 days ago | Europe

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 information interdependencies.

Enterprise Information Challenges

Critical Business Context:

- Decisions require accurate, current information
- **High stakes**: compliance, approvals, financial risk

Current Problems:

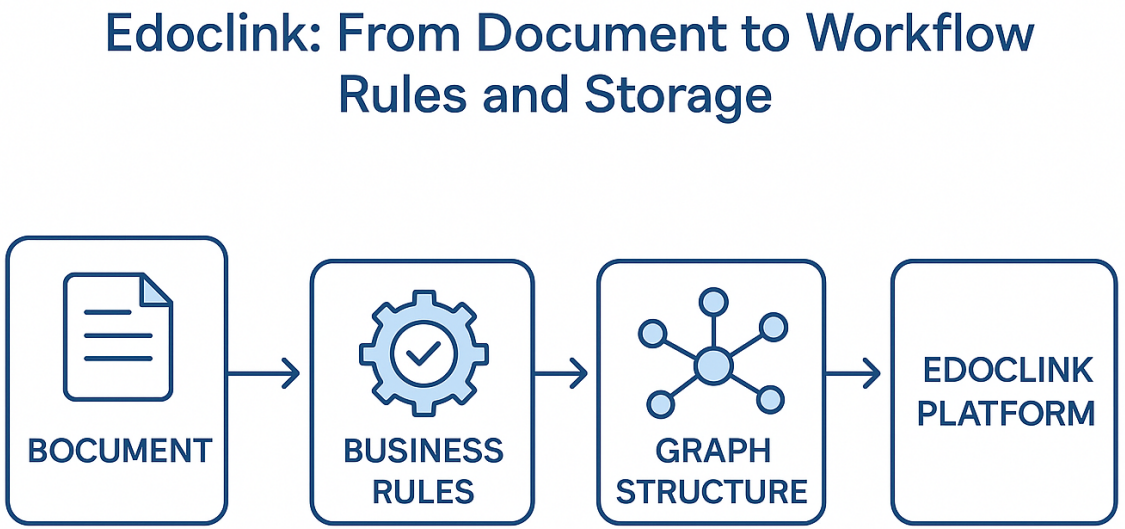
- Time spent searching and reading
- Manual dependency checks = **high cognitive load**
- Reliance on superseded information
- **Modern semantic search fails** to ensure currency/contradictions

Context: Edoclink Enterprise

Document Management Platform:

- Workflow-driven lifecycle and business rules
- Supports evolution from ad-hoc to complex workflow configurations
- End-to-end automation and digitalization
- Features:document lifecycle management, collaboration, ERP integration
- Deployed in public sector and enterprises

Edoclink Workflow Illustration



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

Objectives Overview

- **Represent**: enterprise structure and cross-references
- **Detect**: updates and contradictions
- **Enterprise-ready**:
  - **Standardize**: emerging frameworks (MCP)
  - **Scale**: large repositories
- **Evaluate**: efficiency and quality

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)

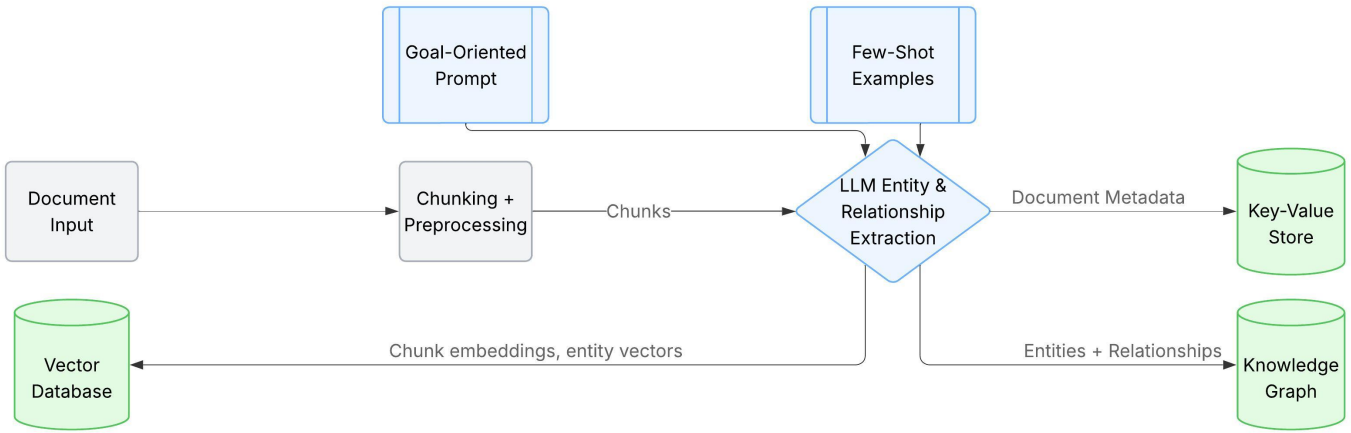
Architecture Overview

Three independent components across the pipeline:

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

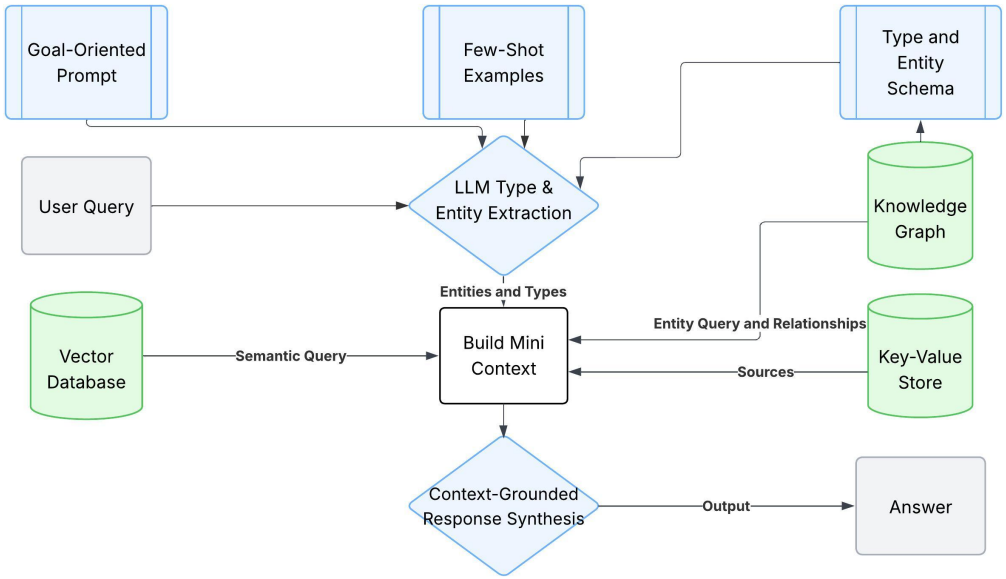
1. Extract entities and cross-references
2. Merge nodes with matching entities
3. Store in a graph



Query Techniques

From lowest to highest latency:

1. Semantic, BM25, and hybrid queries
2. Query reformulation (entities/relations) for graph retrieval
3. ReAct (Agent) via MCP server



Agent Configuration

1. **Instructions:** Document search assistant connected to Weaviate; deliver concise, sourced answers.
2. **Tools:** MCP server translating natural language to schema-valid GraphQL.

Weaviate MCP Server

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: Time-stamped facts; correctness depends on updates/supersessions.

Objective: Test whether MiniRAG profiling + relations improve Token Recall vs 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 expected benefits.

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 benefits from curated collection descriptions that guide semantic queries.

- More accurate 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 & Storage

Selecting relevant text reduces noise and improves retrieval.

Table: Document retrieval and summarization metrics

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

Thresholds via Youden's J statistic.

Edoclink Integration

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

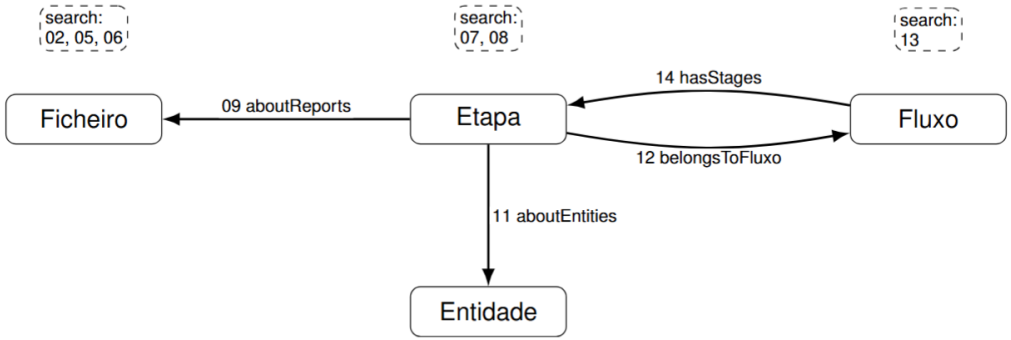


Figure A.2: Q1 walkthrough: base searches at each class and schema-guided hops.

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.

Comparison to Prior Approaches

Multi-hop Agentic RAG for workflow-structured enterprise content (e.g., Edoclink)

- 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

- Integrate with Edoclink, including its access controls
- Beta-test multi-hop Agentic RAG in production
- Release structure-aware retrieval benchmark
- Add caching to reduce multi-hop latency
- Agentic workflows: Graph RAG, outdated info handling, memory
- Additional MCP tools for adjacent applications

Thank You

Thank you!

Francisco Azeredo
Instituto Superior Técnico
Questions?