Construction GraphRAG - System Architecture

# Diagram 1: Document Ingestion Process

| **📄 PDF DOCUMENT** Construction Plans, Specifications, Drawings |
| --- |

**↓**

| **STEP 1: TEXT EXTRACTION**  Tool: PyMuPDF (fitz)  Process in 2-page batches  Extract page-by-page |
| --- |

**↓**

| **STEP 2: TEXT CHUNKING**  Tool: Custom Splitter with Memory Management  Size: 800 characters | Overlap: 150  Chunk each page separately  Track: (chunk, page\_number) |
| --- |

**↓**

| **STEP 3: ENTITY EXTRACTION AND RELATIONSHIP MANAGEMENT**  Tool: Regex matching and LLM extraction  Entities and relationships are extracted using both regex and llm matching. |
| --- |

**↓**

| **STEP 4: GENERATE EMBEDDINGS**  Tool: OpenAI text-embedding-3-small  Output: 1536-dimensional vectors  FIXED: Consistent model (was mixed)  Dimension validation added |
| --- |

**↓**

**Storage (3 Paths):**

| **PATH 1 VECTOR STORAGE**  ChromaDB  Stores:  • Embeddings  • filename  • page  • text | **PATH 2 KEYWORD INDEX**  BM25  Builds:  • TF-IDF index  • Exact matching | **PATH 3 GRAPH STORAGE**  Neo4j  Stores:  • Entities  • Relationships  • Connections |
| --- | --- | --- |

**↓**

| **✅ DOCUMENT READY FOR QUERYING**  ✅ No MemoryError crashes  ✅ Correct embedding dimensions  ✅ Real page numbers tracked |
| --- |

# Diagram 2: Query & Retrieval Process

| **🔍 USER QUERY** "What is refrigeration engine room equipment plan?" |
| --- |

**↓**

| **STEP 1: QUERY EXPANSION & EMBEDDING**  Expand to 3-5 query variations  Construction domain synonyms  Tool: OpenAI text-embedding-3-small  FIXED: 1536 dimensions (consistent) |
| --- |

**↓**

**STEP 2: PARALLEL SEARCH (3 Methods)**

| **METHOD 1 VECTOR SEARCH**  ChromaDB  Cosine Similarity  Top-K: 50  (was: 25)  Purpose:  Semantic matching | **METHOD 2 KEYWORD SEARCH**  BM25  TF-IDF Algorithm  Top-K: 50  (was: 25)  Purpose:  Exact matching | **METHOD 3 GRAPH SEARCH**  Neo4j  Cypher Queries  Limit: 10 facts  Purpose:  Relationships |
| --- | --- | --- |

**↓**

| **STEP 3: RECIPROCAL RANK FUSION (RRF)**  Merge all 3 search results  Deduplication across queries  Weighted scoring  Combined ranking algorithm |
| --- |

**↓**

| **STEP 4: SECTION BOOSTING**  Priority multipliers: (boosts the score based on chunk from document type).  • General Notes: 2.5x  • Equipment Notes: 2.3x  • Diagrams: 1.5x |
| --- |

**↓**

| **STEP 5: CONTEXT SELECTION**  Select top 50 chunks (was: 15)  FULL text per chunk (no truncation)  Total context: ~25,000 characters |
| --- |

**↓**

| **STEP 6: ANSWER GENERATION**  Tool: GPT-4o  Temperature: 0.35 (better reasoning)  Max tokens: 2000 (longer answers)  Input: Query + 50 chunks + Graph facts |
| --- |

**↓**

| **✅ FINAL ANSWER WITH CITATIONS**  ✅ Detailed, comprehensive response  ✅ 5-15 sources cited  ✅ Real page numbers (3, 5, 7...)  ✅ Equipment tags & specifications |
| --- |