

GraphRAG System

Technical Architecture Overview

Version 1.0 | March 2026

1. Abstract

This document presents the technical architecture of a Multimodal GraphRAG System designed for intelligent document parsing and knowledge graph construction. The system integrates MinerU for document parsing, LangExtract for structured entity extraction, and a graph database for knowledge storage and retrieval.

The pipeline supports multiple document formats including PDF, DOCX, PPTX, and image files. Extracted entities and relations are stored as graph nodes and edges, enabling semantic search and question answering over large document collections.

2. System Components

2.1 Document Parsing Module

MinerU Cloud API (v4) serves as the document parsing backend. It accepts PDF, DOCX, PPTX, PNG, JPG, and HTML files. Output includes Markdown text, structured `content_list.json`, and extracted images.

2.2 Entity Extraction Module

LangExtract (v1.1.1) performs structured information extraction from plain text using few-shot prompting with LLM backends (Gemini, OpenAI, or local Ollama). Each extraction includes character-level position anchoring.

2.3 Knowledge Graph Module

Extracted entities and relationships are stored in a graph database. Node types include: Person, Organization, Location, Event, Concept. Edge types include: RELATED_TO, BELONGS_TO, CAUSED_BY, LOCATED_IN.

2.4 Retrieval Module

The retrieval layer supports hybrid search combining vector similarity and graph traversal. Query results are ranked by relevance score and returned with source document references.

3. Data Pipeline

The end-to-end data pipeline consists of the following stages:

Stage 1: Document Ingestion

- Accept raw documents (PDF, DOCX, images, HTML)
- Submit to MinerU API for parsing
- Poll task status until state = done

Stage 2: Content Extraction

- Download and decompress full_zip_url
- Parse content_list.json into Document objects
- Separate text blocks, tables, images, equations

Stage 3: Entity & Relation Extraction

- Feed text blocks to LangExtract
- Extract entities with char_interval positions
- Extract relationships between entities

Stage 4: Graph Construction

- Map extractions to graph nodes and edges
- Store with source provenance (page_idx, bbox)
- Build vector embeddings for semantic search

4. Supported File Formats

Format	Extension	OCR Required	Model
PDF (text)	.pdf	No	pipeline / vlm
PDF (scan)	.pdf	Yes	vlm
Word	.docx	No	pipeline
PowerPoint	.pptx	No	pipeline
Image	.png / .jpg	Auto	vlm
HTML	.html	No	MinerU-HTML

5. API Configuration Reference

The following environment variables must be configured before running the MinerU parsing service:

```
MINERU_API_TOKEN      : Bearer token for API authentication
MINERU_USER_UID       : User UUID for quota management
MINERU_BASE_URL       : https://mineru.net/api/v4
MINERU_MODEL_VERSION  : pipeline (default) | vlm | MinerU-HTML
MINERU_LANGUAGE       : ch (Chinese) | en (English)
MINERU_IS_OCR         : false (text PDF) | true (scanned PDF)
MINERU_ENABLE_FORMULA: true | false
MINERU_ENABLE_TABLE   : true | false
```

Rate Limits:

- Max file size : 200 MB per file
- Max pages : 600 pages per file
- Daily quota : 2000 pages (high priority)
- Batch limit : 200 files per request