Ingestion Pipeline

Purpose

- Describe how MEAI ingests PDF documents into Supabase chunks. (ingest_01_text_to_supabase.py)

Inputs

- CORE_LIBRARY_DIR and docs/system_pdfs are scanned for PDFs. (ingest_01_text_to_supabase.py)
- Directories named policies or references are excluded from ingestion. (ingest_01_text_to_supabase.py)

Processing Steps

- PDFs are read with pypdf and text is chunked before embedding. (ingest_01_text_to_supabase.py)
- Chunks are embedded and inserted into the meai_chunks table. (ingest_01_text_to_supabase.py)
- Ingestion resumes from the last chunk_index per source_file. (ingest_01_text_to_supabase.py)

Chunking

- CHUNK_CHARS is 900 and OVERLAP is 120. (ingest_01_text_to_supabase.py)
- BATCH_SIZE is 10 and SLEEP_SEC is 0.12. (ingest_01_text_to_supabase.py)

Embedding

- Embeddings use the text-embedding-3-small model. (ingest_01_text_to_supabase.py)

Failure Handling

- Each PDF is wrapped in a try/except that logs errors and continues. (ingest_01_text_to_supabase.py)

Last Verified

- Timestamp: 2026-01-03 16:35:09 EST
- Git branch: main
- Files referenced: ingest_01_text_to_supabase.py