Ingestion and Provenance Pipeline

Technical Reference
Rev. 2026-07-24

Every document that enters FIP is license-gated, deduplicated on a content hash, dual-stored (original bytes plus a searchable record), and provenance-stamped, so a retrieved answer traces back to a specific source under a known right to use it. The stages below name the actual components.

Pipeline stages

01 Sources

Federal Register, the 8 regional fishery management councils, 3 interstate marine-fisheries commissions, SEDAR, NOAA (SERO, SEFSC, Institutional Repository, InPort), regulations.gov, GovInfo (bills, hearings, committee reports), and the peer-reviewed literature (OpenAlex, PLOS, BHL). Corpus is 500K+ documents.

02 Fetch and fallback

Per-source ingesters pull originals through the shared `scraper-sdk` (`fetchBinary` / `fetchText`) with timeouts and retries. Walls are defeated by a headless-browser (Playwright) fallback for 403 sites and a Wayback Machine fallback for moved or 404 pages.

03 License gate default-deny

A commercial-use assessor maps each `data_source` to a posture: `PUBLIC_DOMAIN` (US federal works, statutory SOPPs, public-law compacts), `PER_ITEM` (mixed-rights, must carry a rights string), or `BLOCKED` . Unknown or absent rights resolve to blocked. A per-document copyright-language scan reads each file's own notice. Rights and the decision persist as `rights_raw` and `commercial_use_status` .

04 Canonical spine: `ingestDoc` content-hash dedupe

One shared function for every ingester. It computes a content hash and skips byte-identical re-ingests (`dup_content`) and known `external_id` s (`exists`) before any work. Text extraction runs three tiers: clean text-layer `pdf-parse` , then `ocrmypdf` / Tesseract OCR, then Claude vision on a Ghostscript rasterization for scanned pages, tables, and figures. Provenance is stamped: `source_url` , `data_source` , `document_type` , `document_date` , `external_id` , and an extraction-method record.

Storage, three ways

S3 originals
The original source bytes, preserved verbatim. Link-rot-proof and re-processable if extraction improves.

```
bucket: ARCHIVE_S3_BUCKET
key: content-hash
```

Postgres record
The searchable document row: text, metadata, dates, rights, provenance, body codes, and species.

```
table: documents
full_text, rights_raw,
commercial_use_status
```

RAG chunks
Chunked text plus embedding vectors for semantic retrieval. Written by the embedder, not inline.

```
table: document_chunks
embedding_vec (OpenAI)
```

Embed step. `ingestDoc` inserts and extracts only. `document-embedder-backfill.js` chunks and embeds any doc with `full_text` and no chunks; the `fip-embedder-backfill-all` cron runs every 6 hours so every source becomes RAG-reachable regardless of a dedicated per-source job.

Enrich (the growth loop)

- **Reference-mining.** `reference-miner.js` reads a document's bibliography, LLM-extracts each cited work, dedups against the corpus, and writes copyright-gated candidates to `library_candidates` for review and ingest. Mined references become new sources, so the corpus compounds toward 1,000,000 documents.
- **Entity extraction.** Named people, organizations, and affiliations are pulled from document bodies and bibliographies into the person and organization directory, provenance-linked, growing the expert graph.
- **Cross-linking.** Documents are wired to modules and to each other (assessment to permit, coral publication to FMP and EFH, EIS to rule).

Retrieval and governance

- **FIP AI chat.** Hybrid retrieval over `document_chunks` : dense vectors plus a lexical (tsvector) arm. Every answer is source-cited, rights-aware, and callable on demand via `/api/answers` .
- **License audit.** A recurring corpus-wide re-scan re-checks rights as content grows and soft-quarantines failures (`review_status = 'rejected'` , excluded from retrieval), never deleting, so the trail survives.
- **Dedupe and coverage.** Content-hash dedupe blocks bloat at ingest; a coverage audit measures recovery against the full corpus, never a scanned subset.

500K+	3-tier	1	6h	1M
DOCUMENTS	EXTRACTION	CANONICAL SPINE	EMBED CADENCE	CORPUS TARGET