Step 2 — Chunking & Structural Parsing (Technical Summary)

Objective

Transform raw legal documents into **auditable, retrieval-ready chunks** that preserve legal meaning, provenance, and annex content, while remaining robust to noisy PDF extraction.

What this step produces

- **Sections**: best-effort structural units (articles, numbered sections, annexes) with page ranges and character offsets.
- **Chunks**: token-bounded text blocks with overlap, designed for semantic retrieval and LLM consumption.
- **Strong provenance**: each chunk is traceable to its document, section, page range, and hash (`sha256`).

Key engineering decisions

1. **Fail-soft design**
   - Chunking *never* drops content.
   - If section detection fails, the system falls back to full-document chunking.
2. **Structure is advisory, not authoritative**
   - Headings help grouping and explainability.
   - Retrieval and extraction depend on **chunk content**, not perfect sectioning.
3. **Stateful parsing where necessary**
   - Explicit ignore spans for:
     - *Table des matières*
     - *Appendix summary after signature*
   - This avoids false section creation from metadata.

4. **Annex-smart chunking**
    - Annexes are first-class sections.
    - Enumerated annex items (e.g. reserved matters) are chunked by item boundaries, not by raw token splits.
    - Prevents hundreds of fake “sections” while preserving annex evidence.

Tricky problems encountered

- TOC lines indistinguishable from real headings after PDF extraction.
- French contracts using `1. DEFINITIONS` instead of `ARTICLE 1`.
- Appendix summaries listing annex titles without content.
- Enumerated annex lists (`1. ...; 2. ...;`) being mistaken for headings.
- Over-restrictive heuristics leading to empty outputs.

Mitigations

- Soft heading detection with minimal heuristics.
- Ignore spans instead of aggressive regex rejection.
- Strong invariants enforced by tests (never empty output).
- Annex-aware chunking instead of annex over-sectioning.

Current limits

- Heading detection is heuristic, not a formal grammar.
- Page ranges depend on upstream text extraction quality.
- Some exotic document layouts may still produce imperfect section titles.

These limits are acceptable because:

> Retrieval quality depends primarily on chunk content, not section purity.