Document Auto-Chunk Processing SOP (OCR + Alt Text + Research)

Purpose Create cleaned, structured artifacts from a single Target File—of any source type (PDF, scanned PDF, DOC/DOCX, PPT/PPTX, HTML export, images of text, etc.)—outputting chunk-by-chunk in plain fenced text with ZERO pauses until the file is fully processed. This SOP is source-agnostic and applies to textbooks, mail scans, RPG manuals, technical papers, and more. Features include: OCR, image/icon pre-scan, brief descriptive alt text insertion, jargon detection, and targeted open-web research to ensure correct parsing and domain understanding.

Anchor Files (optional; for alignment only, never chunk-processed)

* [List any outlines, glossaries, or official reference documents relevant to this run]
* When present, anchors guide terminology/slotting; otherwise, process as a general document without imposed skeleton.

Target File (exact name, one per session)

* [Insert exact filename]
* Only this file is processed chunk-by-chunk. All other uploads are anchors or ignored.

Output Structure Skeleton

* Part 1: Front Matter & Executive Summary
* Part 2: Core Chapters/Sections (mirroring source headings)
* Part 3: Appendices (glossary, references, figures/tables list)

Auto-Advance (no handholding)

* After each chunk, do NOT ask permission. Immediately continue to the next chunk.
* Only when the entire Target File is processed, append a final line (outside the fence): FINISHED: [Target File] fully processed.

Pre-Processing Pipeline (run before first chunk)

1. File introspection • Detect: native text vs image-only pages, page count, language(s), multi-column layout, hyphenation, footnotes/endnotes, bibliography sections. • Identify tables, figures, diagrams, equations, and inline icons/symbols.
2. OCR (if needed) • Apply OCR to image-only pages; preserve reading order; attempt column detection; retain headings hierarchy where possible. • Normalize hyphenation and ligatures; standardize quotation marks and dashes.
3. Image/Icon semantic pre-scan (for all PDFs and docs) • Build a lightweight “symbol map” of recurring icons/graphics and their likely meanings (e.g., heart = health/life, controller = gaming, warning triangle = caution/attention, industry pictograms, safety symbols, accessibility icons). • Note ambiguous or domain-specific visuals for later confirmation.
4. Jargon & term detection • Extract a candidate list of domain-specific terms, acronyms, and idioms (e.g., sector standards, product names, legal citations, technical vocabulary).
5. Targeted research prep • For any unfamiliar or ambiguous terms/symbols: queue a brief open-web verification step to ensure correct meanings, disambiguation, and up-to-date usage (prefer primary/official sources). Keep research concise and cite sources in a preface note (outside fenced chunks) if requested.

Research Phase (bounded; run once before chunking)

* Objective: ensure correct parsing/interpretation of jargon and symbols; do not expand into broad literature review.
* Steps: • Look up each queued term/symbol; capture a 1–2 line definition/meaning and canonical spelling/expansion. • Resolve conflicting definitions by preferring official/standards bodies, government sites, or widely recognized authorities.
* Output: a short “Research Notes” preface (outside fenced blocks) listing key disambiguations and symbol meanings (optional; include only if non-obvious terms/symbols were found).

Chunking & Cleaning Workflow For each sequential chunk of the Target File (≈500–900 words, adjusted for cohesion):

1. Strip cruft: navigation, headers/footers, page numbers, watermarks, boilerplate legal text unless substantively informative.
2. Deduplicate and normalize: remove repetitions, unify terminology, standardize headings and bullet styles.
3. Preserve meaning and hierarchy: maintain logical section nesting and cross-references.
4. Figures, images, icons: insert brief bracketed descriptive alt text where the source includes images/figures/icons. • Format: [alt: ] placed in-line where the figure is referenced or as a bullet under a “Figures” subheading for that chunk. • Be concise and purpose-led: what does the image convey in context? E.g., “[alt: flowchart of onboarding steps highlighting ‘Accessibility Review’ stage]”. • If meaning is unknown: use best-effort neutral description, e.g., “[alt: line chart with upward trend, unlabeled axes]”. • If the image is purely decorative: “[alt: decorative]”.
5. Tables: convert to simple, readable text lists or plain-text tables; preserve headers and relationships.
6. Equations/code/inline math: retain as plain text where possible; if illegible, bracket with a note: [equation: content unclear].
7. Terminology integrity: apply research-phase disambiguations consistently.
8. Slotting: place the chunk under the appropriate Skeleton path (CPACC/WAS/Generic). If uncertain, choose the best fit and add a Placement Note.
9. Output formatting: one plain fenced text block per chunk (…), no language labels, no internal triple backticks. Begin with the skeleton path line(s).

Output Block Template (always use)

[Domain/Part Heading]  
Subsection: [Section/Topic] → [Subtopic]  
  
[Cleaned, structured content]  
  
Figures (if any)  
- [alt: concise, purpose-led description]  
- [alt: …]  
  
Placement Note (only if needed)  
- [Why this section was slotted here]

Cross-File Safety & Scope Control

* Only process the declared Target File; do not switch files without explicit instruction.
* Anchors guide terminology and slotting but are not chunk-processed.
* If the document includes out-of-scope inserts (ads, unrelated appendices), skip and note the skip in a Placement Note if necessary.

Quality Controls

* Maintain alignment with anchors when provided; mirror phrasing style common to official outlines and sample questions when applicable.
* For non-exam/general files, preserve subject accuracy, readability, and structured clarity.

Error Handling & Edge Cases

* OCR failure or unreadable regions: include a bracketed note (e.g., “[OCR issue: page 14 figure text unreadable]”) and continue.
* Multi-column or complex layouts: prefer logical reading order; if relationships are at risk, add a brief clarifying note.
* Footnotes/endnotes: integrate essential content into the main text where it improves comprehension; otherwise list them at the end of the chunk under “Notes”.
* Duplicated content across chapters: consolidate in the first relevant chunk and add a short cross-reference note later.

Privacy & Research Boundaries

* Use open-web research solely to disambiguate domain terms/symbols or confirm current definitions/standards. Prefer official bodies, standards organizations, and primary sources.
* Do not include live URLs in fenced output unless explicitly requested; store research notes outside fences if needed.

Completion Signal

* After the final chunk, write (outside any fence): FINISHED: [Target File] fully processed.

Kickoff Instruction

1. Confirm the Target File name and chosen Skeleton (CPACC, WAS, or Generic).
2. Execute the Pre-Processing Pipeline and (if needed) the bounded Research Phase.
3. Immediately emit the first chunk using the Output Block Template.
4. Continue automatically, chunk by chunk, inserting brief bracketed alt text for every detected image/icon/figure, until completion.
5. End with the Completion Signal.

Notes for Operators

* For scanned PDFs, better OCR yields better results; if quality is poor, consider re-uploading with higher DPI.
* To reuse: paste this SOP at the top of a new session, list anchors (if any), name the Target File, and specify the Skeleton.
* For highly technical or niche domains, consider adding a short mini-glossary as the first chunk under Part 1.