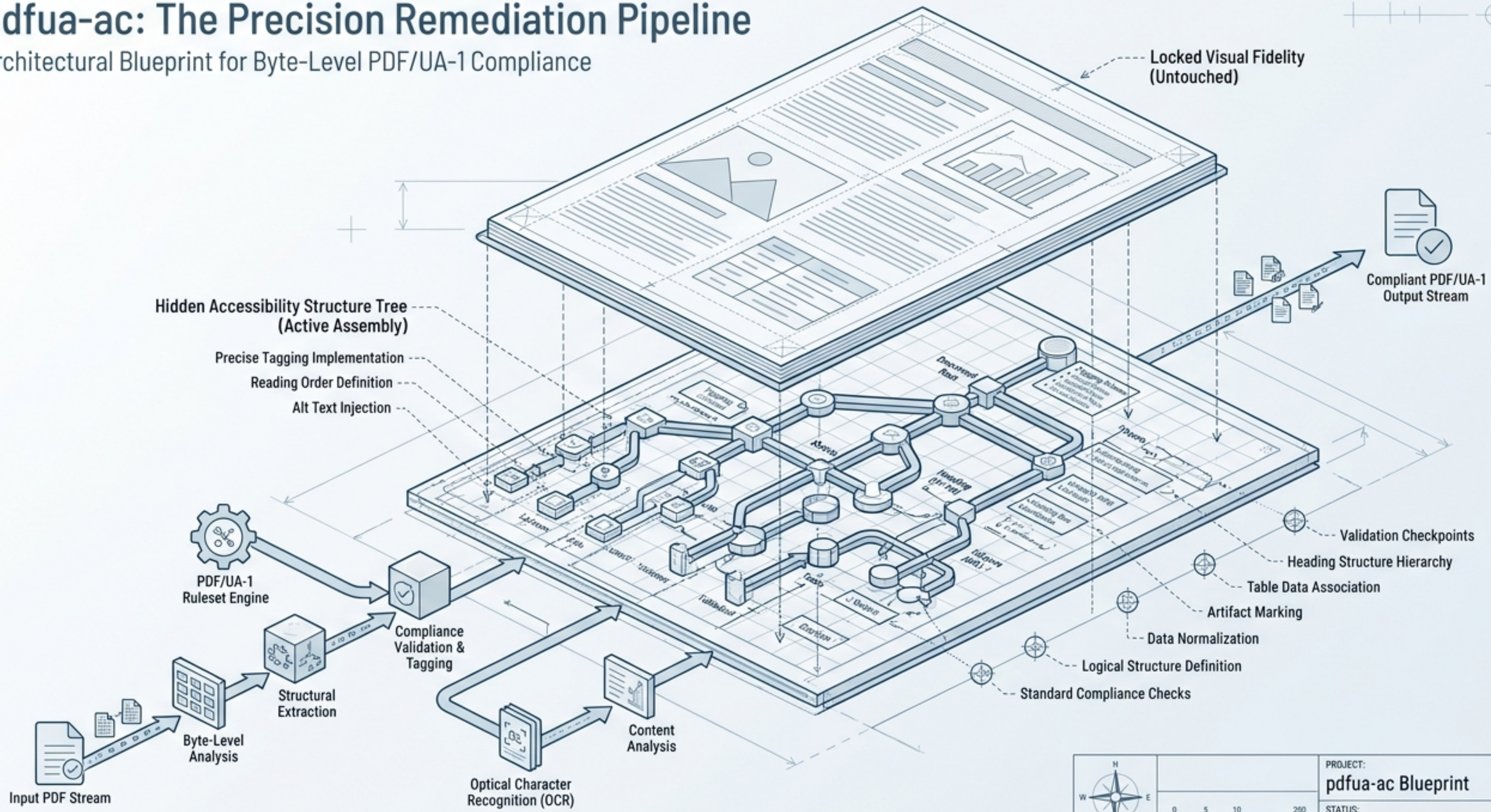
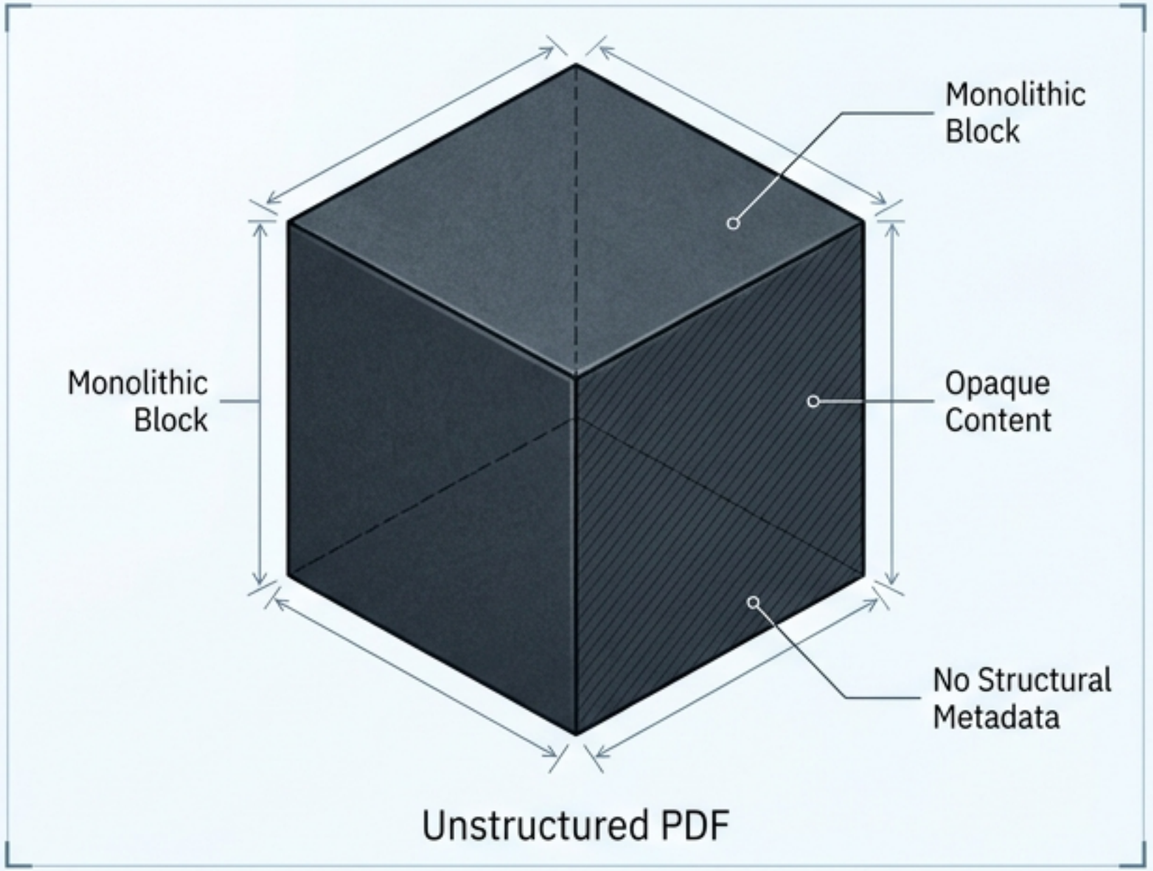
pdfua-ac: The Precision Remediation Pipeline

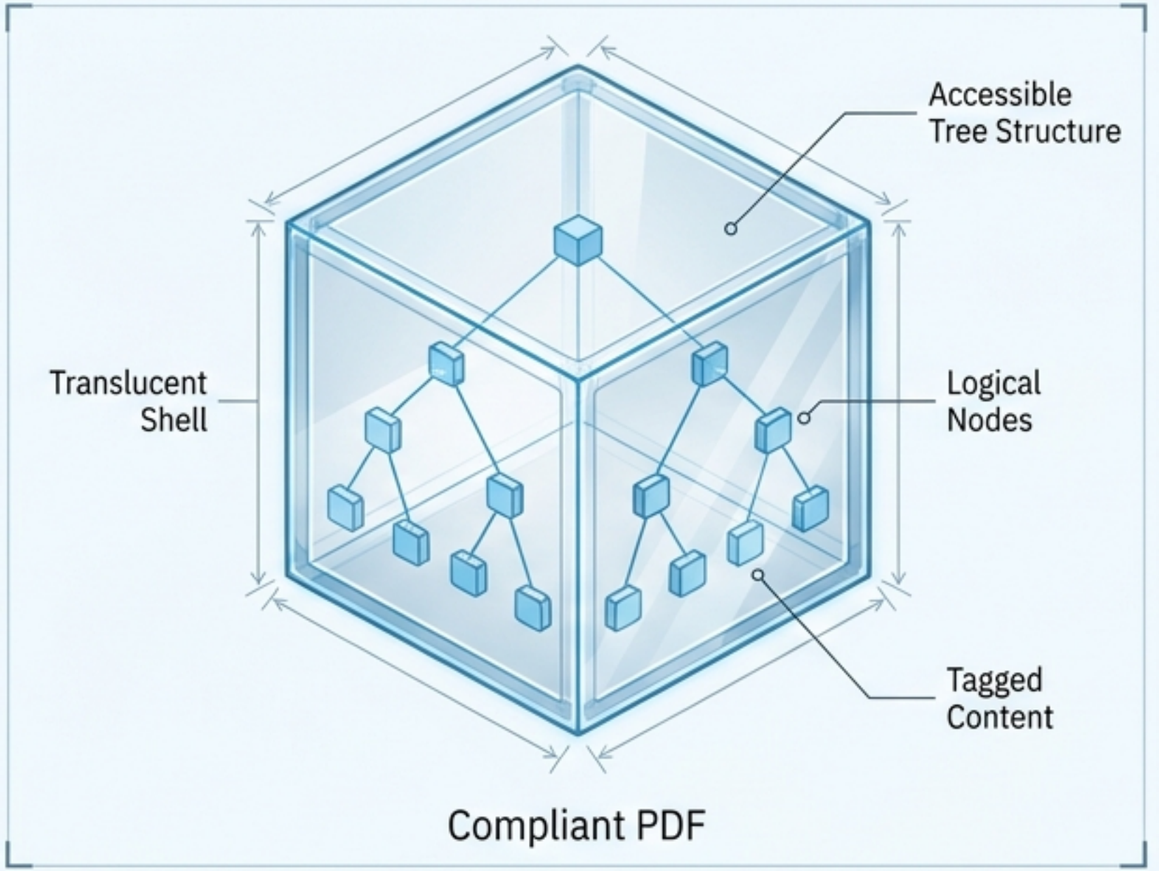
Architectural Blueprint for Byte-Level PDF/UA-1 Compliance



Retrofit, Don’t Regenerate



Unstructured PDF



Compliant PDF

Visual Fidelity is Locked

The pipeline never regenerates the document from scratch. The original appearance remains entirely untouched.

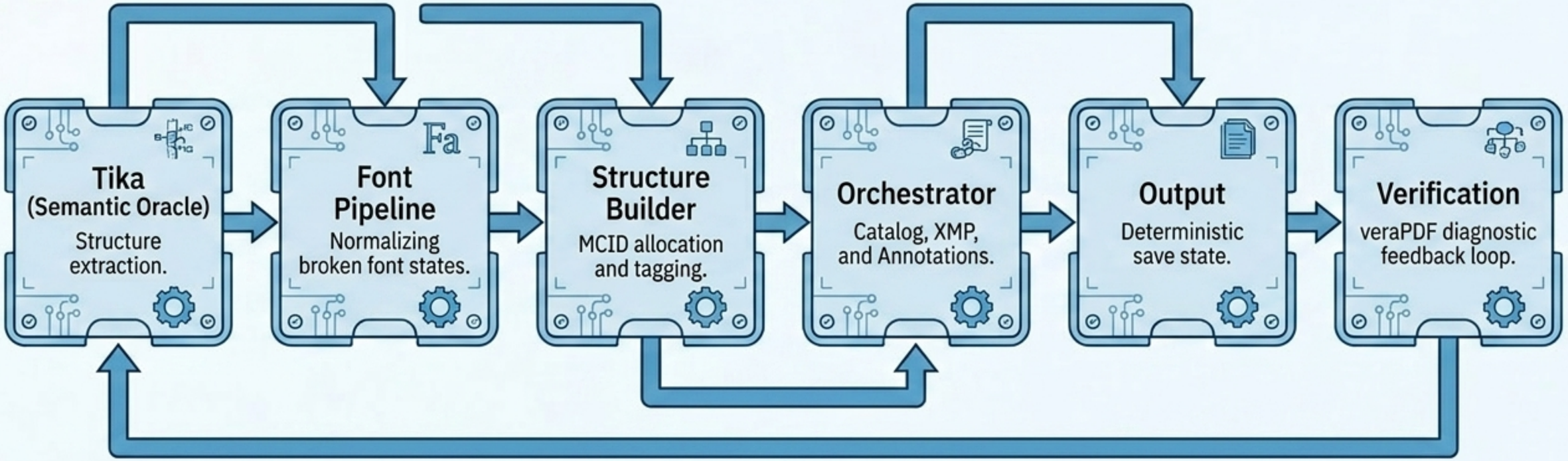
Byte-Level Object Graph Manipulation

Builds the missing accessibility tree by surgically altering the underlying PDF object graph.

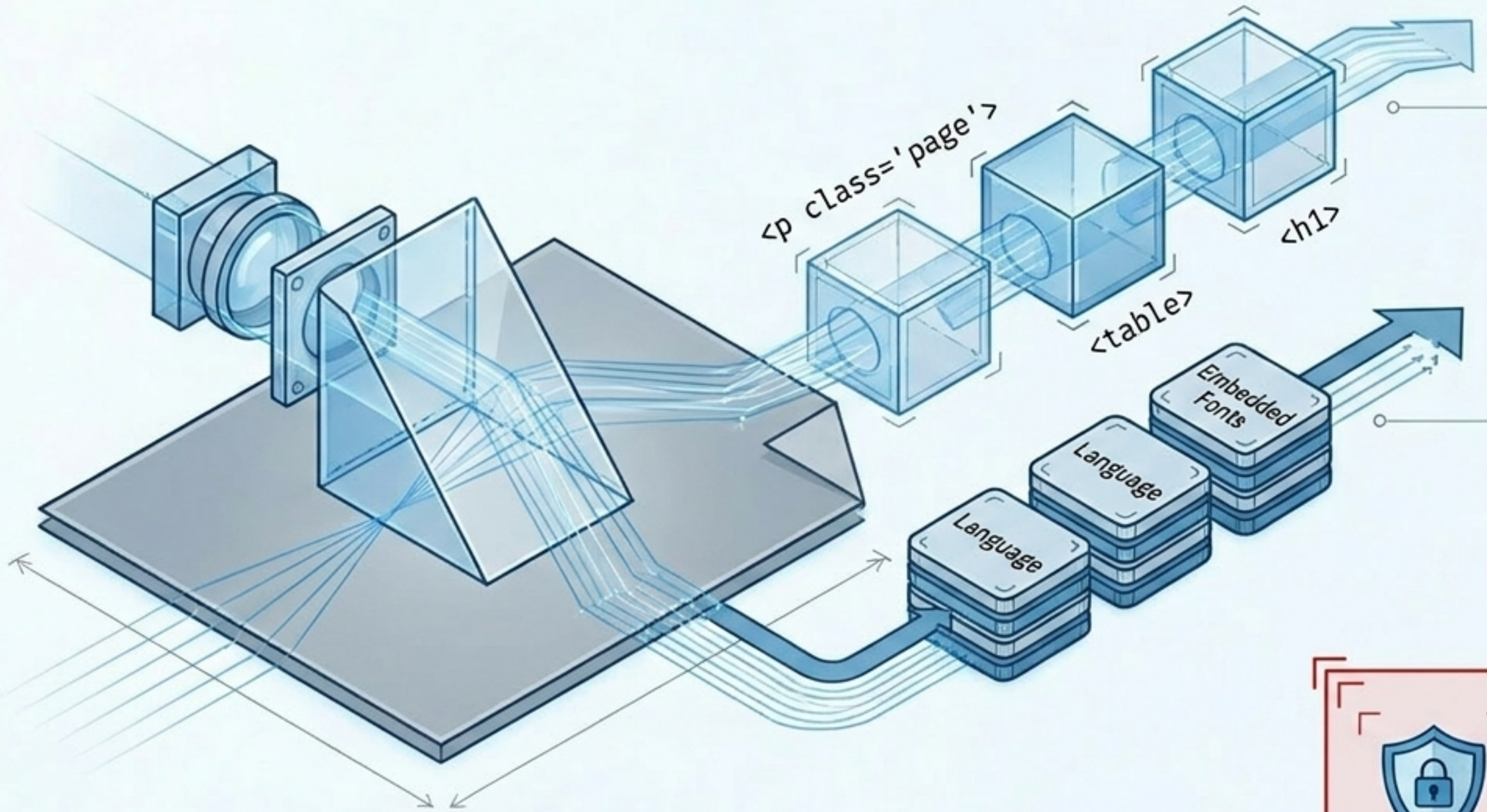
Deterministic Output

Every run produces the exact same object IDs and structure, guaranteeing predictable PDF/UA-1 compliance.

The Six-Stage Remediation Pipeline



Stage 1: Tika as the Semantic Oracle



Core Mechanism

Determines structural roles via HTTP client to local Apache Tika server (localhost:9998).

Outputs

`extract_structure(pdf)` yields XHTML with class hints. `extract_metadata(pdf)` yields JSON.

The B48 Defense:

Retry and exponential backoff mechanism built to handle Tika's intermittent Out-Of-Memory failures on massive PDFs.

Stage 2: The Font Repair Diagnostic Matrix

PDFs ship with fonts in numerous half-broken states, triggering different veraPDF failures. The `fonts.py` pipeline normalizes them before tagging begins.

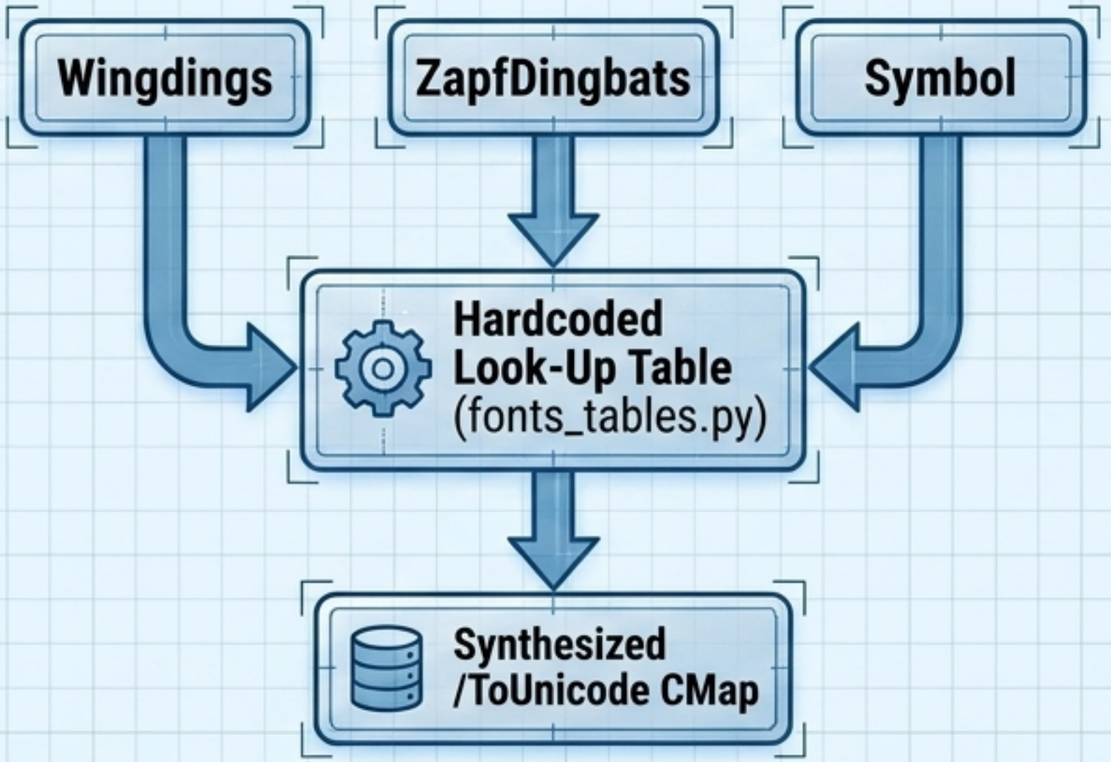
Symptom	Remediation Function	Output State
Broken CID-to-GID mapping	`fix_cidtogidmap()`	Valid CID mapping
Missing/incorrect CIDSets (subsetted fonts)	`fix_cidset()`	Rebuilt CIDSets
Unreadable text by screen readers	`fix_tounicode()` / `augment_tounicode_coverage()`	Valid CMaps
Invalid .notdef glyph references	`fix_notdef_glyphs()` / `strip_notdef_operators()`	Clean glyph references

Stage 2: Format Coverage & CMap Synthesis

Format Handling Matrix

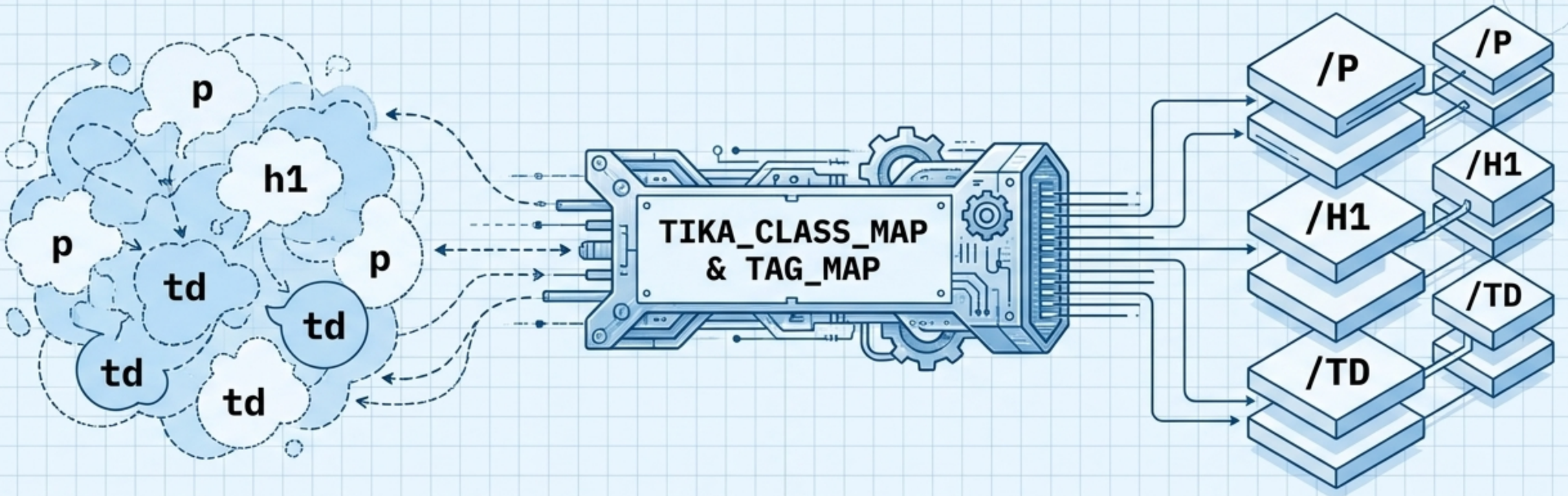
- `.ttf` → Maps to /FontFile2
- `.otf` → Maps to /FontFile3
- `.ttc` → Requires face extraction

The Symbol Font Synthesizer



Symbol fonts lack inherent text mapping. The pipeline invents a /ToUnicode CMap from glyph-name conventions.

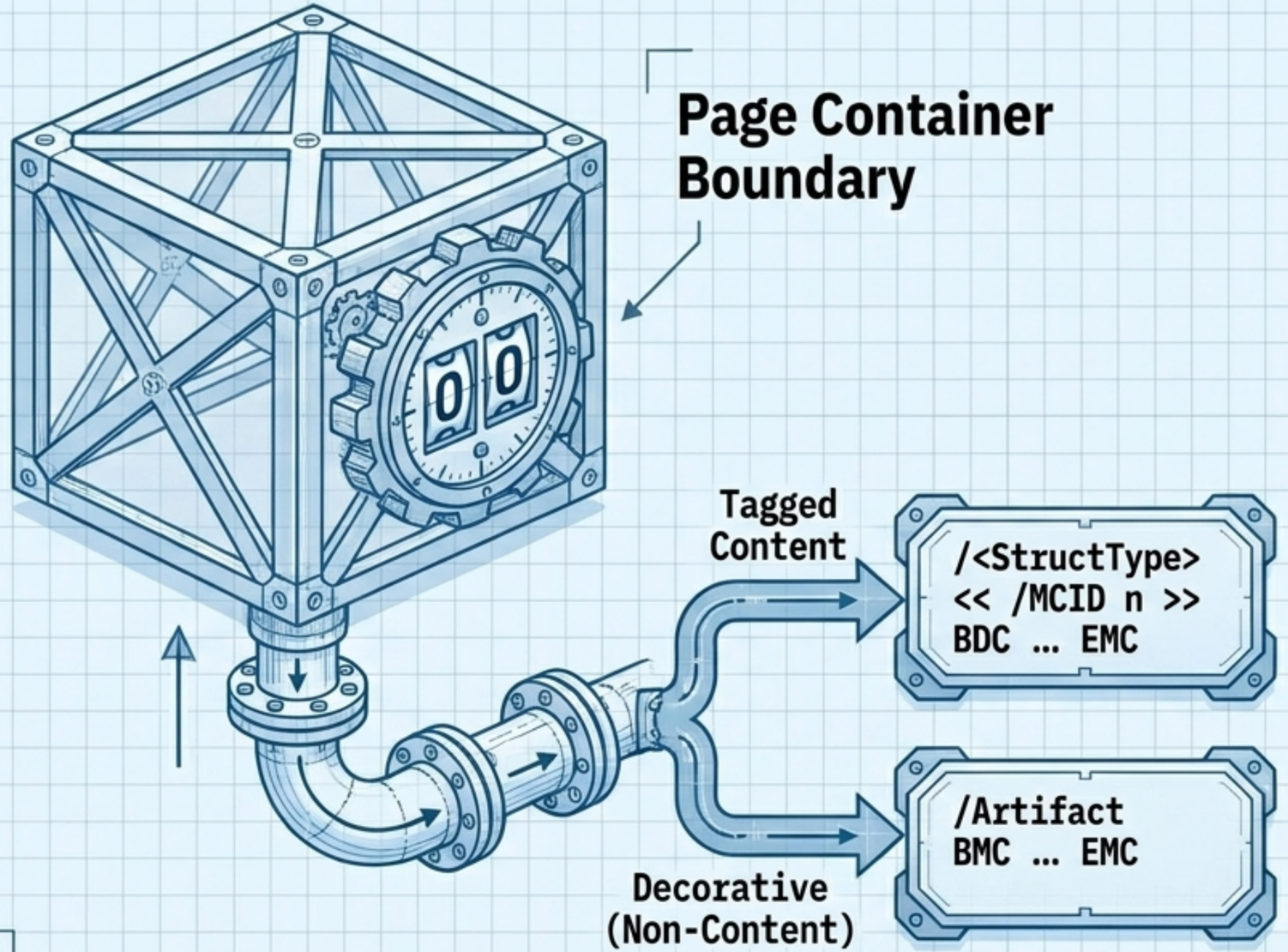
Stage 3: Structure Builder & Tag Translation



Strict PDF/UA Enforcement Rules

- **Heading Clamping (B5 / B23):** Forces the first heading to /H1 and clamps level jumps to previous + 1 (PDF/UA forbids skipping levels).
- **List Structure (B8):** Lbl nodes deliberately receive NO MCID; labels are part of the marker, not the content stream.
- **Orphan Cells (B64):** Bare TH/TD/TR without a table ancestor are wrapped in synthetic Table/TR nodes.

Stage 3: The MCID Paging Model & Stream Rewriting



Page-Local Boundaries (B4)

MCIDs are strictly page-local.

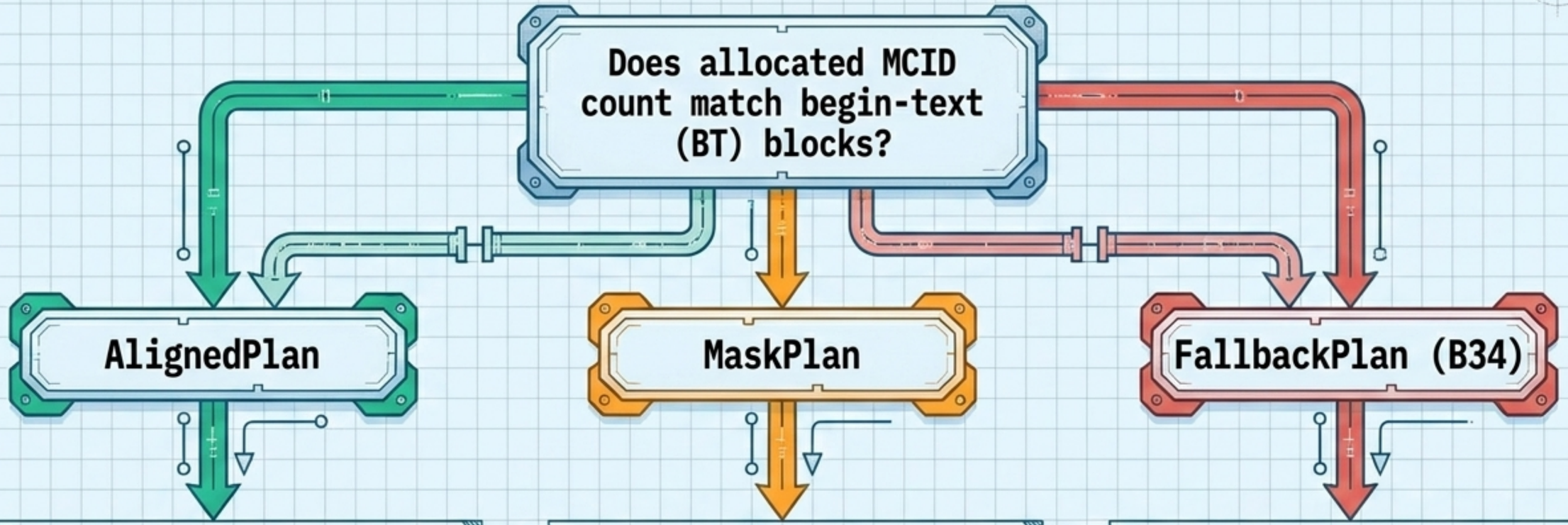
Each page restarts at MCID **0**. The **/ParentTree** coordinates them via an index of (page, mcid).

Form XObject MCIDs are explicitly stripped (**B6**).

Content Stream Rewriting

rewrite_content_streams() walks every page's content stream and forces every drawing operator into one of the two wrappers shown.

Stage 3: The MCID Reconciliation Strategy



Clean 1:1 map. Retains high granularity. The normal expected case.

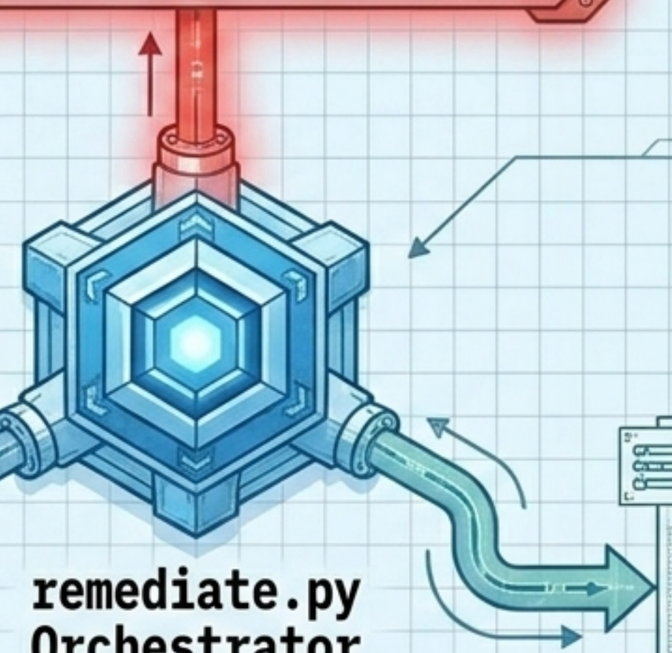
Partial alignment. Maps what it can, masks the gaps as **/Artifact**. Medium granularity retained.

Total misalignment. Bails out and wraps entire page's content as **/Artifact**. Granular tagging is lost, but structural validation is preserved.

Stage 4: Orchestration & Wiring (remediate.py)

The Scorched-Earth Rule

strip_existing_tags() runs first. Completely wipes any pre-existing **/StructTreeRoot** and markers. Builds from scratch, never patches.



remediate.py Orchestrator

Catalog & XMP Wiring

Sets **/ViewerPreferences**, writes **/Lang** from Tika, and writes full XMP packet (pdfuaid:part = 1).

Annotation Wiring

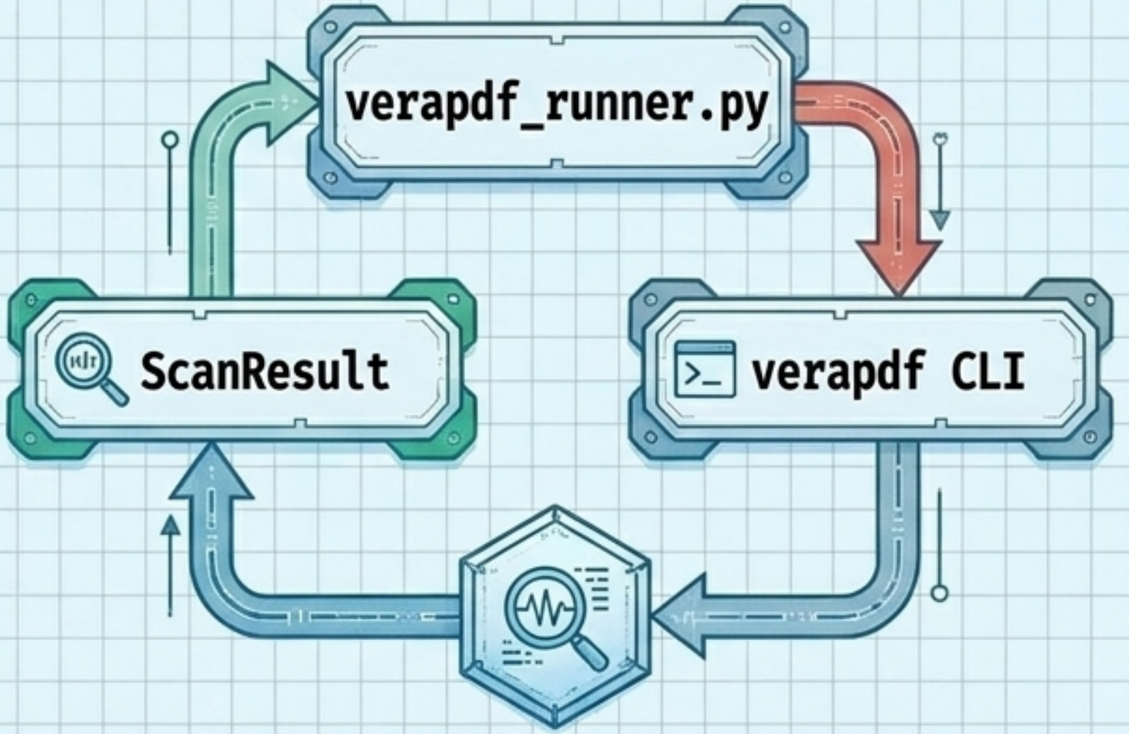
Link annotations get **/StructParent**. Widgets get **/TU**. **B2 Fix** deletes forbidden **/TrapNet** annotations. **B3 Fix** rebuilds **/ParentTree**.

Stages 5 & 6: Deterministic Output & Verification

Stage 5 Output State (pikepdf.save)

- Saved with **linearize=False** for deterministic object IDs.
- Contains a full **/StructTreeRoot**.
- Contains complete **XMP** and **Catalog** metadata.
- Embedded fonts feature valid **/ToUnicode** mappings.
- 100% of visible operators are tagged or marked **/Artifact**.

Stage 6 Verification Loop



Subprocess wrapper returning passed/failed rules. Acts as both a pre-flight diagnostic gate and a post-flight confirmation check.

The Empirical Backbone: The B-Comment Scheme

// B34: FallbackPlan

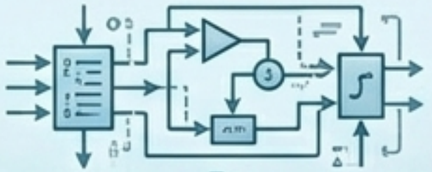
The Trigger

A specific broken document.

The Rule

The veraPDF rule failure that flagged it.

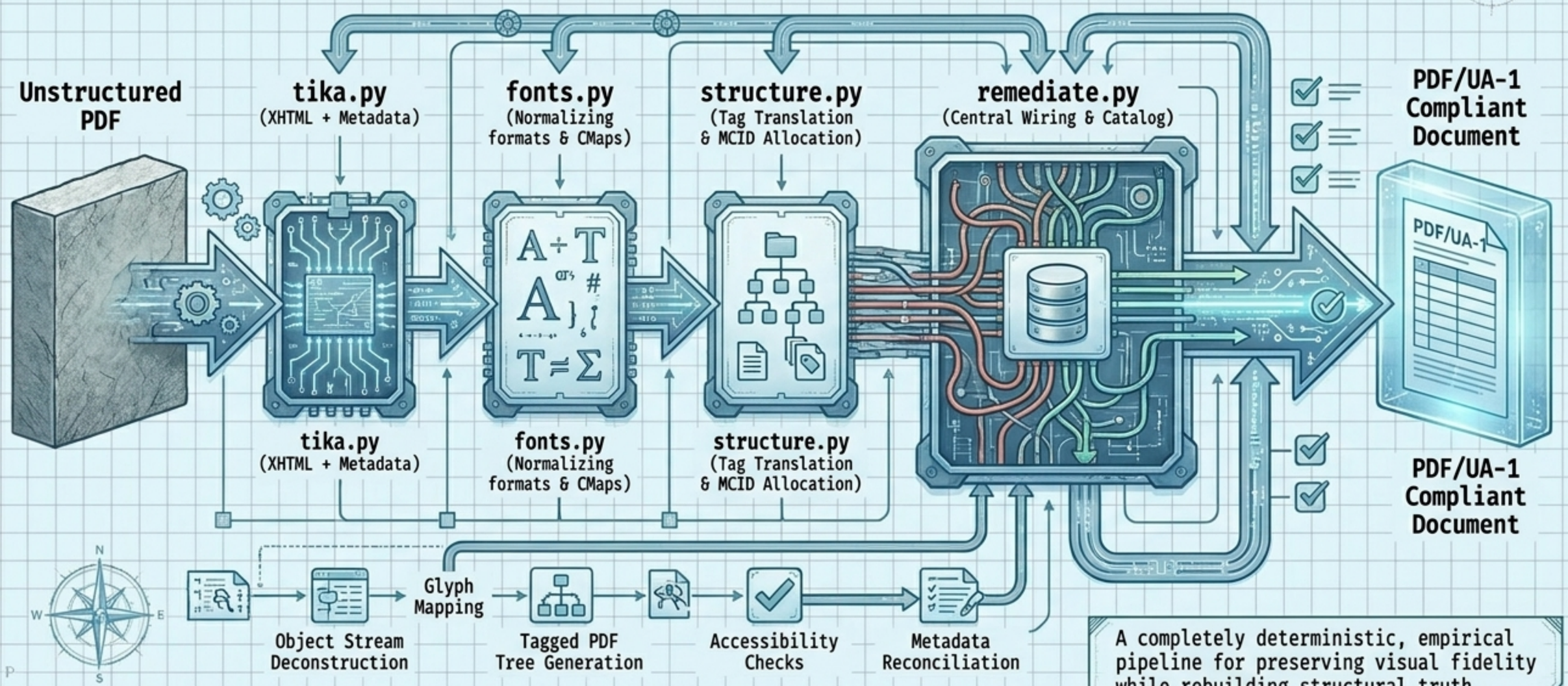
The Patch



The surgical code logic applied to defend against it.

Every fix in the codebase (B1–B66+) is reverse-engineered from a real-world document failure. Deleting the code re-breaks a real document.

The Complete pdfua-ac Blueprint



A completely deterministic, empirical pipeline for preserving visual fidelity while rebuilding structural truth.

PRECISION_BLUEPRINT_V2.0 | 2024.10.27