

# Same bytes, same meaning: how WDF makes documents verifiable for humans and machines

*Every WDF package carries two views of one document — a rendered page for people and a Markdown layer for machines — and a proof that they cannot drift apart.*

## The problem: documents that lie to machines

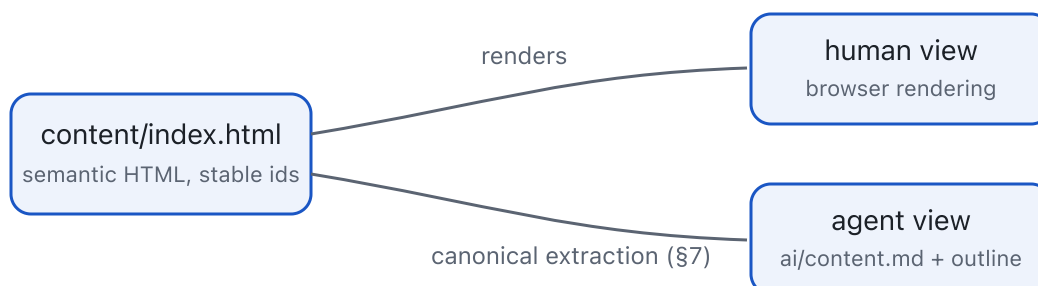
An entire industry exists to reconstruct structure from PDFs before feeding them to language models. The results are probabilistic: headings guessed from font sizes, tables re-assembled from x/y coordinates, citations pointing at page numbers rather than statements. As *the PDF Association* itself puts it, the format is “visually focused, not AI-friendly”.

The failure is structural. When the machine-readable view is *derived by heuristics*, nobody can say whether it matches what the human saw — and when money or law is involved, “probably the same” is not an answer.

## The WDF approach

WDF inverts the relationship. The package contains constrained semantic HTML as the single source of truth, and a **canonical extraction algorithm** — normative, step by step — derives the machine view from it. Determinism is the contract:

Same input bytes, same output bytes, on every platform. A validator re-runs the extraction and rejects the package on any difference.



One source, two synchronized views. Tampering with either side is detectable.

Every citable element carries a stable identifier, so an agent can answer with a citation such as `wdf:urn:uuid:...#tbl-spesa-2025` that anyone can resolve and verify — down to the paragraph, H<sub>2</sub>O-molecule precision instead of page<sup>2</sup>-level guesswork.

## What verification looks like

The reference implementation exposes the whole check as one call:

```
import { readPackage, verifyPackage } from '@wdf/core';

const pkg = readPackage(bytes);
const result = await verifyPackage(pkg);
// result.integrity    - every file matches its SHA-256 digest
// result.determinism  - ai/ is the canonical extraction of content/
// result.verified     - both, or the package is rejected
```

Note what this rules out: an attacker who edits the Markdown and diligently recomputes every hash still fails, because the *determinism* check re-derives the agent view from the human view and compares byte by byte.

## Small glossary

### canonical extraction

The normative algorithm (spec §7) mapping WDF-HTML to Markdown and an outline, with no locale-, time- or platform-dependent behavior.

### citable element

A section, heading, paragraph, table, figure, blockquote or top-level list item, carrying a mandatory stable `id`.

### standalone profile

A single HTML file embedding viewer and package: double-click to open, offline, and the canonical bytes remain extractable.

The specification and tooling are open: [repository](#) and WDF Core 0.1 draft. Feedback is welcome at [info@example.org](mailto:info@example.org).