Document Verification Scoring System

Siddartha

Last Updated: 2025-11-28 | Version 2.1

Abstract

This document formalises the multi-factor scoring framework used in the Companies House document verification pipeline. It gives precise formulas, normalization rules, interpretation guidance and decision thresholds used by the automated verification engine. The goal is to provide a clear, auditable and reproducible description suitable for engineers, auditors and reviewers.

Contents

1. Overview

The verification system computes a composite score by combining five positive components and subtracting a forensic penalty. The unbounded maximum is 130 points but the final score is capped to the $[0, 100]$ range.

Total Possible Score: 130 points (capped at 100).
Score Formula:

$$\text{Final Score} = S_{ocr} + S_{registry} + S_{provided} + S_{ocr\_match} - P_{forensic}$$

2. Score Components

2.1. OCR Confidence Score (0–30)

Purpose Measures OCR extraction quality using AWS Textract per-document confidence.

Computation If C_{ocr} is the Textract confidence expressed as a percentage (0–100), then:

$$S_{ocr} = \frac{C_{ocr}}{100} \times 30$$

Interpretation

- 25–30: High confidence (clear scan)
- 15–24: Medium confidence (minor issues)
- 0–14: Low confidence (poor scan / failures)

2.2. Registry Score (0–40)

Purpose Validates that the company number from OCR corresponds to Companies House registry. This is the most critical component.

Normalization Company numbers are normalised to an 8-character canonical form by left-padding digits with zeros and preserving any alphabetic prefix (e.g., SC, OC). Examples:

- “640918” → “00640918”
- “3035678” → “03035678”
- “SC555555” → “SC555555” (preserve prefix)

Computation Let r denote the fuzzy similarity between the normalized OCR company number and the Companies House number (range $[0, 1]$). Then:

$$S_{registry} = 40 \times r$$

In practice, exact matches yield $r = 1$ and $S_{registry} = 40$. Partial numeric similarities are scored proportionally; no match gives 0.

2.3. OCR Data Match Score (0–30)

Purpose Measures how closely OCR-extracted fields (company name, number, address) match Companies House records. This is an accuracy score distinct from raw OCR confidence.

Weights and algorithm Three fields are compared with the following weights: Name 50%, Number 30%, Address 20%. String similarity is computed using Python's `difflib.SequenceMatcher` producing ratios in $[0, 1]$.

$$S_{ocr\_match} = 30 \times \left(0.5 \cdot s_{name} + 0.3 \cdot s_{num} + 0.2 \cdot s_{addr}\right)$$

Strict name validation If $s_{name} < 0.98$ a penalty or scaling factor is applied to ensure the name component is enforced strictly. Typical behavior:

- $s_{name} \geq 0.98$: no penalty (full name credit)
- $0.90 \leq s_{name} < 0.98$: mild penalty applied via an interpolation factor
- $s_{name} < 0.90$: heavier penalties to reflect likely OCR error or mismatch

Notes Number similarity is computed after normalization (same normalization as Registry Score). Address similarity is more lenient because addresses can change over time.

2.4. Provided Score (0–30)

Purpose Measures similarity between merchant-provided data and Companies House records (if merchant data is supplied).

Weights Name 40%, Number 40%, Address 20%.

$$S_{provided} = 30 \times \left(0.4 \cdot p_{name} + 0.4 \cdot p_{num} + 0.2 \cdot p_{addr}\right)$$

where each p_* is a similarity ratio in $[0, 1]$.

2.5. Data Match Score (Informational: 0–100%)

Purpose Aggregate percentage that summarises overall similarity across OCR and merchant-provided fields. This is informational only and not used directly in the final scoring calculation.

Computation Average of available similarity ratios (six comparisons when both OCR and merchant data exist):

$$\text{DataMatch\%} = 100 \times \frac{\sum similarity\_ratios}{N}$$

2.6. Forensic Penalty (0–15, subtracted)

Purpose Deducts points for signs of tampering or manipulation detected by image-forensics.

Checks included

1. EXIF metadata analysis for suspicious or missing metadata
2. Error Level Analysis (ELA) to detect localised recompression
3. JPEG/Compression consistency checks
4. Copy-move / cloning detection

Penalty is an integer or float in $[0, 15]$ and is subtracted from the positive score sum. Higher penalties correspond to more severe forensic findings.

3. Final Score and Decision Logic

Computation The engine computes the raw sum and then clamps it to the legal range:

$$\text{Raw} = S_{ocr} + S_{registry} + S_{provided} + S_{ocr\_match} - P_{forensic}$$

$$\text{Final Score} = \min\left(100, \max(0, \text{Raw})\right)$$

Decision thresholds

Final Score	Decision	Meaning
≥ 75	PASS	Document is authentic and verified
50–74	REVIEW	Requires manual review (discrepancies)
< 50	FAIL	Verification failed (likely fraudulent or incorrect)

4. Worked Examples

Example 1 — Perfect Match (PASS)
OCR Confidence: 97% ⇒ $S_{ocr} = 29.1$
Registry Score: 40.0 (exact number match)
OCR Data Match: 28.5
Provided Score: 0 (no merchant data)
Forensic Penalty: 0.0

$$\text{Raw} = 29.1 + 40 + 28.5 - 0 = 97.6 \Rightarrow \text{Final} = 97.6\ (PASS)$$

Example 2 — Needs Review
OCR Confidence: 80% ⇒ $S_{ocr} = 24.0$
Registry Score: 0.0 (number mismatch)
OCR Data Match: 15.0
Provided Score: 0.0
Forensic Penalty: 2.0

$$\text{Raw} = 24 + 0 + 15 - 2 = 37 \Rightarrow \text{Final} = 37\ (REVIEW)$$

5. Implementation Notes

Similarity algorithm The system uses `difflib.SequenceMatcher` from Python's standard library to produce similarity ratios; this is deterministic and performs well for short strings. For production, I will be considering augmenting with domain-specific token normalization (strip punctuation, expand abbreviations like "Co." to "Company", etc.) and optionally using Jaro-Winkler for names.

Company number handling Always normalise company numbers before comparison. Preserve alpha prefixes. Ensure leading-zero padding is applied for numeric-only numbers to 8 characters.

Forensic tooling I will be using a pipeline of forensic detectors. Store the raw forensic signals (e.g., ELA heatmap metrics, EXIF review table) in the audit record to support human review.

6. Best Practices

- Use high-resolution scans (300 DPI recommended) and avoid lossy edits prior to upload.
- Prompt merchants to provide company data during the upload flow to increase score coverage.
- Surface the key reasons when a document is in `REVIEW` status (e.g., number mismatch, low name similarity, forensic penalty).

7. API Response Schema

```
{
  "ocr_score": 29.1,
  "registry_score": 40.0,
  "ocr_comparison_score": 28.5,
  "provided_score": 0.0,
  "data_match_score": 85.3,
  "final_score": 97.6,
  "decision": "PASS",
  "forensic_penalty": 0.0
}
```

8. Change Log

- 2.1 (2025-11-28) — Clarified distinction between OCR Confidence Score and OCR Data Match Score; tightened name similarity thresholds.
- 2.0 (2025-07-01) — Initial public release.