COMPUTER FORENSICS LAB

London digital forensics & e-discovery · Established 2007

OCR: When Searchable Does Not Mean Accurate

Scanned Documents, Recognition Errors and the Text Layer Your Searches Actually Run Against

Every disclosure exercise contains documents that are pictures of words: scans, faxes, photographed pages, image-only PDFs. Optical character recognition gives them a text layer so searches can reach them: and that text layer is a machine's best guess, not the document. Recognition misreads characters, mangles tables, skips handwriting, inverts columns and silently produces gibberish from poor scans: while the review platform reports the document as "searchable" with a straight face. The gap between searchable and accurate decides cases: the guarantee that "was never found" because its reference number OCRed with a zero for an O; the null result certified over an archive whose text layer was noise. This guide explains OCR for UK lawyers: how recognition works and fails, quality measurement and triage, the searching consequences, handwriting and legacy formats, and the protocol language that keeps text quality honest.

⏱ Estimated reading time: 24 min read

§ ABOUT THE AUTHOR

PREPARED BY COMPUTER FORENSICS LAB E-DISCOVERY TEAM

Computer Forensics Lab is a London-based digital forensics and electronic disclosure practice established in 2007. Its eDiscovery arm, **e-discovery.uk**, runs OCR as a measured processing step: engine selection, quality scoring, re-OCR triage and the posture declarations that keep search certificates honest: while the forensic side handles the hard cases: degraded scans, handwriting, and disputes about what a produced text layer did to the underlying document.

ESTABLISHED 2007 · LONDON | ISO 17025-ALIGNED PROCEDURES | OCR & TEXT-QUALITY ENGINEERING | SEARCH VALIDATION | CPR PART 35 EXPERT REPORTS | FULL CHAIN-OF-CUSTODY DOCUMENTATION

§ CONTENTS

In this guide

§ 0 1 · ORIENTATION

Executive summary

THE HEADLINE POINT: OCR TEXT IS AN ESTIMATE: MEASURE IT, TRIAGE IT, AND NEVER CERTIFY A SEARCH OVER TEXT NOBODY SCORED

OCR converts document images into searchable text by recognising character shapes: and its output quality is a spectrum, from near-perfect (clean laser-printed originals) to useless (third-generation faxes, photographed pages, carbon copies), with the review platform treating every point on that spectrum identically as "searchable". Four disciplines manage the gap. **Posture:** the corpus's image-only fraction is known, OCR is applied deliberately, and the fraction and method are stated: guide 84's protocols and guide 85's null diagnoses both depend on this line existing. **Measurement:** text layers are scored (engine confidence, dictionary-word rates, character statistics) so quality is a number per document, not an assumption per corpus. **Triage:** low-scoring populations route to re-OCR with better engines and settings, image enhancement, or: where recognition cannot succeed: human reading and the §8 alternatives, with the residual unsearchable population listed like guide 81's exceptions. **Search adaptation:** terms running against OCR text compensate (fuzzy matching, character-confusion variants, identifier patterns), and critical searches over critical archives are validated by sampling the images, not the text. The one-line rule: a null result over unmeasured OCR is not a result.

- **Know the fraction:** how much of the corpus is image-derived text, and how much has no text at all: the posture line every certificate needs.
- **Score, don't assume:** per-document quality metrics turn "searchable" into a measured property.
- **Triage the tail:** re-OCR, enhancement, or honest listing as unsearchable: never silent inclusion as noise.
- **Adapt the searches:** fuzzy operators and confusion variants for OCR populations; image-side sampling for the critical ones.
- **Declare it all:** engine, settings, fraction, residuals: in the protocol, per guide 84 §11's wording.

§ 0 2 · FIRST PRINCIPLES

The problem in plain English: pictures of words

A scanned contract is a photograph. It contains no words a computer can read: only pixels arranged in word-shaped patterns: and every search in every platform runs against text, not pictures. OCR bridges the gap by guessing which characters the pixels show, and on a clean scan of typed text the guess is excellent. On the documents litigation actually cares about: the 1998 fax of the guarantee, the photocopied-thrice board minutes, the phone photo of the whiteboard: the guess degrades, sometimes to static, and nothing in the review interface says so. The document opens beautifully (you are looking at the image); the searches fail silently (they are reading the guess).

The trap is the invisibility. A corrupted file announces itself as an exception (guide 81); a bad text layer announces nothing: it just stops matching. Whole archives can sit in a corpus, technically searchable, functionally invisible: and the party certifying "no documents found" over them is honest, diligent and wrong. Guide 84 planted the flag: the OCR posture is part of every search methodology: and guide 85's null diagnosis routinely lands here. This guide is what to actually do about it.

§ 0 3 · THE MACHINERY

How OCR works and how it fails

- **The pipeline:** image cleanup (deskew, despeckle, contrast), layout analysis (columns, tables, reading order), character recognition (pattern models and language dictionaries), and output assembly: modern engines, including the neural generation, are dramatically better than their ancestors and still bounded by the same inputs.
- **Character confusions:** the classic substitutions: 0/O, 1/l/I, 5/S, 8/B, rn/m, cl/d: harmless in prose, lethal in identifiers: invoice numbers, account codes, references: exactly the strings litigation searches for verbatim.
- **Layout failures:** tables flattened into word soup, columns read across instead of down, stamps and marginalia interleaved mid-sentence: the text exists but the sequence lies, defeating phrase and proximity searches (guide 84 §4's workhorses).
- **Input-quality dependence:** resolution, generation loss (copies of copies), skew, bleed-through, coloured backgrounds and dot-matrix faces each tax recognition: quality of output is mostly decided before the engine runs.
- **Language and era effects:** engines assume languages and eras: mixed-language pages, archaic typefaces and non-Latin scripts need configuration, not defaults: the guide 84 language census reaching the text layer.

§ 0 4 · THE MEASUREMENT

Measuring text quality: scoring, triage and re-OCR

- **Engine confidence:** recognisers emit per-character and per-page confidence: captured, not discarded: the first quality signal, free at processing time.
- **Statistical scoring:** dictionary-word rates, character-distribution sanity, word-length profiles and symbol density: computable over every text layer, flagging the gibberish populations without reading a page.
- **The quality map:** scores aggregated by source, custodian and era: the corpus's text quality as a picture: the 2015-onward estate excellent, the scanned legacy archive red: driving effort where it matters.
- **Re-OCR triage:** the red populations re-run with better engines, tuned settings, and image enhancement: the same page often recovering from noise to usable in one modern pass: with before/after scores recording the gain.
- **The honest residual:** pages recognition cannot rescue: listed, counted and disclosed as functionally unsearchable, per guide 81's exception discipline: reviewed by eye where materiality warrants, never left masquerading as searched text.

§ 0 5 · THE CONSEQUENCES

The searching consequences: terms against imperfect text

- **Exact match is brittle:** one confused character defeats a verbatim term: identifier searches over OCR populations enumerate confusion variants (the reference with 0-for-O, 1-for-l) or use pattern searching where the platform supports it.
- **Fuzzy operators, deliberately:** edit-distance matching widens reach over noisy text and buys noise doing it: applied to OCR populations knowingly, with precision sampled per guide 85 §4 rather than assumed either way.
- **Phrase and proximity degrade with layout:** where tables and columns scrambled sequence, connector searches under-perform: material questions against layout-heavy documents (contracts schedules, ledgers) validate against the images.
- **The dual-validation rule:** for case-critical searches over OCR archives, sample the un-hit images by eye (guide 85's elusion logic, applied at the text-quality boundary): the only direct test that the text layer is not eating your documents.
- **Null diagnosis, completed:** guide 85 §3's null list resolves here for image-heavy estates: the zero-hit term checked first against the quality map before anyone believes the zero.

§ 0 6 · THE HARD CASES

The hard cases: handwriting, faxes, stamps and legacy paper

- **Handwriting:** conventional OCR does not read it; handwriting recognition (including modern AI transcription) is improving and is treated as assisted drafting of an estimate: material manuscript evidence (attendance notes, diary entries, signatures) is read by humans and, where contested, examined forensically: transcription never substitutes for the image in evidence.
- **Faxes and thermal paper:** low resolution, banding and fade: the classic red population: enhancement plus modern engines recover much; the rest joins the honest residual with its materiality assessed.
- **Stamps, annotations and signatures:** overlays that garble surrounding text and carry their own evidential weight (dates received, approval marks): flagged populations where the image, not the text layer, is the document.
- **Carbon copies, dot matrix and pre-laser eras:** era-specific degradation met with era-aware settings: and expectations set honestly on recovery rates.
- **Photographed documents:** phone photos in message threads (the guide 63-65 estates are full of them): perspective, shadow and crop taxing recognition: routed through enhancement, scored like everything else, and remembered when messaging corpora report suspicious nulls.

§ 0 7 · THE PAPERWORK

OCR in the protocol: postures, disclosure and disputes

- **The posture declaration:** guide 84 §11's clause given teeth: image-only fraction, engine and settings, scoring approach, re-OCR policy, residual listing: stated in the DRD or protocol so both sides' searches mean something.
- **Productions and text layers:** produced documents carry their text (extracted or OCR) per the protocol: with OCR-derived text identified as such, because the receiving party's searches inherit its errors: and the receiving party entitled to know what it is searching.
- **Challenging a null over images:** the §11 questions: their fraction, their scores, their residuals: guide 84 Example 3's application, generalised: a certificate over unmeasured text quality is the challengeable kind.
- **Defending your own:** the quality map, triage records and dual-validation samples as the answer filed before the question: the guide 85 protocol file gaining an OCR chapter.
- **Contested transcriptions:** where what a poor document says is itself disputed, the question leaves the pipeline: image enhancement and document examination under CPR Part 35, per the laboratory disciplines: the text layer was only ever a finding aid.

§ 0 8 · THE WIDER MAP

Source architecture: where else the evidence lives

Per this series' source-architecture discipline: a bad scan is rarely the only form the document ever had, and OCR failure is a routing instruction to its better selves. The content may also live as: the born-digital original (the Word file behind the printed-and-scanned contract, in the DMS with its version ladder per guide 75: the single highest-value re-pull); the counterpart's copy (the same agreement as sent or received, cleanly, per the standing counterpart method); other generations of the same paper (the first photocopy where you hold the fourth; the sender's file copy of the fax); rendered and quoted forms (the schedule retyped into an email, the clause quoted in correspondence); the transactional record beside the paper (the invoice's data in the ERP per guide 74, the payment in the bank feed): often a better answer than the image ever was; and the human layer (the witness who can read the annotation, the author who kept the notebook). The unsearchable residual of §4 is triaged against this map: many of its members matter only until their better self is found.

WHEN THE SCAN DEFEATS OCR...	BORN-DIGITAL ORIGINAL	COUNTERPART COPIES	EARLIER GENERATIONS	RENDERED / QUOTED FORMS	STRUCTURED RECORDS	HUMAN READING
Contracts and agreements	Y: DMS ladders	Y: as exchanged	Execution copies	Quoted clauses	Obligations in systems	Last resort, documented
Invoices and financials	Accounting exports	Counterparty's set	n/a	Statements, schedules	Y: guide 74's rows	Rarely needed
Faxes and letters	Sometimes drafts survive	Y: sender's file copy	Y: earlier copies	Replies quoting content	n/a	Y: enhanced image
Handwritten material	n/a	Sometimes	Y: the clearest copy	Typed-up versions	n/a	Y: the primary route
Photographed pages in chat	The photographed original	Other recipients' copies	n/a	The thread discussing it	n/a	Y

Fallback strategy in one line: when recognition fails, hunt the document's better self: the born-digital original, the counterpart copy, the structured record: before spending on rescuing pixels.

§ 0 9 · IN THE WILD

Worked examples

EXAMPLE 1 · THE GUARANTEE FOUND BY ITS OWN MISREADING

A lender's claim depended on locating a 2009 guarantee referenced as GTE-1107/B across a scanned deeds archive. Verbatim searches returned nothing, and the borrower's team pressed the absence hard. The §5 discipline found it in an afternoon: the archive's quality map showed mid-grade OCR; confusion-variant searches were generated for the reference (0/O, 1/l/I, 7/T, B/8); and "GTE-11O7/8" hit twice: the guarantee and its side letter, both cleanly legible as images, both invisible to exact match for fifteen years of that archive's life. The re-OCR pass that followed scored the whole deeds population, corrected 900 documents' text layers, and the eventual disclosure certificate stated the archive's measured posture: the null that nearly was, converted into the exhibit that decided the claim.

EXAMPLE 2 · THE CERTIFICATE OVER STATIC

A defendant certified that searches across its legacy archive returned no documents concerning the disputed commission arrangements: technically true. The claimant's challenge ran §7's questions: what fraction of the archive was image-derived, what were the scores? The answers unravelled the certificate: 60,000 documents were scans of 1990s carbon copies and faxes, OCRed once in 2011 on a then-current engine, never scored: sampled by the court-appointed exercise, their text layers averaged under 40% dictionary-word rate: functionally static. Modern re-OCR recovered most of the population; the re-run terms found the commission ledger and eleven covering letters; and the costs judgment distinguished, precisely, between an honest certificate and a meaningful one. The defendant's provider now scores everything.

EXAMPLE 3 · THE HANDWRITTEN MARGIN THAT OUTWEIGHED THE TYPED PAGE

A partnership dispute turned on a typed 2014 memorandum: OCRed perfectly, searched, reviewed, relied on by both sides: whose produced image carried a handwritten marginal note nobody's searches could see: "NB: JH says split stays 60/40 regardless: agreed on call". The note surfaced only when the §6 discipline flagged annotation-bearing populations for image review; forensic examination confirmed the ink annotation was contemporaneous with the document's circulation copy, and the counterpart's file held the same page without the note: dating the annotation to the claimant partner's own copy and corroborating his account of the call. The text layer had been flawless and the document's decisive content was never in it: the standing reminder that OCR searches text, and documents are more than their text.

§ 10 · WHERE IT GOES WRONG

Common mistakes and technical limitations

Common mistakes

- **"Searchable" read as "searched":** the platform's checkbox mistaken for a quality statement: §2's invisible trap.
- **Certificates over unscored text:** Example 2's static: honest, diligent, meaningless.
- **Identifier searches run verbatim only:** Example 1's fifteen-year blindness: confusion variants never generated.
- **The residual left masquerading:** unrescuable pages neither listed nor eye-reviewed: guide 81's exception discipline stopping at the text layer's edge.
- **Annotations unsurveyed:** Example 3's margin: image-only content in "perfectly searchable" documents, unflagged.
- **One OCR pass forever:** a 2011 engine's guesses inherited by every later search: re-OCR economics never revisited.
- **Fuzzy search unsampled:** edit-distance operators deployed with precision assumed: guide 85's discipline abandoned at the noisy boundary.
- **Better selves unhunted:** budgets spent rescuing pixels while the born-digital original sat in the DMS: §8 unread.

Technical limitations

- **Recognition is bounded by the image:** information destroyed by generations of copying is not recoverable by software: enhancement helps; it does not resurrect.
- **Handwriting remains hard:** machine transcription of manuscript is an estimate for triage, not evidence: the human read governs where it matters.
- **Layout reconstruction is imperfect:** complex tables may never OCR into faithful sequence: table-critical questions work from images or re-keyed data.
- **Scores are proxies:** dictionary-rate metrics flag gibberish well and subtle errors less well: critical populations still sample against images.
- **Every engine differs:** two engines, two text layers, two search results from one archive: the engine and version belong in the protocol for exactly this reason.

§ 11 · INTERROGATORIES & DRAFTING AIDS

Questions to ask · Suggested wording

Ask your client

- Which estates are paper-derived: legacy archives, deeds, scanned files: and do born-digital originals exist for the ones that matter?
- Are there identifier-critical searches (references, account numbers) heading for OCR populations: for the §5 variant treatment?
- Which document classes carry annotations, stamps or manuscript content the text layer will never hold?

Ask your opponent

- Please state what proportion of your disclosure corpus comprises image-derived text, the OCR engine and settings applied, and what quality measurement has been undertaken.
- Please identify any populations determined to be unsearchable or of low text quality, and state how they have been addressed in your search methodology.
- Where a produced document's text was generated by OCR, please identify it as such, and produce the image at sufficient resolution for independent recognition.

Ask your eDiscovery / forensic provider

- What does the quality map show, by source and era: and where is the red?
- What would re-OCR of the red populations recover, at what cost: before/after on a sample?
- Which §8 better selves exist for the residual: and which residual members are material enough to read by eye?

SUGGESTED WORDING · OCR POSTURE PARAGRAPH FOR THE DRD / PROTOCOL

> "Each party shall identify the proportion of its corpus comprising image-only or image-derived documents and shall apply OCR to such documents prior to the application of search terms, stating the engine, version and material settings used. Text quality shall be measured by stated metrics, and populations of low quality shall be re-processed with current-generation recognition or, where recognition cannot produce usable text, identified in a schedule of functionally unsearchable documents with their treatment stated. Search methodologies applied to OCR-derived text shall account for recognition error, including in respect of identifiers and references, and search results over such populations relied upon for any certification shall be validated by sampling against document images. OCR-derived text in productions shall be identified as such."

§ 12 · QUICK CONTROL

Checklist and red flags · When to involve a digital forensic expert

The OCR checklist

- ☐ Image-only fraction known per source; posture declared
- ☐ Engine, version and settings recorded in the protocol file
- ☐ Text layers scored; quality map built by source and era
- ☐ Red populations re-OCRed with before/after measurement
- ☐ Honest residual listed, counted, and materially triaged
- ☐ Identifier searches variant-expanded over OCR populations
- ☐ Fuzzy operators sampled for precision where deployed
- ☐ Critical nulls validated against images, not text
- ☐ Annotation and manuscript populations flagged for image review
- ☐ Better selves hunted per §8 before rescue spend

Red flags

- Certificates silent on OCR posture: Example 2's shape.
- Zero hits for references known to exist in paper estates: Example 1's signal.
- Archives OCRed once, years ago, never re-scored.
- "Searchable" legacy estates with no quality numbers attached.
- Produced text layers unlabelled as OCR-derived.
- Annotation-heavy classes reviewed by text alone: Example 3's miss.
- Search-adequacy disputes where neither side has scored anything.

When to involve a digital forensic expert

At processing, where posture, scoring and triage are built into the pipeline (guide 81's stage, text-quality chapter); at search design and testing, where variant strategies and image-side validation protect the critical questions (guides 84-85's disciplines meeting this one); at the hard cases, where enhancement, handwriting examination and contested-transcription work run under CPR Part 35; and at disputes, where certificates over image estates are challenged or defended with measurements. Through **e-discovery.uk**, OCR runs as a scored, declared step: the difference between searchable and searched, kept visible by design.

§ 1 3 · COMMON QUESTIONS

Frequently asked questions

Our platform says the documents are searchable. Isn't that enough?

The flag means a text layer exists, not that it is any good: a page of recognition static is "searchable" in exactly that sense. The question that matters is the score: dictionary rates, confidence, the quality map: and platforms do not volunteer it; §4 is how you get it.

How accurate is modern OCR really?

On clean, recent, typed originals: excellent, routinely near-perfect for search purposes. On the paper litigation cares about: faxes, generations of photocopies, photographed pages: accuracy falls with the image, sometimes to noise: which is why the honest answer is always a measurement of your corpus, not an industry percentage. The spread, not the average, is the story.

Searches found nothing in our scanned archive. Can we rely on that?

Not until the archive is scored and the critical terms are variant-expanded and image-validated: Examples 1 and 2 are the two ways that reliance ends. A null over measured, remediated text with a sampled image check behind it is solid; the same null over an unscored 2011 text layer is a certificate waiting to be distinguished.

Can OCR read handwriting, signatures and stamps?

Conventional OCR, no; modern handwriting recognition produces useful estimates for triage and is treated as exactly that: material manuscript content is read by humans, examined forensically where contested, and never reduced to its machine transcription in evidence. Stamps and annotations garble nearby text and carry their own weight: flagged populations, per §6, reviewed as images.

Should we re-OCR an archive that was processed years ago?

Score it first: if the map shows red, current-generation engines routinely transform recovery on the same images, and the before/after sample prices the decision in an afternoon: Example 2's defendant would have paid for that afternoon many times over. Green populations can stand; the point is knowing which is which.

The other side produced scans whose text garbles the key clause. What do we do?

Work the ladder: demand the image at proper resolution and run independent recognition; demand identification of OCR-derived text per §7; hunt the document's better self (their DMS original, your own counterpart copy) per §8; and where what the page says is itself the dispute, take it to examination under CPR Part 35: enhancement and document analysis, not argument about a machine's guess. The text layer is a finding aid: the moment it becomes the battlefield, the image is the evidence.

§ 14 · REFERENCE

Glossary

Character confusion
Systematic misreads (0/O, rn/m); the identifier killer.

Confidence score
The engine's own per-character certainty estimate.

Dictionary-word rate
The fraction of output words that are real words; a gibberish detector.

Dual validation
Checking critical search results against images, not text.

Fuzzy search
Edit-distance matching tolerant of recognition error.

Image-only document
A document with pictures of text and no text layer.

Layout analysis
The engine's reconstruction of columns, tables and order.

Quality map
Text scores aggregated by source, custodian and era.

Re-OCR
Re-recognition with better engines or settings.

Text layer
The searchable text attached to a document image.

Sources and authoritative references

The OCR disciplines in this guide draw on materials published by our laboratory and e-discovery practice, referenced here as sources:

- e-discovery.uk, EDRM Phase 04 · Processing (OCR as a scored pipeline step; posture declaration; residual handling): e-discovery.uk/edrm/processing · Phase 05 · Review: e-discovery.uk/edrm/review
- cflab.uk, digital forensics services (image enhancement, document examination, contested transcription, CPR Part 35 reporting): cflab.uk/digital-forensics-services · guides library: cflab.uk/guides
- e-discovery.uk, practice method and Knowledge Centre: e-discovery.uk · e-discovery.uk/knowledge
- PD 57AD and the Disclosure Review Document: justice.gov.uk, PD 57AD · CPR Part 35: justice.gov.uk, Part 35
- Forensic Science Regulator, Code of Practice (questioned document and image work in criminal contexts): gov.uk, FSR Code
- ICO, UK GDPR guidance: ico.org.uk · ISO/IEC 27037: iso.org, 27037

§ HOW A SPECIALIST LABORATORY CAN ASSIST

Working with Computer Forensics Lab

OCR at **e-discovery.uk** is a measured step, not a checkbox: fractions known, engines and settings recorded, every text layer scored into the quality map, red populations re-recognised with before/after evidence, and the honest residual listed and triaged against its §8 better selves: so the searches of guides 84 and 85 run over text somebody can vouch for, and certificates state postures instead of hoping. The laboratory side takes the hard cases: enhancement of degraded images, handwriting and annotation examination (Example 3's margin), contested transcriptions under CPR Part 35: and both sides of the Example 2 dispute, scoring archives to challenge or defend what was certified over them. Conflicts checks, confidential scoping discussions and fixed-fee estimates are routine; and if a scanned archive sits under your live search protocol unscored, the quality map is this week's two-day job: before the null, not after it.

INSTRUCT THE LAB

Speak to a forensic examiner, not a salesperson.

Describe your matter in confidence and an examiner will respond the same working day. Conflicts checks and scoping calls with US counsel are routine.

NEW ENQUIRIES
+44 (0)20 7164 6971

EMAIL
info@cflab.uk

E-DISCOVERY
e-discovery.uk

Computer Forensics Lab Ltd · Euro House, 133 Ballards Lane, Finchley, London N3 1LJ, United Kingdom
cflab.uk · Customer service +44 (0)20 7164 6915 · ICO ZB395023 · NPCC · ISO 17025 aligned practice · UK Gov Digital Marketplace