

OCR When Searchable Does Not Mean Accurate

This guide addresses the accuracy of Optical Character Recognition (OCR) for scanned documents in e-discovery. It covers how OCR works, its limitations, and the impact of text quality on searches. It provides guidance on measuring text quality, handling difficult cases, and incorporating OCR considerations into legal protocols and disclosure.

e-discovery.uk, Computer Forensics Lab. Published 2026-08-30.
<https://e-discovery.uk/library/ocr-when-searchable-does-not-mean-accurate>

OCR & TEXT QUALITY · A GUIDE FOR UK LAWYERS OCR: When Searchable Does Not Mean Accurate Scanned Documents, Recognition Errors and the Text Layer Your Searches Actually Run Against COMPUTER FORENSICS LAB

§ ABOUT THE AUTHOR PREPARED BY COMPUTER FORENSICS LAB E-DISCOVERY TEAM SEARCH VALIDATION CPR PART 35 EXPERT REPORT S FULL CHAIN-OF-CUSTODY DOCUMENTATION

§ CONTENTS In this guide 01 Executive summary 02 The problem in plain English: pictures of words 03 How OCR works and how it fails 04 Measuring text quality: scoring, triage and re-OCR 05 The search in g consequences: terms against imperfect text 06 The hard cases: handwriting, faxes, stamps and legacy paper 07 OCR in the protocol: postures, disclosure and disputes 08 Source architecture: where else the evidence lives 09 Worked examples 10 Common mistakes and technical limitations 11 Questions to ask · Suggested wording 12 Checklist and red flags · When to involve a digital forensic expert 13 Frequently asked questions 14 Glossary · References · Disclaimer · How a specialist laboratory can assist

§ 01 · ORIENTATION Executive summary THE HEADLINE POINT: OCRTEXTISANESTIMATE: MEASUREIT, TRIAGEIT, AND NEVER CERTIFYASEARCHOVERTEXTNOBODYSCORED

§ 02 · FIRST PRINCIPLES The problem in plain English: pictures of words

§ 03 · THEMACHINERY How OCR works and how it fails

§ 04 · THEMEASUREMENT Measuring text quality: scoring, triage and re-OCR

§ 05 · THECONSEQUENCES The search in g consequences: terms against imperfect text

§ 06 · THE HARD CASES The hard cases: handwriting, faxes, stamps and legacy paper

§ 07 · THE PAPER WORK OCR in the protocol: postures, disclosure and disputes

§ 08 · THEWIDERMAP Source architecture: where else the evidence lives WHEN THE SCAN COUNTERPART EARLIER STRUCTURED HUMAN DEFEATS COPIES GENERATIONS RECORDS READING OCR BORN- RENDERED / DIGITAL QUOTED ORIGINAL FORMS

§ 09 · IN THE WILD Worked examples EXAMPLE1 · THEGUARANTEEFOUNDBYITSOWNMISREADING EXAMPLE2 ·

THE CERTIFICATE OVERSTATED EXAMPLE 3 ·

THE HANDWRITTEN MARGIN THAT OUTWEIGHED THE TYPED PAGE

§ 10 · WHERE IT GOES WRONG Common mistakes and technical limitations Common mistakes
Technical limitations

§ 11 · INTERROGATORIES & DRAFTING AIDS Questions to ask · Suggested wording Ask
your client Ask your opponent Ask your e Discovery / forensic provider SUGGESTED
WORDING · OCR POSTURE PARAGRAPH FOR THE DRD / PROTOCOL

§ 12 · QUICK CONTROL Checklist and red flags · When to involve a digital forensic expert The
OCR checklist Red flags When to involve a digital forensic expert

§ 13 · COMMON QUESTIONS Frequently asked questions Our platform says the documents are
searchable. Isn't that enough? How accurate is modern OCR really? Searches found nothing in
our scanned archive. Can we rely on that? Can OCR read handwriting, signatures and stamps?
Should we re-OCR an archive that was processed years ago? The other side produced scans
whose text garbles the key clause. What do we do?

§ 14 · REFERENCE Glossary Sources and authoritative references DISCLAIMER

§ HOW A SPECIALIST LABORATORY CAN ASSIST Working with Computer Forensics Lab
Speak to a forensic examiner, not a salesperson. INSTRUCT THE LAB
NEW ENQUIRIES EMAIL - DISCOVERY

Questions and answers

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, some time s 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 forensic all y where contested, and never reduced to its machine transcription in evidence. Stamps and an not at i on s 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 after no on: Example 2's defendant would have paid for that after no on 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

Source: <https://e-discovery.uk/library/ocr-when-searchable-does-not-mean-accurate>. Free to read and share with attribution to Computer Forensics Lab, e-discovery.uk.