

Deduplication Explained

Deduplication reduces data volumes by identifying and suppressing exact copies. This guide explains the mechanics of hashing, the difference between custodian-level and matter-level deduplication, and how to manage suppressed copies for evidential purposes. It addresses common problems and offers practical advice for UK litigators.

e-discovery.uk, Computer Forensics Lab. Published 2026-08-30.
<https://e-discovery.uk/library/deduplication-explained>

DEDUPLICATION · A GUIDE FOR UK LAWYERS Deduplication Explained Hashes, Scope Choices and Why "the Same Document" Is a Legal Question COMPUTER FORENSICS LAB

§ ABOUT THE AUTHOR PREPARED BY COMPUTER FORENSICS LAB E-DISCOVERY TEAM ESTABLISHED 2007 · LONDON ISO 17025-ALIGNED PROCEDURES EDRM PROCESSING PRACTICE

§ CONTENTS In this guide 01 Executive summary 02 The problem in plain English: the copies that eat the budget 03 How it works: hashes and what counts as identical 04 The scope decision: custodian-level vs matter-level 05 Email's special problems: near-copies that are not twins 06 Keeping what suppression hides: custodian tables and copy evidence 07 The forensic cousin: hashing as copy-tracing evidence 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: DEDUPLICATION TO REVIEW ONCE, BUT RECORD WHO ELSE HELD EVERY COPY: SUPPRESSION MUST LOSE COST, NOT EVIDENCE

§ 02 · FIRST PRINCIPLES The problem in plain English: the copies that eat the budget

§ 03 · THE MECHANICS How it works: hashes and what counts as identical 83) handles resemblance; dedup handles identity only.

§ 04 · THE SCOPE DECISION The scope decision: custodian-level vs matter-level

§ 05 · EMAIL 'SQUIRKS' Email's special problems: near-copies that are not twins

§ 06 · WHAT SUPPRESSION MUST KEEP Keeping what suppression hides: custodian tables and copy evidence

§ 07 · THE COUSIN DISCIPLINE The forensic cousin: hashing as copy-tracing evidence

§ 08 · THE WIDER MAP Source architecture: where else the evidence lives QUESTION SUPPRESSION PRESERVED INSTANCE DUP / COUNTERPART FORENSIC TABLE SOURCES METADATA THREAD PRODUCTION LAYER NEAR- LAYER

§ 09 · IN THE WILD Worked examples EXAMPLE 1 · THE MEMO THAT ELEVEN PEOPLE HELD EXAMPLE 2 · THE SCOPE FIGHT SETTLED BY ARITHMETIC EXAMPLE 3 · THE " DUPLICATE

" THATWASTHEFORGERY ' STWIN

§ 10 · WHEREITGOESWRONG 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 · DEDUPLICATION PARAGRAPHFORTHEDISCLOSUREPROTOCOL

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

§ 13 · COMMON QUESTIONS Frequently asked questions Does deduplication delete documents?
Custodian-level or matter-level: which should we agree? Can we still prove who received a
document if their copy was suppressed? The same email seems to appear several times in the
review set anyway. Why? Two productions contain "the same" contract. Do we need to check?
Can deduplication data itself be evidence?

§ 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. INSTRUCTTHELAB
NEWENQUIRIESEMAIL - DISCOVERY

Questions and answers

Does deduplication delete documents?

No: it suppresses copies from review sight while every source stays preserved and every suppressed instance stays recorded and restorable: a visibility decision over intact data. A provider who cannot restore an instance on request is describing a different, and in defensible, process.

Custodian-level or matter-level: which should we agree?

For most matters, matter-level with the full suppression table and restoration rights: Example 2's compromise: taking the larger saving while keeping holdings evidence as a lookup. Custodian-level earns its keep mainly where per-person holdings are the daily currency of the review itself; possession-critical sources get carve-outs under either scope.

Can we still prove who received a document if their copy was suppressed?

Yes, from the table: holder, source, path and dates per suppressed copy, surfaced as review fields and producible on request: Example 1's knowledge schedule is exactly this. The proviso is timing: the table is captured at processing: which is why §6 belongs in the protocol, not the post-mortem.

The same email seems to appear several times in the review set anyway. Why?

Usually legitimately: different recipes' field margins, platform-mutated bodies (disclaimers, encodings), journal vs mailbox forms, or the message quoted within later thread members: which are not duplicates but inclusiveness, guide 83's subject. Persistent unexplained repeats are a settings question for the provider: asked with message IDs in hand.

Two productions contain "the same" contract. Do we need to check?

The moment it is contested, yes, and it costs minutes: hash-match c on firms identity to mathematical standard; divergence triggers the byte and metadata comparison that found Example 3's rebuilt PDF. Assumed identity is the only dangerous kind: the check converts assumption into finding, which ever way it goes.

Can deduplication data itself be evidence?

Frequently: the distribution table dates organisational knowledge (who held the memo, in which folders), circulation breadth grounds notice arguments, and hash matches across estates prove copying per §7: review plumbing promoted to exhibits. Which is the quiet theme of this guide: done properly, the cost-saving machinery and the evidence machinery are the same table read twice. cflab. u k · e-disc ove r y. u k ©2026 Computer Forensics Lab Ltd ·cflab.uk ·e-discovery.uk ·info@cflab.uk ·+44 (0)20 7164 6915 Page 15 of 17

Source: <https://e-discovery.uk/library/deduplication-explained>. Free to read and share with attribution to Computer Forensics Lab, e-discovery.uk.