



C O M P U T E R F O R E N S I C S L A B

London digital forensics & e-discovery · Established 2007

# Deduplication Explained

## Hashes, Scope Choices and Why "the Same Document" Is a Legal Question

Deduplication is the quiet workhorse of every disclosure exercise: without it, review sets double and treble with copies of the same email harvested from every mailbox it touched, and costs follow. But "the same document" is less obvious than it sounds: the same attachment sent twice is two events; the same email in the sender's and recipient's boxes carries different evidential weight; a message trimmed by one platform's export is no longer hash-identical to its twin. The choices: hash method, custodian-level or matter-level scope, what metadata the hash includes, how suppressed copies are recorded: change what reviewers see, what gets disclosed, and what can be proved about who held what. This guide explains deduplication for UK lawyers: how it works, the scope decision and its consequences, where it goes wrong, and how to keep the suppressed copies' evidential value without paying to review them all.

🕒 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**, applies deduplication on every processed matter with recorded method, scope and suppression tables, and its forensic side relies on the same hashing science for evidence integrity, cross-device copy tracing and exfiltration proof: the two uses of one discipline, explained here together.

ESTABLISHED 2007 · LONDON

ISO 17025-ALIGNED PROCEDURES

EDRM PROCESSING PRACTICE

HASH & COPY-TRACE FORENSICS

CPR PART 35 EXPERT REPORTS

FULL CHAIN-OF-CUSTODY DOCUMENTATION

§ 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 |

## Executive summary

**THE HEADLINE POINT: DEDUPLICATE TO REVIEW ONCE, BUT RECORD WHO ELSE HELD EVERY COPY: SUPPRESSION MUST LOSE COST, NOT EVIDENCE**

Deduplication computes a fingerprint (hash) per item and keeps one instance of each: review effort collapses, consistency improves (one privilege call per document, not five conflicting ones), and productions shrink to sense. The governing decisions are legal, not technical. **Scope:** matter-level (one copy across all custodians) maximises savings; custodian-level (one copy per custodian) preserves per-person holdings in the workspace: the modern standard is matter-level *with a custodian table*: every suppressed copy's holder, location and path recorded against the surviving instance, so "who had this" stays answerable without re-review. **What counts as identical:** email hashing uses selected fields (participants, subject, body, attachment hashes), and platform quirks make true twins non-identical: method matters and is recorded. **What never deduplicates away:** family integrity (an attachment suppressed globally still appears within each produced family), and context where the copy itself is the fact: the draft on the defendant's laptop is evidence of possession that a hash match must not silently erase. The same hashing science, run forensically, proves copying across devices: §7's cousin discipline.

- **Matter-level plus custodian table:** the cost of copies removed, the evidence of who-held-what retained: the default worth agreeing.
- **Method is recorded:** hash algorithm, email fields included, tool and version: because different recipes disagree about "identical".
- **Families override global suppression:** attachments produce within every family they belong to: no orphaning by dedup.
- **Possession cases flip the logic:** where holding the copy is the issue, the copies are the evidence: dedup scoped consciously around them.
- **Near-duplicates are a different tool:** similar-but-not-identical documents are grouped for review efficiency, never suppressed as duplicates.

## The problem in plain English: the copies that eat the budget

Send one email to six colleagues and the estate now holds at least seven copies; collect those seven mailboxes and review will read the same words seven times, bill seven units, and: on a bad day: make seven slightly different relevance and privilege calls. Multiply by every message, every shared attachment, every file synced to every laptop, and duplication is routinely half or more of a collected corpus. Deduplication is simply the refusal to pay for the same reading twice.

The reason it deserves a guide rather than a checkbox is that "the same" is doing legal work. The same contract in the CEO's inbox and the auditor's file may be one document and two facts: each holding is a fact. The same wire instructions in five mailboxes may be one fraud and five notice questions. Suppression done naively answers the cost question by quietly deleting the evidence questions; suppression done properly: one instance reviewed, every instance recorded: answers both. The distance between those two is a settings decision and a table, which is what the rest of this guide is about.

## How it works: hashes and what counts as identical

- **Hashing in one paragraph:** an algorithm (MD5, SHA-1, SHA-256) reduces any file to a fixed fingerprint: identical input, identical hash; one byte's difference, a different hash: the same mathematics that verifies forensic images in guide 1 deduplicates review sets here.
- **Loose files:** hashed on binary content: two byte-identical spreadsheets match regardless of name, folder or dates: which is why file-system metadata (the path, the timestamps: often the evidence) must be captured before suppression, per §6.
- **Emails:** hashed on a recipe of fields: typically participants, subject, sent time, body and attachment hashes: because the raw message files differ per mailbox even for one sending: the recipe is the method, and it is disclosed with the production.
- **Attachments:** hashed individually and as family members: deduplicated globally for review, but produced within each family: the two roles reconciled by the platform, verified at load per guide 81 §5.
- **What hashing cannot see:** meaning: the re-saved PDF of the same contract, the forwarded body with one added line, the printed-and-rescanned page: all different hashes: near-duplicate and threading analysis (guide 83) handles resemblance; dedup handles identity only.

## The scope decision: custodian-level vs matter-level

- **Custodian-level:** one instance per custodian: each person's holdings stay visible in the workspace; cross-custodian copies still multiply review: the conservative legacy default.
- **Matter-level (global):** one instance across everyone: maximum saving, single consistent call per document: and, alone, it erases who-held-what from the reviewer's sight: never run without §6's table.
- **The modern answer:** matter-level with full suppression recording: the DRD-stage agreement to seek, because it gives both sides the cost profile and preserves the holdings evidence: Example 2's arithmetic.
- **Carve-outs by design:** possession-critical sources (the accused's devices in an IP theft; the recipient boxes in a notice dispute) excluded from global scope or flagged for instance-level treatment: the legal judgement encoded in settings.
- **Priority ordering:** which custodian's copy survives (and therefore whose metadata fronts the review) follows a recorded priority list: key custodians first: a small setting with outsized effects on what reviewers perceive.

## Email's special problems: near-copies that are not twins

- **One sending, many bodies:** platforms rewrite messages in transit and export: disclaimers appended, encodings shifted, images detached: true twins arriving hash-different: met by field-based recipes and, residually, by near-duplicate grouping.
- **Journal and archive forms:** envelope-journaled copies and archive stubs differ structurally from mailbox originals: mapped in processing so the same message is recognised across its forms.
- **BCC and distribution facts:** the sender's copy knows the BCC list; recipients' copies do not: suppression priority keeps the most informative instance, and the custodian table keeps the rest of the story.
- **Threads vs messages:** the same words quoted inside later replies are not duplicates of the original: that is inclusiveness, guide 83's subject: dedup must not be stretched to suppress earlier messages because later ones quote them.
- **Time zone traps:** sent-time fields rendered in different zones by different exports can break naive recipes: normalisation per guide 81 §3 precedes hashing for exactly this reason.

## Keeping what suppression hides: custodian tables and copy evidence

- **The custodian (dedup) table:** per surviving instance: every suppressed copy's custodian, source, path and native dates: the answer to "who held this, where" without re-processing: the non-negotiable companion to matter-level scope.
- **All-custodians fields in review:** the table surfaced as workspace fields, so a reviewer reading the one instance sees its full distribution at a glance: notice and knowledge questions answered in-line.
- **Path and location evidence:** the folder the copy sat in (Drafts; "Personal/Keep"; the USB root) is often the finding: captured into the table before the binary vanishes from view.
- **Production behaviour:** what is produced (instances vs distribution data) follows the protocol: with the table available to answer the opponent's who-held-what requests without a re-collection.
- **Reversibility:** suppression is a review-layer decision over preserved sources: any copy can be restored to sight if its instance-specific features become an issue: nothing was deleted, and the report says so.

## The forensic cousin: hashing as copy-tracing evidence

- **Same mathematics, opposite use:** review dedup hides copies to save cost; forensic hashing hunts copies to prove movement: the client file's hash found on the leaver's USB (guide 41), the pricing book in the personal cloud (guide 111's territory), the repository objects on the NAS (guide 76's Example 2).
- **Hash sets as search instruments:** a defensible list of the taken material's hashes swept across devices, cloud accounts and productions: presence proved to mathematical standard, absence bounded honestly (renamed-and-edited copies change hashes: fuzzy matching extends reach, stated as such).
- **Cross-production intelligence:** the opponent's production hashed against your client's estate: the "independent" document that is byte-identical to yours, per guide 76's shared-ancestry logic.
- **Notice and circulation analysis:** the dedup table itself as evidence: the memo's distribution across custodians dating organisational knowledge: the review artefact promoted to exhibit.
- **Integrity, throughout:** the same hashes verifying that what was reviewed and produced is what was collected: guide 1's discipline closing the loop.

## § 0 8 · THE WIDER MAP

## Source architecture: where else the evidence lives

Per this series' source-architecture discipline, adapted to the stage: when deduplication questions arise: who held this, which copy differs, where did the instance come from: the answers live beyond the surviving instance. They live in: the suppression table (the first stop: holders, paths, dates per copy); the preserved sources themselves (every suppressed binary intact and restorable); instance-level metadata captured at processing (per-copy paths and file-system dates); the near-duplicate and thread layers (the resemblance map around the identical core); counterpart productions (their copy of "the same" document, hash-compared: agreement authenticates, divergence is a finding); and the forensic layer (device artefacts dating when a copy arrived somewhere: creation and access times the review platform never needed). A dedup dispute is answered from these layers in minutes when §6 was done, and only by re-processing when it was not.

| QUESTION                                | SUPPRESSION TABLE        | PRESERVED SOURCES      | INSTANCE METADATA     | NEAR-DUP / THREAD LAYER   | COUNTERPART PRODUCTION | FORENSIC LAYER        |
|-----------------------------------------|--------------------------|------------------------|-----------------------|---------------------------|------------------------|-----------------------|
| <b>Who held this document?</b>          | Y: the direct answer     | Verifiable             | Per-copy detail       | n/a                       | Their holdings         | Device presence proof |
| <b>Which copy came first / differs?</b> | Dates per copy           | Y: binaries comparable | Y: paths + timestamps | Variant map               | Hash comparison        | Creation artefacts    |
| <b>Was suppression sound?</b>           | Method + recipe recorded | Re-hash verifiable     | Y                     | Non-identical caught here | Cross-check            | n/a                   |
| <b>Did they hold ours?</b>              | n/a                      | Your hash set          | n/a                   | Edited-copy reach         | Y: swept               | Y: the §7 hunt        |
| <b>When did a copy arrive?</b>          | Native dates recorded    | n/a                    | File-system dates     | n/a                       | n/a                    | Y: the dating layer   |

Fallback strategy in one line: build the suppression table at processing and every later dedup question is a lookup; skip it and every later question is a project.

## Worked examples

### EXAMPLE 1 · THE MEMO THAT ELEVEN PEOPLE HELD

A regulatory-exposure claim turned on when the board knew of a compliance warning. The warning memo survived dedup as one reviewed instance from the company secretary's mailbox: unremarkable: until the all-custodians field was read: the suppressed copies sat in eleven mailboxes, including three directors', with one copy filed to a folder named "Q3 board prep" six weeks before the pleaded date of knowledge. The dedup table, produced under the §11 request, became the knowledge schedule: eleven holders, dated receipts, telling folder names: one document, eleven facts. Matter-level suppression had saved ten reviews and, because the table existed, lost none of the eleven holdings.

### EXAMPLE 2 · THE SCOPE FIGHT SETTLED BY ARITHMETIC

At the DRD stage, a claimant demanded custodian-level dedup "to preserve holdings evidence"; the defendant's costs model showed matter-level saving 46% of a 730,000-document review. The compromise this guide recommends resolved it in one call: matter-level scope, full suppression table, all-custodians fields surfaced in both parties' review platforms, and instance restoration on request for any document where copy-specific features were put in issue. The claimant's holdings evidence survived intact (Example 1's kind of schedule remained a lookup); the defendant's saving survived almost intact; and the court's order recited the mechanism, which has since travelled to the parties' next dispute unlitigated.

### EXAMPLE 3 · THE "DUPLICATE" THAT WAS THE FORGERY'S TWIN

Two productions each contained "the same" signed side letter: the parties' bundles treated them as duplicates and reviewed one. The hash comparison, run as §8's counterpart check when the letter became central, said otherwise: the binaries differed. Byte-level and metadata examination found the divergence: the defendant's copy was a rebuilt PDF, producer string and creation date post-dating the letter by two years, its signature image resampled: while the claimant's copy carried the original 2019 producer chain. The "duplicate" label had nearly buried the case's best document-authenticity finding. The lesson is §3's closing point run in reverse: identical hashes prove identity; assumed identity proves nothing, and the assumption is checked the moment a shared document becomes contested.

## Common mistakes and technical limitations

### Common mistakes

- **Matter-level without the table:** the savings taken, the holdings evidence silently discarded: Example 1, unavailable.
- **Scope defaulted, not agreed:** the DRD conversation skipped, the opponent discovering the settings from the production's shape.
- **Possession cases deduped globally:** the copies that were the case suppressed into one anonymous instance.
- **Contested documents assumed identical:** Example 3's near-miss: the hash check never run when it started to matter.
- **Near-duplicates suppressed as duplicates:** resemblance treated as identity: variant wording: the evidence: hidden from review.
- **Attachment orphaning:** global suppression allowed to break families: guide 81 §5's arithmetic unrun.
- **Method undisclosed:** productions exchanged with no statement of algorithm, recipe or scope: the §11 questions answered by nobody.
- **Priority order unconsidered:** whose copy fronts the review left to chance, with the least-informative instances surviving.

### Technical limitations

- **Hashes see bytes, not meaning:** trivially altered copies escape; the fuzzy layers extend reach and say so.
- **Email recipes are conventions:** different tools' field choices disagree at the margin: cross-production comparisons account for method before crying foul.
- **Algorithm collisions are a theoretical footnote:** MD5's weaknesses matter to cryptographers; for dedup and integrity work the practice is modern algorithms and, where challenged, dual-hashing: stated, then moved past.
- **Tables are as good as capture:** paths and dates not extracted at processing cannot populate the table later without re-processing: §6 is a day-one setting.
- **Chat resists the concept:** messaging unitisation (guide 81 §6) determines what can even be a duplicate: dedup follows unitisation, never leads it.

## § 1 1 · INTERROGATORIES &amp; DRAFTING AIDS

## Questions to ask · Suggested wording

### Ask your client

- Is who-held-what likely to matter here (notice, knowledge, possession): so scope and carve-outs are set accordingly?
- Which custodians should head the priority order, so the most probative instances front the review?
- Are there documents whose copies we will want to hunt rather than hide: the §7 hash-set candidates?

### Ask your opponent

- Please state your deduplication method and scope: algorithm, email field recipe, custodian-level or matter-level, tool and version: and confirm family integrity was preserved.
- Please produce, for [documents/the production], the suppressed-instance data: each copy's custodian, source, path and native dates.
- Please confirm that any document whose instance-specific characteristics are in issue can and will be restored to instance-level production on request.

### Ask your eDiscovery / forensic provider

- Is the suppression table complete and surfaced as review fields: and what would restoring any instance take?
- Which contested documents should be hash-compared across productions before we rely on "same"?
- What does the distribution data already show on the notice and knowledge questions?

### SUGGESTED WORDING · DEDUPLICATION PARAGRAPH FOR THE DISCLOSURE PROTOCOL

"Deduplication shall be performed at matter level using [SHA-256] hashing, with email deduplication on stated fields, and the method, tool and version disclosed on request. For every suppressed instance, each party shall record and retain the custodian, source, original path and native date fields, and shall make such data available in review and produce it on request in respect of any document. Family relationships shall be preserved notwithstanding global suppression, and no attachment shall be produced outside its family context. Near-duplicate identification shall not be used to suppress non-identical documents. Any party may require instance-level production of a document whose copy-specific characteristics are in issue, and suppressed instances shall remain restorable from preserved sources throughout the proceedings."

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

### The deduplication checklist

- ☐ Scope decided as a legal question and agreed at the DRD stage
- ☐ Method recorded: algorithm, email recipe, tool, version
- ☐ Suppression table captured at processing: holder, source, path, dates per copy
- ☐ All-custodians fields surfaced in the review workspace
- ☐ Priority order set so probative instances survive
- ☐ Possession-critical sources carved out or instance-flagged
- ☐ Family integrity verified against global suppression
- ☐ Near-duplicate grouping kept distinct from suppression
- ☐ Contested "identical" documents hash-verified across productions
- ☐ Restoration path documented: nothing deleted, everything recallable

### Red flags

- Productions silent on dedup method and scope.
- Who-held-what questions answered "we would have to reprocess".
- Attachments appearing without parents: suppression breaking families.
- Key documents present once where distribution should be wide: table unread.
- "Duplicates" of contested documents never hash-compared: Example 3's trap.
- Suspiciously heavy suppression against a possession-based claim.
- Counts moving between productions with no dedup delta explained.

### When to involve a digital forensic expert

At protocol drafting, where scope, recipe and the table are specified before they are argued about; at processing, where §6's capture is a setting, not a retrofit; at review, where distribution data is turned into knowledge and notice schedules (Example 1) and contested documents are hash-verified (Example 3); and on the forensic side, where hash sets hunt copies across devices, clouds and productions per §7 and report under CPR Part 35. Through **e-discovery.uk**, deduplication runs with the table as standard and the savings of Example 2 without its fight.

## § 1 3 · COMMON QUESTIONS

## Frequently asked questions

**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 indefensible, 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 confirms 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, whichever 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.

## § 1 4 · REFERENCE

# Glossary

## All-custodians field

The review field listing every holder of a document's copies.

## Custodian-level dedup

One instance kept per custodian.

## Dedup recipe

The field set hashed to identify email duplicates.

## Fuzzy / near-duplicate matching

Grouping similar, non-identical documents; never suppression.

## Hash

A content-derived fingerprint; identical bytes, identical hash.

## Instance

One physical copy of a document in one location.

## Matter-level dedup

One instance kept across the whole matter.

## Priority order

The custodian ranking deciding which copy survives.

## Suppression table

The record of every suppressed copy's holder, path and dates.

## SHA-256

The current-practice hashing algorithm for integrity and dedup.

## Sources and authoritative references

The deduplication 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 (deduplication method, scope and suppression recording as practised): [e-discovery.uk/edrm/processing](https://e-discovery.uk/edrm/processing) · Phase 05 · Review (distribution data in review): [e-discovery.uk/edrm/review](https://e-discovery.uk/edrm/review)
- cflab.uk, digital forensics services (hash-set copy tracing, cross-production comparison, document authenticity, CPR Part 35 reporting): [cflab.uk/digital-forensics-services](https://cflab.uk/digital-forensics-services) · guides library: [cflab.uk/guides](https://cflab.uk/guides)
- e-discovery.uk, practice method and Knowledge Centre: [e-discovery.uk](https://e-discovery.uk) · [e-discovery.uk/knowledge](https://e-discovery.uk/knowledge)
- PD 57AD and the Disclosure Review Document: [justice.gov.uk](https://justice.gov.uk), PD 57AD · CPR Part 35: [justice.gov.uk](https://justice.gov.uk), Part 35
- NIST, Secure Hash Standard (FIPS 180-4): [csrc.nist.gov](https://csrc.nist.gov), FIPS 180-4
- ICO, UK GDPR guidance: [ico.org.uk](https://ico.org.uk) · ISO/IEC 27037: [iso.org](https://iso.org), 27037

## DISCLAIMER

This publication provides general information about deduplication as at August 2026. Tool behaviour and method details vary by platform and version; verify the current position for the pipeline in the matter at hand. The worked examples are fictional. This guide is not legal advice, does not create a professional relationship, and should not be relied upon in place of advice on the facts of a specific matter from a qualified lawyer or a suitably qualified forensic practitioner. Links were live at the date of publication.

## § HOW A SPECIALIST LABORATORY CAN ASSIST

## Working with Computer Forensics Lab

Deduplication at **e-discovery.uk** runs to this guide's standard because both of its readings matter: matter-level savings taken with the suppression table captured at processing, all-custodians data surfaced in review, families verified, and any instance restorable on request: Example 2's mechanism as default practice. The laboratory side runs the cousin discipline: hash sets hunting taken material across devices, clouds and productions, cross-production verification of contested "identical" documents (Example 3's check), and distribution schedules built from the table for knowledge and notice arguments (Example 1's schedule), reported under CPR Part 35. Conflicts checks, confidential scoping discussions and fixed-fee estimates are routine; and if your live matter's production cannot say who else held its key document, the §11 request is ready to send.



## I N S T R U C T T H E L A B

### 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