



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

What Happens During eDiscovery Processing?

The Black Box Between Collection and Review, Opened for Lawyers

Between "we collected 400 GB" and "there are 180,000 documents in the review platform" sits processing: the stage most lawyers never see and every disclosure statement silently depends on. Containers are unpacked, families extracted, text and metadata pulled, exceptions logged, duplicates suppressed, dates normalised and filters applied: and each of those steps involves decisions that change what reviewers will and will not see. When numbers do not reconcile, when an attachment is missing, when the opponent asks why 400 GB became 180,000 documents, processing answers: if it was documented. This guide opens the box for UK lawyers: the pipeline stage by stage, the decisions that matter (exceptions, deNISTing, date filters, container handling), the reconciliation arithmetic that makes counts defensible, and the questions to ask a provider before and after the button is pressed.

⌚ Estimated reading time: 25 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 processing as a documented, reconciled stage on every matter: container handling, exception management, defensible culling and load-ready output: with the audit trail that lets counts be explained to opponents, courts and clients. This guide is the stage's owner explaining it.

ESTABLISHED 2007 · LONDON

ISO 17025-ALIGNED PROCEDURES

EDRM PROCESSING PRACTICE

EXCEPTION & RECONCILIATION REPORTING

CPR PART 35 EXPERT REPORTS

FULL CHAIN-OF-CUSTODY DOCUMENTATION

§ CONTENTS

In this guide

01	Executive summary
02	The problem in plain English: the stage nobody watches
03	The pipeline: from collected data to reviewable documents
04	The decisions that matter: exceptions, deNISTing and filters
05	Reconciliation: making the numbers add up
06	Special formats: mobile data, chat, cloud exports and structured sources
07	Processing disputes: challenging and defending the counts
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: PROCESSING TURNS DATA INTO DOCUMENTS THROUGH RECORDED DECISIONS: THE RECORD IS WHAT MAKES YOUR DISCLOSURE DEFENSIBLE

Processing is a pipeline: ingest the collected containers; expand them (PSTs into emails, zips into files, emails into families of message plus attachments); extract text, metadata and embedded items; normalise dates and encodings to one standard; hash everything for deduplication (guide 82's subject); apply the agreed cull criteria (deNISTing system files, date ranges, file-type decisions); log every item that fails: encrypted, corrupt, unrecognised: as a tracked exception; and load the survivors, families intact, into review. Two disciplines make this defensible rather than merely done. **Exception management:** the failed items are not noise: the encrypted archive on the finance director's laptop may be the case: every exception is listed, triaged and resolved or consciously accepted, in writing. **Reconciliation:** collected volume to expanded items to culled items to loaded documents, every reduction attributed to a named step with counts: so that "400 GB became 180,000 documents" has an arithmetic answer, which is precisely what PD 57AD cooperation, opponents' questions and your own disclosure statement require.

- **Families are sacred:** emails and their attachments travel together through every step: a family broken in processing is a disclosure defect found at the worst time.
- **Exceptions are a report, not a shrug:** encrypted, corrupt and unsupported items listed, triaged and resolved: the buried treasure is usually here.
- **Culling is a recorded decision:** deNISTing, date filters and type filters applied per the agreed protocol, with counts per criterion.
- **Reconciliation is the defence:** every stage's in and out counts retained: the arithmetic that answers "where did the rest go".
- **Settings are choices:** time zones, deduplication scope, container depth: agreed before the button, because they shape what review sees.

The problem in plain English: the stage nobody watches

Collection gets forensic ceremony; review gets teams and budgets; processing happens overnight in a data centre and is reported in one line: "processing complete, 182,414 documents loaded." Yet processing made more decisions about your disclosure than any reviewer will: it decided that 30,000 system files were irrelevant by definition, that two "identical" emails were one, that an unreadable container was an exception for the appendix, that dates would be displayed in UTC. Each was probably right. All of them are your disclosure statement's silent premises.

The lawyer's job is not to operate the pipeline but to govern it: agree the settings that embody legal judgements (what is culled, what time zone, whose deduplication), demand the exception and reconciliation reports that record what happened, and read them: because the alternative is discovering at the PD 57AD cooperation stage, or in a witness box, that nobody can say where a quarter of the collection went. Processing done well is invisible; processing governed well is defensible: and only the second survives contact with a competent opponent.

The pipeline: from collected data to reviewable documents

- **Ingest and verify:** collected media and exports received against chain-of-custody records, hashes verified: the join to guides 1 and 71's disciplines: nothing processes until provenance reconciles.
- **Expansion:** containers opened recursively: PST/OST/mbox into messages, archives into files, messages into parent-plus-attachment families, embedded objects surfaced: with container depth and embedded-item policies set deliberately.
- **Extraction and normalisation:** text (native or via OCR routing for image-only items: guide 86's subject), metadata fields mapped to a uniform schema, dates converted to a single zone with the original preserved, encodings unified.
- **Hashing and deduplication:** per-item hashes computed; duplicates identified and suppressed per the agreed scope: custodian-level or matter-level: guide 82 treats the choices and their consequences.
- **Culling and load:** deNISTing, date and type filters applied with per-criterion counts; survivors indexed, families intact, and loaded to the review platform with a load file the platform's fields actually match.

The decisions that matter: exceptions, deNISTing and filters

- **Exception handling:** password-protected and encrypted items (credential hunts, cracking under authority, custodian requests), corrupt files (repair attempts, source re-pulls), unsupported formats (specialist tooling): each logged with disposition: resolved, replaced, or accepted-and-listed: the report the opponent is entitled to ask about.
- **DeNISTing and system-file culls:** known operating-system and application files removed by hash list: near-universally sound, but the list version and counts are recorded, and the rare matter where system artefacts matter (guides 41, 61) opts out knowingly.
- **Date filtering:** the range from the protocol applied against chosen date fields: which field (sent, modified, created) is a legal decision with edge cases (undated items are kept and flagged, not silently dropped).
- **File-type decisions:** inclusions and exclusions by type per the matter's needs: with the awareness that type detection reads signatures, not extensions, and mislabeled files are caught, not trusted.
- **Time zone and display settings:** one zone for review consistency, originals preserved for the record: the guide 70 reference-clock discipline, implemented here: agreed in the protocol so both sides' chronologies match.

Reconciliation: making the numbers add up

- **The chain of counts:** collected volume and item counts; expanded item counts; per-criterion cull counts (deNIST, date, type); deduplication suppressions; exceptions by category; loaded documents and families: one table, no gaps.
- **Volume vs count literacy:** gigabytes do not predict documents: containers expand, duplicates collapse, system files vanish: the table is what converts the opponent's suspicious "400 GB to 180,000?" into a five-line answer.
- **Family integrity checks:** parent and child counts reconciled; orphaned attachments and childless parents investigated: the classic processing defect caught by arithmetic before review finds it by accident.
- **The processing report:** settings, tool and version, hash lists used, per-stage counts, exception log: the standing deliverable on every e-discovery.uk matter, and the exhibit if processing is ever challenged.
- **Reprocessing discipline:** when settings change mid-matter (a new date range, a dedup scope change), the delta is run, documented and reconciled: never silently merged over the original.

Special formats: mobile data, chat, cloud exports and structured sources

- **Mobile extractions:** UFDR and similar packages processed with message threads, media and app data mapped to reviewable units: the unitisation question (what is "a document" in a five-year chat?) decided by protocol, per guide 63-65's platform realities.
- **Chat and collaboration data:** Slack and Teams exports (guides 66, 51-57) sliced into review units (channel-day, thread, or conversation) with edits, deletions and reactions represented: the slicing choice changes what reviewers perceive, so it is agreed, not defaulted.
- **Cloud and platform exports:** Google Takeout, Purview exports, platform APIs: each with format quirks (labels vs folders, version handling) that processing maps deliberately to the review schema.
- **Structured data:** database slices and log extracts (guides 71-74, 79) processed as data, not flattened into page images: loaded to analytical tools beside the document review, joined by the guide 70 chronology.
- **Media-heavy sets:** audio, video and image collections routed to transcription, technical review or specialist handling (guide 77's disciplines) rather than forced through a text pipeline that cannot see them.

§ 0 7 · WHEN COUNTS ARE CONTESTED

Processing disputes: challenging and defending the counts

- **Challenging theirs:** the §11 questions: settings, exception report, reconciliation table: put to the producing party: a production that cannot explain its own arithmetic invites the inference that the arithmetic was never done.
- **Defending yours:** the processing report answers most challenges in a letter: which is why it is written contemporaneously, not reconstructed under fire.
- **The classic attack surfaces:** broken families, missing attachments, over-broad culls, exceptions never resolved, dedup suppressing unique-in-context copies (guide 82's territory), date-field choices excluding the undated: each preventable by this guide's disciplines, each discoverable by its questions.
- **Expert engagement:** processing disputes resolve fastest specialist-to-specialist: settings compared, deltas identified, reprocessing scoped: the guide 71 §7 agreed-queries spirit applied to the pipeline.
- **Proportionality honestly:** reprocessing costs are real: challenges target material defects (the missing custodian's exceptions) rather than cosmetic ones, and courts reward the party whose complaints come with counts attached.

§ 0 8 · THE WIDER MAP

Source architecture: where else the evidence lives

Per this series' source-architecture discipline, adapted to a workflow stage: when processing cannot render an item, the item is rarely the only copy of its content. The exception's content may also live in: the source system, re-pulled in a friendlier format (the corrupt PST's mailbox re-exported; the broken file's DMS version ladder per guide 75); counterpart copies (the unreadable attachment as sent, intact in the recipient's production); earlier or later versions of the same document elsewhere in the estate; rendered forms (the print, the PDF, the email quoting the contents); credential holders for the encrypted (custodians, IT, password managers: sought before cracking); and the backup layer preceding the corruption. Exception resolution is a small source-architecture exercise per item: which is why the exception report lists sources tried, not just failures logged.

EXCEPTION TYPE	SOURCE RE-PULL	COUNTERPART COPIES	OTHER VERSIONS / RENDERS	CREDENTIALS & KEYS	BACKUPS	NET RESULT
Encrypted / password items	Sometimes unencrypted at source	Y: sent copies often clear	Rendered outputs exist	Y: the primary route	Pre-encryption states	Mostly resolvable
Corrupt files	Y: re-export beats repair	Y	Version ladders	n/a	Y: the classic rescue	Usually recoverable
Unsupported formats	Native application export	Converted copies in circulation	PDF/print renders	n/a	n/a	Specialist tooling resolves
Empty / stub items	Y: archive stubs re-hydrated at source	Full copies with recipients	n/a	n/a	Journal archives	Source-side fix
Missing attachments	Family re-extraction	Y: the sent family intact	Attachment as standalone file	n/a	Y	Arithmetic finds, sources fix

Fallback strategy in one line: treat every material exception as a mini investigation across the estate: the pipeline failing to read an item is the beginning of the work, not the end of it.

Worked examples

EXAMPLE 1 · THE EXCEPTION THAT WAS THE CASE

A £6m fraud claim processed 61 GB from the defendant company's own systems under a court-supervised protocol. The exception report: 214 items: contained one line that mattered: a 2.1 GB password-protected 7-Zip archive named "old backups" in the finance director's user share, last modified two days after the freezing order. Credential requests failed; the court ordered its production or the password's; and the eventually opened archive held the parallel invoice ledger the culled-and-reviewed 180,000 documents never mentioned. Had the exception report been a count rather than a list: "214 items could not be processed": nobody would ever have asked. The item-level exception log is the whole lesson.

EXAMPLE 2 · THE 400 GB THAT BECAME 41,000 DOCUMENTS, DEFENSIBLY

An opponent attacked a production's arithmetic: 400 GB collected, 41,000 documents produced, "plainly incomplete". The reconciliation table answered in one page: 400 GB expanded to 1.9m items; deNISTing removed 1.1m by the published hash list (counts and list version cited); the agreed 2019-2023 date range removed 410,000 (field: sent/modified per protocol; 3,200 undated items retained and reviewed); matter-level deduplication suppressed 240,000 with duplicate custodians recorded; 108,000 entered review; 41,000 met the agreed issues. Every reduction had a name, a criterion and a count. The application evaporated at the CMC, with the judge observing that the table was what a disclosure certificate is supposed to look like underneath.

EXAMPLE 3 · THE FAMILIES THAT CAME APART

Midway through a review, a claimant team noticed privilege calls being made on emails whose attachments were nowhere in the workspace. The family-integrity check that should have run at load ran late instead: 6,100 attachments had been orphaned by a container-handling setting that expanded embedded messages as standalone items, breaking parent links. The fix was mechanical: reprocess the affected containers with corrected settings, reconcile parent-child counts, overlay the corrected families: but 9,000 documents needed re-review with their context restored, and two early privilege determinations reversed. The costs of the re-review were borne where the settings decision had been made. Family arithmetic at load is cheaper than family archaeology at review.

Common mistakes and technical limitations

Common mistakes

- **Settings by default:** time zones, dedup scope and container depth left to the tool's presets: legal decisions made by software nobody instructed.
- **Exception counts without lists:** Example 1 inverted: the archive that was the case, summarised into invisibility.
- **No reconciliation table:** the arithmetic reconstructed under challenge instead of recorded in course: Example 2's defence, unavailable.
- **Families unchecked at load:** Example 3's re-review: parent-child counts not reconciled until reviewers stumble.
- **Culls beyond the protocol:** type or date filters applied unilaterally: the cooperation obligation breached by settings.
- **OCR routing skipped:** image-only documents indexed as empty text and invisible to every search thereafter: guide 86's problem, born here.
- **Chat unitisation defaulted:** five-year threads as single documents, or single messages stripped of context: the slicing decision nobody made.
- **Reprocessing merged silently:** setting changes overlaid without deltas: two incompatible arithmetics in one workspace.

Technical limitations

- **Tools disagree:** different platforms expand, count and extract differently: cross-platform count comparisons need translation, not alarm.
- **Extraction is imperfect:** exotic formats, damaged items and embedded oddities defeat any pipeline: the exception process exists because perfection does not.
- **Metadata is as good as the source:** processing preserves what collection captured: it cannot restore fields a bad export never carried (the guide 75 flattening, met downstream).
- **Text is not meaning:** extraction feeds search; it does not understand: the guide 84-90 layers build on what this stage yields.
- **Scale costs time:** terabyte matters process in days, not hours: the timetable's dependency, declared early.

§ 1 1 · INTERROGATORIES & DRAFTING AIDS

Questions to ask · Suggested wording

Ask your client

- Do the collected sources carry any known awkward formats: legacy archives, encrypted stores, chat exports: the pipeline should plan for?
- Who holds credentials for protected material, and is the request to them made before cracking is priced?
- Which culling decisions are ours to make, and which need the opponent's agreement under the protocol?

Ask your opponent

- Please state the processing settings applied to your production: tool and version, time zone, deduplication scope and method, deNIST list version, date fields and ranges, and file-type inclusions or exclusions.
- Please produce your exception report, listing items that could not be processed by category and their disposition, and your reconciliation of collected volume to produced documents with counts per reduction step.
- Please confirm that parent-child family relationships have been preserved and reconciled, and state the unitisation applied to any chat or messaging data.

Ask your eDiscovery / forensic provider

- Which settings need our instruction rather than your defaults: and what does each choice change?
- What is in the exception list, item by item: and which exceptions have a §8 route worth pursuing?
- Does the reconciliation table balance: and will it be maintained through every reprocessing?

SUGGESTED WORDING · PROCESSING PARAGRAPH FOR THE DISCLOSURE PROTOCOL

"Processing shall be performed with settings recorded and disclosed on request, including tool and version, time zone (which shall be [UTC] for review display with original values preserved), deduplication method and scope, deNIST hash list and version, date fields and ranges applied, and any file-type inclusions or exclusions. Parent-child relationships shall be preserved and reconciled at load. Each party shall maintain and produce on request: (a) an exception report listing items not successfully processed, by category and disposition; and (b) a reconciliation of collected data through each processing step to documents loaded for review, with counts per step. Items bearing no date within the applied fields shall be retained and flagged, not excluded. Changes to processing settings after first load shall be documented and their effects reconciled separately."

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

The processing checklist

- ☐ Provenance verified at ingest against collection records
- ☐ Settings instructed, not defaulted: zone, dedup, depth, filters: and recorded
- ☐ Culls applied per protocol with per-criterion counts
- ☐ Exception report maintained at item level, triaged, dispositions recorded
- ☐ Material exceptions worked through the §8 sources
- ☐ Families reconciled at load: parents, children, orphans
- ☐ Undated items retained and flagged
- ☐ Chat and mobile unitisation agreed and documented
- ☐ Reconciliation table maintained through every stage and reprocessing
- ☐ Processing report written contemporaneously

Red flags

- Productions whose arithmetic cannot be stated: Example 2's target.
- Exception "reports" that are a single number: Example 1's near-miss.
- Attachments missing from produced families: Example 3's symptom.
- Large encrypted items in key custodians' shares, unresolved.
- Culling criteria appearing nowhere in the agreed protocol.
- Empty-text documents clustering where scans should be: OCR skipped.
- Counts that changed between productions without a documented delta.

When to involve a digital forensic expert

At protocol drafting, where the settings that need agreement are named before they are argued about; at the pipeline, where exception triage and the §8 source-architecture work turn failures into evidence (Example 1's archive); at challenge time, in either direction, where reconciliation tables are built or audited specialist-to-specialist; and at reporting, where processing methodology is evidenced under CPR Part 35 when it becomes an issue. Through **e-discovery.uk**, processing runs as a documented EDRM stage with the report as standard deliverable: the black box, kept open by design.

§ 13 · COMMON QUESTIONS

Frequently asked questions

Why did 400 GB of data become only 40,000 documents?

Because volume and count measure different things: containers expand into items, then deNISTing removes the operating system's furniture, date and scope filters remove the out-of-range, and deduplication collapses the copies: each step with a recordable count. The right response to the question is Example 2's table; suspicion is only warranted when no table exists.

What actually happens to files that cannot be processed?

They become exceptions: logged by category (encrypted, corrupt, unsupported, empty), triaged, and either resolved through the §8 routes: credentials, source re-pulls, counterpart copies, specialist tooling: or consciously accepted and listed. What must never happen is silent disappearance: the exception list is disclosure work-product, and Example 1 is why.

Is it safe to remove "system files" from the review set?

Almost always: deNISTing removes known-hash operating system and application files with no user content, and doing so is standard, published-list, recorded practice. The caveat is the matter where system artefacts are the evidence: user-activity and exfiltration cases per guides 41 and 61: which opt out of the cull for the relevant sources, knowingly and in writing.

Our opponent's production has emails whose attachments are missing. Innocent?

Sometimes (source-side stubbing, genuine transmission failures): but it is exactly what broken family handling looks like, and the question is answerable: ask for their family reconciliation and unitisation settings per §11. A producing party that preserved families can say so in a sentence; one that cannot has Example 3 ahead of it.

Can we change processing decisions after review has started?

Yes, with discipline: new ranges, scopes or unitisation are run as documented deltas: reprocessed, reconciled, overlaid: never silently merged, so the workspace's arithmetic stays coherent and prior review work maps cleanly onto the corrected set. Example 3's fix is the model; its costs are the argument for getting settings instructed the first time.

Does processing alter the underlying evidence?

No: it derives from preserved originals: text, images, normalised fields: while the collected sources stay sealed and hashed, and any produced native is traceable back to them. What processing does alter is visibility: which is the whole point of governing it, since a document invisible to search and review is, functionally, undisclosed.

§ 1 4 · REFERENCE

Glossary

DeNISTing

Removing known system files by published hash list.

Exception

An item the pipeline could not process, logged for triage.

Expansion

Opening containers recursively into individual items.

Family

A parent document and its attachments, kept together.

Load file

The structured index delivering documents and fields to review.

Normalisation

Unifying dates, zones and encodings to one standard.

Orphan

An attachment separated from its parent; a defect signal.

Reconciliation

The per-stage count table from collection to load.

Stub

A placeholder whose content lives in an archive system.

Unitisation

Deciding what constitutes one reviewable document, especially in chat.

Sources and authoritative references

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

- e-discovery.uk, EDM Phase 04 · Processing (the pipeline, exception management and reconciliation as practised): e-discovery.uk/edrm/processing · Phase 03 · Collection and Phase 05 · Review (the neighbouring stages): e-discovery.uk/edrm/collection · e-discovery.uk/edrm/review
- cflab.uk, digital forensics services (exception investigation, encrypted-item handling, processing dispute support): 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 (including the Disclosure Review Document's technology sections): justice.gov.uk, PD 57AD · CPR Part 35: justice.gov.uk, Part 35
- NIST National Software Reference Library (the deNIST hash source): nist.gov, NSRL
- ICO, UK GDPR guidance: ico.org.uk · ISO/IEC 27037: iso.org, 27037

DISCLAIMER

This publication provides general information about eDiscovery processing as at August 2026. Tool behaviour, format support and platform capabilities vary by product and version and change on vendors' schedules; 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

Processing at **e-discovery.uk** is run as this guide describes because the alternative is indefensible: settings instructed and recorded, families reconciled at load, exceptions listed item by item and worked through the source-architecture routes, and the reconciliation table maintained from first ingest to final production: the deliverables of Examples 1 and 2 as standard practice, and Example 3's defect designed out at the settings stage. The laboratory side handles what the pipeline surfaces: encrypted archives, damaged media, forensic re-pulls: and both sides of processing disputes, building or auditing the arithmetic specialist-to-specialist and evidencing methodology under CPR Part 35 where it becomes an issue. Conflicts checks, confidential scoping discussions and fixed-fee estimates are routine; and if your current provider cannot produce an exception list and a reconciliation table for your live matter this week, that is the finding.



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