



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

Early Data Assessment Before Disclosure Costs Escalate

Estimate First, Commit Second

The most expensive disclosure decisions (custodian counts, search scope, processing commitments, review budgets) are traditionally made at the moment of greatest ignorance. Early data assessment reverses that: light-touch analysis of volume, custodians, file types, duplication, communication patterns, privileged material and likely-relevant datasets, before full processing is bought, so the DRD, the costs budget and the strategy rest on measured numbers rather than guesses.

© Estimated reading time: 29 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**, is built on the economics this guide describes: review is where budgets are spent, and the phases before review are where budgets are made. Our matters open with assessment (custodians, repositories, scope, volumes) precisely so that culling can be aggressive and clients pay for substance rather than volume. Our examiners also produce CPR Part 35 and CrimPR Part 19 compliant expert reports and give evidence in UK courts and tribunals.

ESTABLISHED 2007 · LONDON

ISO 17025-ALIGNED PROCEDURES

ISO 27001-ALIGNED SECURITY

ACPO / NPCC DIGITAL EVIDENCE PRINCIPLES

EARLY CASE & DATA ASSESSMENT

RELATIVITY-READY HOSTED REVIEW

FULL CHAIN-OF-CUSTODY DOCUMENTATION

§ CONTENTS

In this guide

- 01 Executive summary
- 02 The problem in plain English: decisions made blind
- 03 Why early assessment matters legally
- 04 What early data assessment actually is
- 05 Estimate 1 · Volume
- 06 Estimate 2 · Custodians
- 07 Estimate 3 · File types
- 08 Estimate 4 · Duplication
- 09 Estimate 5 · Communication patterns
- 10 Estimate 6 · Privileged material
- 11 Estimate 7 · Potentially relevant datasets
- 12 The ten-day EDA workflow and the assessment report
- 13 The post-assessment decision tree
- 14 Worked example: Marwell v Casterbridge, with numbers
- 15 Data protection, privilege and the assessment itself
- 16 Common mistakes and technical limitations
- 17 Questions to ask · Suggested wording for instructions
- 18 Checklist and red flags · When to involve a digital forensic expert
- 19 Frequently asked questions

Executive summary

THE HEADLINE POINT: WHY EARLY DATA ASSESSMENT MATTERS

Early data assessment (EDA) is light-touch, metadata-led analysis of the client's data estate, performed before full processing and review are bought, to measure what the traditional workflow only discovers after the money is spent: how much data there really is, which custodians actually communicated on the issues, what file types and duplication rates lurk in the population, where privilege concentrates, and how large the genuinely relevant core is likely to be. It typically costs a low single-digit percentage of the disclosure budget and moves the biggest decisions (custodian lists, search scope, staging, budgets, and sometimes settlement posture) from guesswork onto measured numbers.

- **The cost curve is a ratchet.** Each EDRM phase multiplies the previous one: collection is priced per source, processing per gigabyte, review per document, and review dwarfs everything. Decisions taken before processing set the size of everything after it, which is why assessment belongs before commitment.
- **Seven estimates cover the ground.** Volume, custodians, file types, duplication, communication patterns, privileged material and likely-relevant datasets: each is measurable early, cheaply and to useful accuracy from metadata, samples and platform reports.
- **EDA output is procedural currency.** The DRD asks for data volumes and cost estimates; costs budgets must state disclosure figures the court can test; proportionality arguments need numbers on both sides. An assessment report converts all three from advocacy into evidence.
- **Assessment work is never wasted.** Done properly, the indexed seed data and its analytics promote directly into full processing for whatever survives the decisions, so EDA is a down payment, not a duplicate cost.
- **Assessment also serves minimisation.** Measuring in place, and processing only what the numbers justify, is what UK GDPR data-minimisation looks like in disclosure practice: assess broadly, ingest narrowly.

The problem in plain English: decisions made blind

The traditional disclosure timetable has an awkward shape. The decisions with the largest financial consequences arrive first: which custodians, which sources, what date ranges, what terms, what budget to put in Precedent H, what to propose in the DRD. The information that would make those decisions well arrives last, as processing reports and review statistics, months later and after the spend. The result is familiar to every litigation partner:

- Custodian lists negotiated on instinct, then found (in month four) to include three people who barely feature and to omit the coordinator who ran everything.
- Budgets built on a guessed document count that the first processing report doubles.
- Search terms agreed blind, producing the 1.4 million-hit population that the timetable cannot absorb.
- Proportionality submissions made in adjectives ("vast", "disproportionate") that the court, quite reasonably, discounts.
- Settlement decisions taken without knowing whether the documents help or hurt, because nobody has looked at even a sample.

Early data assessment is the correction: a short, deliberately cheap phase between identification and full processing whose entire purpose is to measure before committing. It borrows its tools from processing and analytics but applies them lightly (metadata and samples rather than full extraction) and it ends in a report and a set of decisions, not a review population.

Why early assessment matters legally

- **The DRD assumes you have the numbers.** PD 57AD's Disclosure Review Document asks parties to describe data volumes, sources and the likely cost of the proposals they advance, and the court is directed to consider proportionality when choosing Models. A party with an assessment report completes those sections with measurements; a party without one either guesses in a court document or delays.
- **Costs budgeting is a commitment.** Disclosure figures in Precedent H are scrutinised, and revising them upward mid-matter needs justification. Budgets built on EDA numbers (documents per custodian, dedupe rates, forecast review populations) survive contact with reality; budgets built on analogies to the last case do not.
- **PD 31B expects an informed discussion.** The pre-CMC obligation to discuss sources, tools and search strategy presupposes each party knows its own estate; the Electronic Documents Questionnaire is, in substance, a request for assessment outputs.
- **Proportionality arguments are won with numbers.** "Restoring those archives adds an estimated 9,000 unique documents at a processing and review cost of £62,000, against live sources already covering 96 per cent of the period" is a submission; the same point without numbers is a complaint.
- **Early merits insight is legitimate strategy.** Sampling the core custodians early tells the client what the documents look like before positions harden: Part 36 decisions, admissions, and the framing of Issues for Disclosure all improve when someone has actually read a sample.
- **Criminal and regulatory parallels.** Investigators' scoping of digital material, and the Disclosure Management Document's description of strategy, are assessment products by another name; regulator deadlines (an FCA information requirement, for instance) are survivable precisely to the extent the estate is already measured.

§ 0 4 · DEFINITIONS

What early data assessment actually is

EDA is a method spectrum, not a single product. Four levels, in ascending intrusiveness and precision:

LEVEL	WHAT IS DONE	WHAT IT YIELDS
1 · In-place reporting	Tenancy and platform statistics without moving data: mailbox sizes, item counts, Purview / Vault search estimates, drive and site inventories, chat volumes	Volume and per-custodian scale in hours, at near-zero cost; first-cut term hit estimates from platform search
2 · Metadata-only ingestion	Collect targeted seed sources; index metadata (dates, senders, recipients, types, sizes, hashes) without full text extraction of everything	File-type profiles, dedupe rates, timelines, communication maps, exact per-custodian counts
3 · Sampled content analysis	Extract text for samples and priority sets; test candidate terms; read structured samples of core custodians	Term hit forecasts with relevance rates, richness estimates, privilege concentration, merits feel
4 · Assessment platform / ECA workspace	Seed data in an early-case-assessment environment with analytics (clustering, threading, maps) ahead of any review commitment	The full assessment report of \$12, with data ready to promote into processing

Two distinctions keep the discipline honest. **EDA is not full processing:** it defers exception resolution, full OCR and family-perfect extraction, which is why it is cheap and fast, and also why its numbers carry stated tolerances. **EDA is not review:** sampled reading informs decisions; it does not code documents. The working principle: **assess on metadata and samples; commit on evidence; promote, never redo.**

Volume

The first question every budget and every DRD section asks: how much is there?

- **Measure, per custodian and per source:** mailbox sizes and item counts (platform admin reports give these in minutes), drive and site sizes with file counts, chat message counts, archive and PST inventories, and structured-system record counts via the system owner.
- **Convert gigabytes to documents carefully.** Document yield per gigabyte varies enormously by data type (mail-heavy custodians yield thousands of items per gigabyte; CAD and media directories yield dozens), which is why the file-type profile in §7 is measured alongside volume rather than assumed. State conversions as ranges with their basis.
- **Expect expansion.** Containers expand on processing (archives, embedded attachments, mail stores), so collected gigabytes understate processed volume; the assessment should apply and disclose an expansion allowance per source type rather than be surprised later.
- **Present it as a table the DRD can absorb:** source → custodian → size → estimated items (range) → basis. That table is also the costs budget's disclosure assumption sheet.

Custodians

Assessment tests the custodian list before it is negotiated, using the data itself:

- **Correspondent analysis on a seed.** Index two or three obviously core mailboxes and rank their correspondents on the issues (by volume, by period, by counterparty domain). The ranking surfaces the heavy participants the org chart missed and demotes the seniors who barely feature.
- **Per-candidate footprint counts.** For each proposed custodian, in-place reports give issue-period item counts before any commitment: a proposed custodian with 400 items in the window is a different proposition from one with 40,000, in both directions.
- **Overlap estimation.** Hash and thread overlap between candidate custodians (measurable at metadata level) predicts how little unique material marginal custodians add, which is the evidence for resisting padded lists.
- **The output is a tiering evidence pack:** proposed list, per-candidate counts, correspondent rankings and overlap rates, feeding the tiering method in our companion guide *Identifying the Right Custodians* and the DRD negotiation directly.

File types

The file-type profile predicts processing complexity, review speed and exception load before any of them is bought:

- **Profile the population by extension and signature:** mail items, office documents, PDFs (split text-born from scanned), images, chat exports, spreadsheets, media, CAD and technical formats, archives, databases and unknowns.
- **Read the profile for cost signals:** a high scanned-PDF share means OCR volume and slower review; heavy spreadsheet presence means native review and non-keyword handling; chat-heavy populations need participant-and-date scoping rather than terms; exotic formats (CAD, proprietary systems) mean viewer costs and possible expert handling; large media volumes mean review-hour multiplication that budgets routinely miss.
- **Forecast exceptions early:** encrypted items, password-protected archives and corrupt files visible at metadata level predict the exception log; a 4 per cent encrypted rate discovered at assessment is a plan, discovered at processing a delay.
- **Feed it forward:** the profile prices processing realistically per source, shapes the review plan (who reviews spreadsheets, how chat is presented) and flags where forensic help will be needed before it is urgent.

Duplication

Duplication is the gap between collected volume and review burden, and it is measurable from hashes alone:

- **Compute exact-duplicate rates on the seed** (hash comparison at metadata level), within and across custodians, and near-duplicate and thread-repetition rates where level 3 or 4 assessment runs.
- **Typical patterns to test against:** cross-custodian mail duplication rises steeply with each added custodian on the same threads (which is why marginal custodians add less than their gigabytes suggest); project drives duplicate mail attachments; backup-derived data duplicates live data almost entirely. Measured rates of 30 to 60 per cent are routine, and the exact figure moves review budgets by six figures in volume matters.
- **Model the funnel with the measured rate:** collected items → post-dedupe → post-threading unique reading burden, per proposal on the table (each custodian-count and source-mix scenario gets its own funnel), so negotiations trade in review-hour consequences rather than raw counts.
- **Record the dedupe-scope decision now** (global versus per custodian) because it changes both the numbers and the evidential meaning of custodian fields later, as covered in the custodian guide.

Communication patterns

Metadata alone (senders, recipients, dates, channels) draws the shape of the case before anyone reads a document:

- **Timelines:** message volume over time, per custodian and per counterparty, shows the spikes around the pleaded events, the silences that need explaining, and whether the proposed date ranges actually bracket the activity. A range that clips a visible spike is redrawn at assessment, not litigated later.
- **Maps:** who-communicates-with-whom networks on the issues confirm or embarrass the custodian list, reveal the side channels (the director whose counterparty traffic runs through a personal domain), and identify the bridging individuals whose absence would break threads.
- **Channel mix:** the proportion of issue-period communication in mail versus chat versus shared mailboxes tells you where the search effort belongs and whether the mobile and messaging sources flagged in identification are marginal or central.
- **Anomaly signals:** a custodian whose volume drops to zero mid-period (departure? deletion?), counterparty traffic that migrates channels just before the disputed events, out-of-hours clusters: assessment does not conclude anything from these, but it routes them to the forensic escalation path early, which is when routing is cheap.

Privileged material

Privilege review is the most expensive review stripe per document; assessment forecasts its size and shapes its handling:

- **Screen by participants and domains:** counts of items to or from law-firm domains, named in-house lawyers and known advisers, per custodian and period, give a first privilege-population estimate from metadata alone.
- **Map the concentration:** privilege clusters (the GC's mailbox, the deal folder, the litigation period) can be carved into dedicated workflows, senior-reviewed, while the low-privilege bulk runs faster lanes; the assessment tells you where the clusters are.
- **Spot the hard categories early:** lawyers acting in commercial roles, advice forwarded into business threads, internal investigations, and board packs mixing advice with minutes: sampled reads at level 3 estimate how much judgment-heavy material sits in each, which is what actually drives privilege-review cost.
- **Feed protections forward:** the estimate sizes the privilege-review budget, justifies clawback terms in the DRD, and (where seed sampling shows privilege saturating a source) supports handling proposals such as lawyer-only review lanes agreed early rather than argued late.

Potentially relevant datasets

The estimate everything else serves: how big is the genuinely relevant core, and where does it live?

- **Test candidate terms on the seed** with full hit reports and sampled relevance reads, exactly as the search guide in this series prescribes, but before the CMC rather than after: the assessment's term forecasts are the negotiation's evidence.
- **Estimate richness per source:** sampled relevance rates by custodian and source (the core mailbox running at 18 per cent, the shared drive at 2 per cent) rank where review money should go and which sources justify staging or parking.
- **Model the review population per scenario:** for each combination of custodians, ranges and terms under discussion, the assessment produces a forecast population and review cost, which converts DRD negotiation from positional to computational.
- **Read a structured merits sample:** a few hundred documents from the richest cells, read by the case team, is the cheapest strategic intelligence in litigation: it informs Issues for Disclosure, settlement posture and witness planning while every option is still open, and it is where known adverse documents tend to first surface, engaging the PD 57AD duty early rather than awkwardly.

The ten-day EDA workflow and the assessment report

DAYS	WORK
1–2	In-place reporting across the estate (level 1): platform statistics per custodian and source; leaver and volatility checks alongside (preservation runs in parallel, never behind, assessment)
3–5	Targeted seed collection (typically 3 to 5 core custodians plus the key repository) under full chain of custody; metadata indexing (level 2); dedupe, type and timeline analytics
6–8	Sampled content work (level 3): term testing with hit reports, relevance and privilege sampling, communication maps on the issues
9–10	Scenario modelling, the assessment report, and the decision meeting that turns numbers into the DRD position, the budget and the staging plan

The assessment report is the phase's deliverable, and it should contain, as a minimum: the volume table per source and custodian with conversion bases; the file-type and exception profile; measured duplication rates and the funnel model per scenario; timeline and communication-map findings with any anomalies flagged; the privilege concentration estimate; term hit forecasts with sampled relevance; review-population and cost scenarios; and stated tolerances and blind spots (§16). Its annexes are the DRD's Section 2, the budget's assumptions and the negotiation's exhibits.

§ 1 3 · THE DECISION

The post-assessment decision tree

Run per source (and per custodian tier) once the report lands:

QUESTION 1

Does the measured richness and uniqueness of this source justify full processing now (core custodians, spiking periods, high sampled relevance)?

YES

Promote. The assessed seed data and settings move into full processing; nothing is re-collected or re-indexed.

NO

Continue to Q2.

QUESTION 2

Is it plausibly material but second-order (moderate richness, high overlap with promoted sources)?

YES

Stage. Preserve, hold at assessment level, and define in the DRD the trigger (review findings, surviving issues) on which it processes in tranche two.

NO

Continue to Q3.

QUESTION 3

Do the numbers show cost out of proportion to yield (low richness, near-total duplication, expensive accessibility)?

YES

Park, openly. Preserve; put the measured cost-versus-yield case in the DRD as a reasoned exclusion the court can endorse.

NO

Continue to Q4.

QUESTION 4

Did assessment flag an anomaly on this source (gaps, deletions, channel migration, encryption)?

YES

Escalate. Forensic preservation or examination per the escalation protocol before any further ordinary handling; the assessment finding is the instruction's scoping document.

NO

Preserve-only, and record why nothing further is proposed.

Figure 1 · Promote, stage, park, escalate: every source leaves assessment with one of four labels and a recorded basis, which is the staging plan, the DRD narrative and the budget's structure in one pass.

Worked example: Marwell v Casterbridge, with numbers

A fictional but representative shape. Marwell Logistics defends a £6m contract termination claim by Casterbridge Foods. Fourteen custodians are proposed by the claimant; Marwell's instinct is to agree and process everything: an estimated 610 GB. Instead, a ten-day assessment runs on a four-custodian seed (68 GB) plus in-place reporting across all fourteen.

WHAT THE ASSESSMENT MEASURED

In-place counts showed five of the fourteen proposed custodians held fewer than 900 issue-period items each, with hash-overlap against the core four above 80 per cent. The seed's file profile was mail-dominated but contained 9 per cent scanned PDFs (an OCR and slow-review flag) and a 3 per cent encrypted-container rate concentrated in one custodian (escalated). Measured cross-custodian duplication ran at 44 per cent. Timelines showed the dispute's traffic spiking August to November 2024, with the claimant-proposed range starting eighteen months earlier through a period of near-zero issue traffic. Term testing cut the claimant's 22 proposed terms to 14, three rewritten with proximity operators after "delivery" alone hit 96,000 items at a sampled relevance of 3 per cent. Privilege screening put 6 per cent of the seed on lawyer domains, concentrated in one mailbox. Sampled richness: core custodians 14 per cent, the shared operations mailbox 11 per cent, the proposed archive restoration under 1 per cent unique.

WHAT THE NUMBERS CHANGED

The DRD went in with: nine custodians (four core, five targeted), the five marginal names offered as preserve-only with the overlap data attached; a date range opening February 2024, with the earlier period staged behind a defined trigger; fourteen tested terms with hit reports exchanged; the archive restoration parked on measured cost-versus-yield (£54,000 estimated, under 1 per cent unique sampled relevance); and one custodian's devices routed to forensic examination over the encrypted containers. Forecast review population: 92,000 documents against the 380,000 the original proposal implied. The claimant challenged the custodian reduction; the overlap and footprint tables settled it in correspondence. The budget filed matched the forecast within 8 per cent at completion.

Lesson: ten days and a low five-figure assessment moved roughly a quarter of a million documents out of a review estimate, produced the evidence that won every scope negotiation, and surfaced the forensic issue while it was still cheap to handle. All figures are illustrative, but the shape is the routine one.

Data protection, privilege and the assessment itself

- **Assessment is minimisation in action, and should be run as such.** Measuring in place and ingesting only a seed is the proportionate architecture; the assessment plan should record the lawful basis, keep the seed as narrow as the questions allow, prefer metadata over content wherever content adds nothing, and host in appropriately secured, UK-resident environments with role-based access.
- **Chain of custody applies from the seed onward.** Assessment data becomes disclosure data on promotion, so it is collected forensically, hash-verified and logged from day three, not re-done later to a higher standard.
- **Privilege inside the assessment:** sampled reading is done by the legal team or under its supervision; the assessment report describes privilege in aggregates and locations, not by quoting borderline documents; and where the seed includes another entity's or a custodian's personal material, the protocols from the control and custodian guides in this series apply from the first index.
- **The report itself is litigation work product:** prepared for the dominant purpose of the proceedings, circulated accordingly, and drafted knowing that its numbers (though not its advice) will be deployed in the DRD and correspondence. Keep the deployable tables severable from the strategic commentary.

Common mistakes and technical limitations

Common mistakes

- **Skipping assessment to save time**, then spending the saved fortnight three times over in re-negotiated terms, revised budgets and re-run processing.
- **Assessment without preservation**. EDA is analysis, not a hold: volatile sources are preserved in parallel from day one, or the measured estate shrinks while being measured.
- **Letting the seed skew**. A seed of only the friendly custodians, or only mail, produces confident numbers about the wrong population; seed choice is a design decision, recorded with reasons.
- **Presenting estimates as counts**. Assessment numbers carry tolerances (expansion, exceptions, OCR yield); a DRD that states them as exact invites a completeness argument later. State ranges and bases.
- **Assessment as a one-off**. The numbers date: re-run the light reports when pleadings amend, custodians are added or tranche triggers are considered.
- **Buying assessment twice**. If the provider cannot promote assessed data into processing without re-ingestion, the phase's economics halve; confirm promotion mechanics before instruction.
- **Ignoring the anomalies**. The timeline gap and the encrypted cluster are the assessment's most valuable outputs when routed to the forensic path, and its most expensive when noted and shelved.

Technical limitations

- **Metadata sees shape, not meaning**. Levels 1 and 2 estimate volumes, types and patterns but cannot judge relevance; only sampled content (level 3) grounds richness and privilege rates, which is why the levels combine.
- **Platform estimates differ from processed counts**. In-place search statistics use the platform's own indexing and estimation, which will not match a processing platform item-for-item; treat level 1 numbers as scale, not census.
- **Exceptions hide until touched**. Encrypted, corrupt and exotic items are visible as flags at metadata level but their contents, and their cost, resolve only in processing; the assessment forecasts the exception load, it does not clear it.
- **Chat and structured data resist item counting**. Messages, records and rows are not documents; assessment reports these populations in their own units with a stated approach to how they will be reviewed and produced.
- **Samples have confidence bounds**. Small samples on rare categories (privilege in a low-lawyer population, relevance in a thin source) carry wide intervals; the report says so, and decisions on the margin get bigger samples.

Questions to ask · Suggested wording for instructions

Ask your client

- Who can pull platform statistics today: mailbox sizes, item counts, site inventories, chat volumes per custodian?
- Which three to five people were unarguably central, for the seed?
- What in the estate is volatile or scheduled for change while we assess (leavers, migrations, retention timers)?
- What did previous disputes or audits measure about this estate that we can reuse?

Ask your opponent

- What assessment underlies your proposals: volumes per custodian and source, and on what basis were they measured?
- Provide per-custodian issue-period item counts for your proposed list, and overlap analysis if performed.
- Have your proposed terms been tested against your own population, and will you exchange hit reports and sampled relevance rates?
- For sources you propose we should process (archives, backups): what is your estimate of unique yield, and its basis?

Ask an eDiscovery / forensic provider

- What does your assessment offering include at each level, priced how, and delivered in what timescale?
- Does assessed data promote directly into processing and review on your platform, or is there re-ingestion?
- What does the assessment report contain as standard, and can we see a specimen?
- How are anomalies (gaps, encryption, deletion signals) flagged and handed to the forensic side?
- What tolerances do you state on volume, dedupe and hit forecasts, and how have past forecasts tracked outcomes?

SUGGESTED WORDING · ASSESSMENT INSTRUCTION TO A PROVIDER (CORE PARAGRAPHS)

"You are instructed to carry out an early data assessment, in advance of any decision on full processing, comprising: (1) in-place reporting across the sources at Schedule 1 (per-custodian and per-source volumes and item counts); (2) forensically sound collection and metadata indexing of the seed sources at Schedule 2, under full chain of custody; (3) analysis and reporting of file-type and exception profiles, duplication rates within and across custodians, activity timelines and communication patterns on the issues, and privilege-domain concentrations; (4) testing of the candidate search terms at Schedule 3, with hit reports and sampled relevance rates; and (5) review-population and cost scenarios for the custodian, range and term combinations we identify. Deliverable: a written assessment report with stated tolerances and bases, in a form deployable (as to its tables) in the Disclosure Review Document and costs budget. Assessed data and settings are to be capable of promotion into full processing without re-collection or re-ingestion. Preservation steps continue in parallel and are not to await assessment."

SUGGESTED WORDING · DRD NARRATIVE BUILT ON ASSESSMENT (CORE PARAGRAPHS)

"The Defendant's proposals are based on an early data assessment completed on [date], the volume, duplication and search-term tables from which are annexed. In summary: the estate comprises [n] GB across the sources listed, an estimated [range] documents after de-duplication (measured cross-custodian duplication: [x] per cent); the [m] proposed custodians hold [counts] issue-period items each, and the [k] further individuals proposed by the Claimant hold [counts], with hash-overlap against the proposed custodians of [y] per cent, on which basis preservation without collection is proposed for them; the annexed terms have been tested with hit reports and sampled relevance rates as shown; and [source] is proposed for staged treatment / reasoned exclusion on the measured cost and yield figures annexed. The Defendant will update these figures if the issues or custodian list change materially."

§ 18 · QUICK CONTROL

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

The assessment checklist

- ☐ Preservation running in parallel from day one; volatile sources secured before or during, never after, assessment
- ☐ In-place reports pulled across the full estate; per-custodian issue-period counts obtained
- ☐ Seed chosen deliberately (core custodians plus key repository), recorded with reasons, collected forensically
- ☐ All seven estimates produced: volume, custodians, types, duplication, patterns, privilege, relevant core
- ☐ Terms tested with hit reports and sampled relevance before any agreement
- ☐ Scenarios modelled per live proposal; funnel and cost per scenario
- ☐ Anomalies logged and routed to the forensic path, not shelved
- ☐ Assessment report delivered with tolerances and bases stated; deployable tables severable from strategic commentary
- ☐ Every source labelled promote / stage / park / escalate with recorded basis
- ☐ DRD Section 2, costs budget and staging plan drafted from the report; re-assessment triggers diarised
- ☐ Promotion mechanics confirmed: assessed data flows into processing without re-work

Red flags

- Any budget or DRD volume figure that nobody can trace to a measurement.
- An opponent's proposals unaccompanied by counts, overlap data or tested terms, resisting requests for them.
- Assessment findings (a timeline gap, an encrypted cluster, a channel migration) noted in the report and absent from any instruction.
- A seed that quietly became "the custodians whose data was easiest to get".
- Estimates hardening into exact counts as they travel from report to DRD to submission.
- A provider whose assessment platform cannot promote into its processing platform.
- Preservation waiting politely behind the assessment timetable.

When to involve a digital forensic expert

Assessment findings are frequently the first sight of forensic issues, and the hand-off should be immediate: timeline gaps and mid-period silences (possible deletion), encrypted containers and password-protected clusters, channel migrations before disputed events, custodians whose in-place counts fall far below their role's plausible footprint, and any source whose measured state conflicts with the client's account. Each is an examination question, scoped by the assessment data itself, and each is cheapest at exactly this stage. The escalation machinery, and the wider triage between disciplines, are covered in our companion guides *eDisclosure or Digital Forensics?* and *The EDRM for UK Litigation Teams*.

§ 19 · COMMON QUESTIONS

Frequently asked questions

Why does early data assessment matter?

Because the largest disclosure decisions are otherwise taken blind. Custodian lists, search scope, processing commitments, budgets and DRD positions are all fixed early; assessment measures the estate first (volumes, duplication, file types, communication patterns, privilege concentration and likely relevant core) so those decisions rest on numbers, at a cost that is a small fraction of what any one bad blind decision wastes.

What does it cost, and how long does it take?

Typically ten working days in the workflow this guide sets out, at a low single-digit percentage of the projected disclosure budget, scaling with seed size and levels used. In-place reporting alone (level 1) is near-free and takes hours; it is the minimum no matter how small the case.

Is assessment wasted work if we process everything anyway?

No, provided promotion is confirmed at instruction: the seed's forensic collection, indexing and analytics carry into full processing, the tested terms and settings apply directly, and the report remains the budget's and DRD's evidence base. What assessment prevents is paying to process what the numbers would have parked.

Can we assess before proceedings are issued?

Yes, and it is often the ideal time: pre-action assessment sizes the dispute for protocol correspondence, informs whether and how to plead, supports early settlement judgments, and starts the preservation and chain-of-custody record early. The same minimisation disciplines apply, and the work product protections should be set up deliberately.

Will the court or our opponent see the assessment?

The deployable tables, yes, by design: volumes, counts, overlap rates, hit reports and cost figures are exactly what the DRD and proportionality argument need. The strategic commentary and merits sampling remain privileged work product; the report should be structured from the outset so the two separate cleanly.

How accurate are the estimates?

Accurate enough for the decisions they serve, if tolerances are respected: volumes and duplication measure tightly; document-count conversions and exception loads carry stated ranges; richness and privilege rates depend on sample sizes the report discloses. The discipline is to present ranges with bases, decide on the ranges, and re-measure when the case moves.

Glossary

Assessment report

The EDA deliverable: measured volumes, profiles, rates, forecasts and scenarios, with tolerances and bases stated.

Communication map

A network analysis of who exchanged messages with whom on the issues, drawn from metadata.

ECA / EDA

Early case assessment / early data assessment: light-touch measurement of an estate before full processing and review commitments.

Expansion factor

The growth from collected to processed volume as containers and embedded items are extracted.

Hash overlap

The proportion of one custodian's items duplicated in others' collections; the measure of marginal custodians' unique yield.

In-place reporting

Platform statistics gathered without moving data: sizes, counts and search estimates from the tenancy's own tools.

Metadata-only indexing

Ingesting dates, participants, types, sizes and hashes without full text extraction; the engine of level 2 assessment.

Promotion

Moving assessed data and settings into full processing without re-collection or re-ingestion.

Richness

The sampled rate of relevant documents in a population or source; the ranking metric for review spend.

Seed

The deliberately chosen core sources collected for assessment, from which rates and forecasts are measured.

Scenario model

The forecast review population and cost for a given combination of custodians, ranges and terms.

Tolerance

The stated range around an assessment estimate, reflecting expansion, exceptions and sample confidence.

Sources and authoritative references

Assessment and culling practice in this guide draws on materials published by our e-discovery practice and laboratory, referenced here as sources:

- e-discovery.uk, EDRM Phase 04 · Processing ("Review is where budgets are spent. Processing is where budgets are made"; terms iterated against samples; hit and exception reporting; suppressions reversible and audited): e-discovery.uk/edrm/processing
- e-discovery.uk, practice principles including proportionality ("we aggressively cull noise before it reaches review; you pay for substance, not volume") and EDRM Phase 01 · Identification: e-discovery.uk · e-discovery.uk/edrm/identification
- e-discovery.uk, Knowledge Centre practice notes, including "Writing search terms that survive negotiation" and the case note on negotiated terms reducing a review set by 71 per cent: e-discovery.uk/knowledge · e-discovery.uk/expertise
- cflab.uk, digital forensics services (laboratory handling of the encrypted, damaged and deleted material that assessment flags but cannot open): cflab.uk/digital-forensics-services
- cflab.uk, guides for lawyers, including the treatment of deleted and unallocated data and chronology anomalies that assessment findings feed into: cflab.uk/guides
- Practice Direction 57AD and the Disclosure Review Document (data descriptions, cost estimates, Models and proportionality): justice.gov.uk, [PD 57AD](https://justice.gov.uk/pd-57ad)
- CPR Part 31 and PD 31B (reasonable search factors; pre-CMC discussions; the Electronic Documents Questionnaire): justice.gov.uk, [PD 31B](https://justice.gov.uk/pd-31b)
- CPR Part 3 and Practice Direction 3D–3E territory: costs management and Precedent H budgeting: justice.gov.uk, [Part 3](https://justice.gov.uk/part-3)

- Attorney General's Guidelines on Disclosure (rev. 2024: digital material strategy and Disclosure Management Documents): [gov.uk](https://www.gov.uk/government/publications/attorney-general-guidelines-on-disclosure), [AG's Guidelines](#)
- ICO, UK GDPR guidance including data minimisation: ico.org.uk · Data Protection Act 2018: legislation.gov.uk/ukpga/2018/12 · Data (Use and Access) Act 2025: legislation.gov.uk/ukpga/2025/21

DISCLAIMER

This publication provides general information about early data assessment in United Kingdom disclosure practice as at August 2026. *Marwell v Casterbridge* is a fictional worked example and all figures are illustrative. 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. Rules, guidance and legislation change; links were live at the date of publication.

§ HOW A SPECIALIST LABORATORY CAN ASSIST

Working with Computer Forensics Lab

Assessment is how our e-discovery practice (**e-discovery.uk**) prefers every substantial matter to begin, because it is how culling gets aggressive safely and clients pay for substance rather than volume: in-place reporting across tenancies, forensically sound seed collection that promotes straight into processing, the seven estimates of this guide delivered as a deployable assessment report, and scenario modelling that turns DRD negotiation into arithmetic. The laboratory side receives what assessment flags: encrypted clusters, timeline gaps, deletion signals and devices whose measured state contradicts the account, examined under chain of custody and reported under CPR Part 35 / CrimPR 19 where needed. Fixed-fee assessment packages, confidential scoping discussions and conflicts checks are routine, and we will tell you candidly when an estate is small enough that level 1 reporting and a sensible protocol are all it needs.



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