COMPUTER FORENSICS LAB

London digital forensics & e-discovery · Established 2007

What Is a Reasonable Electronic Search?

Custodians, Dates, Keywords, Analytics and Defensibility

The law does not demand a perfect search; it demands a reasonable and proportionate one that can be described, defended and tested. This guide works through the seven levers that define a search (custodians, date ranges, sources, keywords, analytics, proportionality and sampling), shows what overly narrow and excessively broad searches look like in practice, and sets out the documentation that makes a search defensible at the CMC and beyond.

⏱ Estimated reading time: 30 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**, treats search design as a negotiated, evidence-driven discipline: search terms iterated against samples rather than run once, hit reports produced at every step, and culling that compresses review populations defensibly, with suppressions logged and reversible. In one recent matter, negotiated terms reduced a personal-device review set by 71 per cent. Our examiners also produce CPR Part 35 and CrimPR Part 19 compliant expert reports where methodology itself is challenged.

ESTABLISHED 2007 · LONDON | ISO 17025-ALIGNED PROCEDURES | ISO 27001-ALIGNED SECURITY | ACPO / NPCC DIGITAL EVIDENCE PRINCIPLES | SEARCH-TERM NEGOTIATION SUPPORT | TAR / CAL WITH STATISTICAL VALIDATION | FULL CHAIN-OF-CUSTODY DOCUMENTATION

§ CONTENTS

In this guide

01 Executive summary
02 The problem in plain English: two ways to fail
03 The legal standard: CPR 31.7, PD 31B and PD 57AD
04 Lever 1 · Custodians
05 Lever 2 · Date ranges
06 Lever 3 · Sources
07 Lever 4 · Keywords
08 Lever 5 · Analytics and technology-assisted review
09 Lever 6 · Proportionality
10 Lever 7 · Sampling and validation
11 Worked examples: the overly narrow search
12 Worked examples: the excessively broad search
13 Defensibility: the paper trail behind the search
14 Negotiating search terms with the other side
15 Common mistakes and technical limitations
16 Questions to ask · Suggested wording for instructions
17 Checklist and red flags · When to involve a digital forensic expert
18 Frequently asked questions
19 Glossary · References · Disclaimer · How a specialist laboratory can assist

§ 0 1 · ORIENTATION

Executive summary

THE HEADLINE ANSWER: WHAT IS A REASONABLE SEARCH?

A reasonable electronic search is one whose scope (custodians, date ranges, sources) and method (keywords, analytics, review approach) are rationally connected to the issues, proportionate to the factors in CPR 31.7 and PD 31B (the number of documents, the nature and complexity of the proceedings, the ease and expense of retrieval, and the significance of what a search may yield), tested rather than assumed (sampled hit reports, iterated terms, validated stopping points), and documented well enough to be described in the disclosure statement or DRD and defended when challenged. The law does not require every relevant document to be found; it requires a search a court can approve and an opponent cannot credibly call either a keyhole or a landslide.

- **Reasonableness has two failure modes, and both are expensive.** Too narrow: missed documents, certificate exposure, redone searches at your cost. Too broad: unreviewable populations, exploded budgets, data-protection over-collection, and no extra credit from the court.
- **Seven levers define every search.** Custodians, date ranges, sources, keywords, analytics, proportionality and sampling. Reasonableness is the combined setting of all seven, chosen deliberately and recorded, not a property of any one of them.
- **Keywords are a tool, not a search.** PD 31B warns in terms that keyword searches alone may be inadequate; untested terms miss relevant material and drown teams in noise. Terms earn their place through sampled hit reports and iteration, and they work alongside date, custodian and analytic culling, not instead of it.
- **Validation is what converts a search into a defensible one.** Null-set sampling, richness curves and documented stopping decisions are the difference between "we searched" and "we can show the search worked". English courts have accepted analytics-driven approaches since *Pyrrho* (2016); what they scrutinise is the statistics around the method.
- **The paper trail is half the obligation.** The disclosure statement and DRD must describe the search; hit reports, the search protocol and the methodology memo are what make that description true, and they are produced, not reconstructed, when the challenge comes.

§ 0 2 · FIRST PRINCIPLES

The problem in plain English: two ways to fail

No electronic search finds everything. Language is ambiguous, OCR is imperfect, people misspell, chat is written in fragments and emoji, code words exist, and relevant ideas are expressed without any predictable vocabulary at all. The rules accept this: the duty is a *reasonable* search, and the disclosure statement expressly contemplates limits, provided they are stated. The practical craft is steering between the two failure modes:

THE KEYHOLE (OVERLY NARROW)	THE LANDSLIDE (EXCESSIVELY BROAD)
Two custodians for a five-person deal team; twelve months for a three-year relationship; corporate email only for a business that runs on WhatsApp; terms so precise they match nothing but the pleadings' own vocabulary	Every employee "to be safe"; the relationship's full decade; every backup restored; single common words as terms ("contract", "delay", "spec"); no deduplication strategy; review budgets consumed by noise
How it surfaces: the opponent's disclosure contains your client's own documents that your search missed; specific disclosure applications; a certificate that described a search wider than the one performed	**How it surfaces:** costs budgets rejected, review overruns, missed deadlines, data-minimisation complaints, and a court unimpressed that volume was substituted for thought
The underlying error: treating scope as a cost decision only	**The underlying error:** treating breadth as a safety decision only

The escape from both is the same discipline: connect every setting to the issues, test it against the data, and record why it sits where it does. A narrow search that was tested and reasoned survives; a broad one that was reflexive does not even buy safety, because breadth without method still misses the misspelled, the imaged and the coded.

§ 0 3 · THE LAW

The legal standard: CPR 31.7, PD 31B and PD 57AD

- **CPR 31.7** requires a reasonable search for documents falling within standard disclosure, and lists the factors bearing on reasonableness: the number of documents involved; the nature and complexity of the proceedings; the ease and expense of retrieval of any particular document; and the significance of any document likely to be located. Limits placed on the search must be stated in the disclosure statement (CPR 31.10), which is what makes search design a certified, personal matter.
- **PD 31B** translates the factors for electronic documents: it adds considerations including the accessibility and location of documents and systems, the likelihood of locating relevant data, the cost of recovery and of disclosure, and the availability of documents from other sources; it directs the parties to discuss searches, tools and keyword strategy before the first CMC; and it warns expressly that keyword searches alone may be insufficient and can both over- and under-include, so their use should be discussed and, where possible, agreed.
- **PD 57AD** builds the same idea into its Models: Model D orders a party to undertake a reasonable and proportionate search in relation to the Issues for Disclosure, and Section 2 of the DRD is where custodians, date ranges, sources, terms, analytics and TAR are proposed, negotiated and recorded. The duty to disclose known adverse documents sits alongside and does not depend on any search at all: a known harmful document is disclosable whether or not any term would have caught it.
- **Criminal practice** runs on the same logic under different labels: the reasonable-lines-of-enquiry duty, the Attorney General's Guidelines' treatment of digital material (including search strategy and its documentation in a Disclosure Management Document), and defence engagement on terms and parameters. The vocabulary of this guide transfers directly.

> FOR THE PRACTITIONER
>
> Read the disclosure statement wording before designing the search, not after. It requires the signatory to state the extent of the search, draw attention to its limits, and certify the duty is understood. Every lever setting in §§4 to 10 ends up summarised in that statement or in the DRD; designing the search is drafting the certificate.

§ 0 4 · LEVER 1

Custodians

Who is searched is the largest scope decision and the largest cost decision, and it is made on participation, not seniority: the issues identify events, events identify roles, roles identify people, and interviews plus correspondent analysis test the list. Tier the result (core, targeted, preserve-only) so effort matches involvement, and record every exclusion with a one-line reason. Our companion guide *Identifying the Right Custodians in Electronic Disclosure* covers the method, the difficult categories (shared mailboxes, leavers, organisational data) and the interview set in full; for search design the essentials are:

- **Too narrow looks like:** the deal's two signatories searched, the project manager who wrote everything omitted; the predecessor in a role ignored; shared mailboxes assumed covered by individuals.
- **Too broad looks like:** custodian counts padded with everyone ever copied, multiplying collection and review with material that already exists in the participants' files.
- **Test it with the data:** correspondent analysis on early collections shows who actually traded messages on the issues; run it before the DRD is agreed and again after first-pass review.

§ 0 5 · LEVER 2

Date ranges

Anchor ranges to the pleaded events, not the relationship. The defensible pattern is a core window around the events in issue, extended where a specific issue justifies it, and different ranges for different issues or custodians where the case genuinely splits that way (PD 57AD's DRD accommodates per-issue ranges).

- **Too narrow looks like:** a range starting at the contract date in a misrepresentation claim (the representations predate it); ending at issue of proceedings when quantum and mitigation run on; excluding the negotiation period that explains the disputed clause.
- **Too broad looks like:** "2015 to date" because the parties have dealt since 2015, sweeping in seven irrelevant years of routine trade at review prices.
- **Craft points:** use event anchors with margins (three months either side of the disputed variation, not calendar years); remember date metadata is imperfect (migrated data, wrong-clock devices, undated attachments travel with dated parents), so apply ranges at the family level and treat boundary weeks generously; and state the range per issue in the DRD so extensions are negotiations, not concessions.

§ 0 6 · LEVER 3

Sources

Which repositories are searched is where accessibility economics live: PD 31B expressly makes the accessibility and retrieval cost of particular sources part of the reasonableness assessment. The defensible structure is a source list divided into searched, preserved-but-not-searched (with the trigger that would change that), and excluded with reasons.

- **Too narrow looks like:** corporate email and the file server, in a business whose custodian interviews disclose Teams, WhatsApp and a project SharePoint; the leaver's archived mailbox unlisted; the shared mailbox where the counterparty correspondence actually sits.
- **Too broad looks like:** restoring every backup tape as a first step; collecting whole drives where the project subtree is identifiable; treating the ERP as documents rather than agreeing the four reports that answer the issues.
- **Craft points:** backups are preserve-first, restore-on-demonstrated-need (typically where deletion is in issue and live sources cannot answer); inaccessible legacy systems are parked openly in the DRD with the cost evidence attached; and the search description should say per source what tool searched it and what that tool cannot see. The full source landscape, category by category, is mapped in our companion guide *Finding the Evidence*.

§ 0 7 · LEVER 4

Keywords

Keywords are the most negotiated and most misunderstood lever. They are powerful for culling and dangerous as a proxy for relevance, which is exactly the balance PD 31B strikes. Working craft:

- **Build terms from the data's vocabulary, not the pleadings'.** Read a sample of core documents first: the parties' own shorthand (project codenames, product references, nicknames, the misspelling everyone repeats) outperforms the statement of case's formal language every time.
- **Test before you agree.** Run every candidate term against the processed population and read a sample of hits: a term is kept, refined or dropped on evidence. Hit reports (documents hit, plus families, unique hits per term) are exchanged and iterated; terms are negotiated against samples rather than run once. A term with 190,000 hits is not a term; it is a date range in disguise.
- **Use the syntax deliberately.** Wildcards and stemming for word families (deliver* catches delivery, delivered, deliverables); proximity operators to bind ambiguous pairs ("termination w/10 notice" rather than both words anywhere); phrase quoting for terms of art; and exclusion logic used sparingly and visibly, because silent NOT clauses are where opponents find suppression arguments.
- **Cover the failure modes:** misspellings and variants for names (Halworth, Hallworth, HW), transliterations, product code formats with and without separators, and the chat problem (fragments, abbreviations, emoji reactions carrying meaning) which terms alone cannot solve and which pushes chat toward date-and-participant scoping plus review rather than keyword culling.
- **Know what keywords cannot reach:** images of text without OCR, audio and video, spreadsheets whose meaning is numeric, foreign-language material without translated terms, and the relevant document that simply uses none of the predicted words. This is why terms sit inside a method (analytics, sampling, targeted review) rather than constituting one.

§ 0 8 · LEVER 5

Analytics and technology-assisted review

Analytics are the reasonableness multipliers: they cut volume without cutting scope, and they find what keywords miss.

- **Structural analytics** (deduplication, email threading, near-duplicate grouping) remove repetition, not content: reviewers read each conversation once, at its inclusive message, with suppressed items logged and reversible. These are uncontroversial and should be default in any volume matter.
- **Conceptual analytics** (clustering, concept search, communication mapping) explore the population: clusters surface themes no term predicted, and communication maps test custodian lists and date ranges against reality.
- **Technology-assisted review** (predictive coding, continuous active learning) ranks documents by likely relevance, trained on lawyers' own decisions, and has had English judicial approval since *Pyrrho Investments v MWB Property* [2016] EWHC 256 (Ch). The DRD asks parties whether they will use it; in heavy cases, expect to justify not using it. Its defensibility is statistical (training transparency, validation sampling, a recorded stopping decision), not mystical.
- **Reasonableness framing:** a keyword-culled linear review of 40,000 documents and a CAL-prioritised review of 170,000 can both be reasonable; what makes either defensible is that the choice was made on tested numbers and the residual risk was measured (§10) rather than assumed.

§ 0 9 · LEVER 6

Proportionality

Proportionality is the balancing that sets all the other levers: the value and importance of the case, the significance of what a source or search may yield, and the cost of retrieval and review, weighed openly. Practically:

- **Put numbers on both sides.** "Restoring the 2019 tapes costs £38,000 and the live archive already covers 94 per cent of that custodian's mail" wins arguments that adjectives lose. Provider estimates per source, and review-cost projections per thousand documents, belong in the DRD discussion and the costs budget.
- **Stage the exercise.** Core custodians and sources first; extensions only if the issues survive and the first tranche shows gaps. Staging converts a proportionality argument into a timetable, which courts like.
- **Remember proportionality binds breadth too.** Data minimisation under UK GDPR and the costs rules both cut against reflexive over-collection; a party demanding the landslide can be met with the same factors it would deploy against the keyhole.
- **Significance is issue-relative.** A source marginal to liability may be central to quantum or conduct; the DRD's per-issue structure exists so the balance is struck issue by issue rather than case-wide.

§ 10 · LEVER 7

Sampling and validation

Sampling is how a search proves it worked, and it runs at three points:

- **Before agreement: term testing.** Random samples of each candidate term's hits (does it find relevant material?) and of the population it would exclude (what would we lose?). This is the evidence behind every term kept or dropped, and the answer to an opponent's speculative additions.
- **During review: quality control.** Second-level sampling of coded documents, and richness tracking in ranked review (the falling rate at which the stream yields relevant documents) which signals when continued review stops earning its cost.
- **At the end: the null set.** A random sample of what was never reviewed (term non-hits, low-ranked residue) measures what the method left behind. A documented low residual rate (with the sample size and confidence recorded) is the stopping decision's justification and the certificate's backbone; a high one sends you back to the levers, cheaply and early rather than by court order later.

> FOR THE PRACTITIONER
>
> You do not need to become a statistician; you need to insist the numbers exist and are recorded: sample sizes, what was sampled, the rate found, and who decided what on the strength of it. "A 1,500-document random sample of the unreviewed population found a 0.4 per cent residual relevance rate, reviewed and assessed as immaterial to the issues" is one sentence, and it is the sentence that ends most search challenges.

§ 11 · IN THE WILD

Worked examples: the overly narrow search

EXAMPLE 1 · THE PLEADINGS VOCABULARY TRAP

In a defective-machinery claim, the defendant's terms were built from the pleadings: "defect", "failure", "breach", "warranty". Hit counts looked sensible (11,000 documents) and review was quick. The claimant's disclosure then produced the defendant's own internal emails calling the problem "the gremlin", "the Tuesday issue" and "V2 wobble", none of which any term touched. A sample read of core documents before term design would have surfaced the house vocabulary in an afternoon; instead the defendant re-ran the exercise under a consent order, at its own cost, with an unhappy paragraph in the next certificate.

EXAMPLE 2 · THE RANGE THAT STARTED AT THE CONTRACT

A franchise dispute turned on pre-contract representations. The franchisor searched from the franchise agreement's date forward, reasoning that the claim was about performance. The representations, of course, lived in the nine months before signature. The gap was visible on the face of the DRD, the franchisee's solicitors said so at the CMC, and the court ordered the extended search with an indemnity-costs warning attached. The range had never been mapped to the issues; it had been copied from the previous case's DRD.

EXAMPLE 3 · THE MISSING CHANNEL

An agency's two directors were properly selected as custodians, but the search covered corporate email only. Custodian interviews had never asked about messaging; the directors, it emerged mid-trial through the opponent's evidence, ran the relevant client relationship almost entirely over WhatsApp. The tribunal's observations about the search shaped its view of everything else the agency said. One interview question (channel usage, per custodian) and one targeted handset extraction would have cost a few hundred pounds.

§ 12 · IN THE WILD

Worked examples: the excessively broad search

EXAMPLE 1 · THE SINGLE-WORD TERM LIST

Negotiating under time pressure, the parties in a supply dispute agreed sixteen terms including "order", "invoice", "delay" and "quality", applied across nine custodians and six years. The hit population was 1.4 million documents plus families. Nobody had run the terms before agreeing them; the first hit report arrived after the order was made. The review budget collapsed in month two, the timetable moved twice, and the eventual fix (proximity operators, per-issue ranges, CAL prioritisation) was what a half-day sampling exercise would have produced before the CMC, without the two wasted months.

EXAMPLE 2 · RESTORE EVERYTHING, JUST IN CASE

Fearing criticism, a defendant restored 214 monthly backup tapes spanning five years before any live-source analysis, at six figures, producing terabytes that were 97 per cent duplicative of the live archive once deduplicated. The court's comment at the costs hearing was not gratitude for thoroughness but a question about why restoration preceded the deduplication analysis that would have confined it to the four months where live data had gaps. Breadth had been purchased as insurance and priced as waste.

EXAMPLE 3 · EVERYONE, TO BE SAFE

An employer facing a discrimination claim proposed itself as a model respondent by offering 41 custodians (the claimant's whole department and every HR user). Collection and processing consumed the budget that review needed; reviewers then spent weeks coding canteen rotas and leave requests; and the claimant's advisers, quite reasonably, treated the deluge as an obstruction tactic and sought (and won) a focused re-exercise around the eight people who actually featured in the pleaded events. Volume substituted for analysis, and bought neither safety nor goodwill.

§ 1 3 · THE PAPER TRAIL

Defensibility: the paper trail behind the search

A reasonable search generates its own evidence as it goes. The file that answers any challenge contains:

DOCUMENT	WHAT IT RECORDS
Search protocol	The design: custodians and tiers, per-issue date ranges, source list (searched / preserved / excluded with reasons), term list with syntax, analytics and TAR approach, sampling plan. Versioned as it evolves.
Hit reports	Per term, per iteration: documents hit, with families, unique hits; exchanged in negotiation and filed with dates.
Sampling records	What was sampled, sizes, rates found, decisions taken; including the null-set result behind the stopping decision.
Methodology memo	The narrative: each material decision, its date, its reason, its decider; the single most valuable document in the exercise per hour spent writing it.
Processing and exception reports	Volumes in and out of each culling step; unsearchable items (encrypted, corrupt, no-text) and how each was resolved, because a search description that ignores the exception log is incomplete.
DRD / disclosure statement text	The public summary of all the above, drafted from the protocol rather than from memory, with limits stated as CPR 31.10 requires.

The habit that makes this cheap: write the protocol first and let it drive the work, rather than performing the work and reverse-engineering a protocol for the certificate. Suppressions logged and reversible, volumes reported at each step, and terms iterated against samples are what allow the eventual one-line answer to any challenge: here is the document.

§ 14 · ACROSS THE TABLE

Negotiating search terms with the other side

- **Exchange proposals early, with evidence.** Send terms with hit counts attached; ask for the same. A proposal without numbers is an opening position; a proposal with numbers is a negotiation.
- **Counter with data, not adjectives.** "Your term 'agreement' hits 412,000 documents of which a 200-document sample found two per cent arguably relevant; we propose 'agreement w/15 [counterparty / project term]', which hits 9,300 with a sampled relevance of 61 per cent" is how disputes end before the CMC.
- **Trade breadth for method.** Accepting an opponent's wider term is often cheap if paired with CAL prioritisation and a null-set sample; offering the trade shows the court a party engaging with substance.
- **Keep the record.** Every proposal, counter and hit report exchanged is part of the defensibility file, and conduct in the negotiation is itself something courts notice in both directions.
- **Escalate cleanly.** Genuine disputes go to the CMC as short, numbered issues with the sampling evidence attached; judges resolve tested disagreements quickly and untested ones badly.

§ 15 · WHERE IT GOES WRONG

Common mistakes and technical limitations

Common mistakes

- **Terms agreed before terms tested.** The single most common and most avoidable failure; everything in Example 12.1 follows from it.
- **Search designed from the pleadings' vocabulary** rather than the data's.
- **Keyword monoculture:** terms asked to do the work of dates, custodians, analytics and review combined, despite PD 31B's express warning.
- **Ranges copied between cases** instead of mapped to the issues in this one.
- **The exception log ignored:** encrypted and no-text items silently outside every term, and outside the certificate's description too.
- **No stopping evidence:** review halted when the budget ran out, with nothing measuring what remained.
- **Silent narrowing:** NOT clauses, domain exclusions and de-prioritisations applied without appearing in the protocol or the DRD.
- **The certificate written from memory,** describing the search as designed rather than as performed.

Technical limitations

- **Recall and precision trade off.** Widening terms to miss less always sweeps in more noise; the balance is managed with sampling, never abolished.
- **OCR is probabilistic.** Scanned and image-only documents search only as well as their OCR; quality thresholds belong in the processing report and poor-OCR batches may need review outside the terms.
- **Chat defeats keywords.** Fragmented, abbreviated, emoji-laden and threaded, chat is best scoped by participants and dates and then reviewed, not term-culled.
- **Language and transliteration:** terms must exist in every working language of the population, and names transliterate inconsistently; language identification in processing routes material correctly.
- **Numbers do not keyword.** Spreadsheets, ledgers and structured data answer to report-based extraction and targeted review, not to term lists.
- **Search indexes have edges.** Every platform has file types it cannot index and field-search quirks; the provider should state per source what the index covers, and the protocol should repeat it.

§ 16 · INTERROGATORIES & DRAFTING AIDS

Questions to ask · Suggested wording for instructions

Ask your client

- What did you actually call this project, product, problem and counterparty internally? Nicknames, codenames, abbreviations, habitual misspellings?
- Which channels did the relevant people really use, and in which languages?
- What date anchors the events per issue, and what preceded them that explains them?
- Which sources would be expensive to search, and what do they plausibly add over the live systems?

Ask your opponent

- Provide your proposed terms with hit reports (documents and families, unique hits per term) against your processed population, and your sampling results.
- State your date ranges per issue and the anchor events behind them.
- Which sources are searched, preserved-only and excluded, with the cost evidence for any accessibility exclusion?
- What analytics and TAR will you use, and how will the stopping decision be validated and recorded?
- How were unsearchable exceptions (encrypted, corrupt, image-only) handled, and do the terms' results include or exclude them?

Ask an eDiscovery provider

- What does your term-testing cycle look like: turnaround per iteration, sample sizes, report formats we can exchange?
- What syntax does the platform support (proximity, stemming, fuzzy matching), and what are its indexing limits per file type?
- How is CAL validated on your platform: richness tracking, elusion or null-set sampling, and what documentation do we receive?
- How are suppressions and de-prioritisations logged and reversed?
- Can you support the negotiation directly: joint hit reports, neutral samples, attendance at the CMC if methodology is challenged?

SUGGESTED WORDING · SEARCH PROTOCOL PREAMBLE (CORE PARAGRAPHS)

"The search will be conducted in accordance with this protocol, which will be versioned and retained. Scope: the custodians, date ranges and sources at Schedule 1, with exclusions and their reasons recorded. Terms: the terms at Schedule 2 will be applied only after testing; for each term and iteration, hit reports (documents, families, unique hits) and sample-review results will be recorded and, where agreed, exchanged. Analytics: deduplication [global / per custodian], email threading and near-duplicate grouping will be applied, with suppressions logged and reversible; [continuous active learning] will prioritise review, trained and validated as set out at Schedule 3. Validation: a random sample of not fewer than [1,500] documents from the unreviewed population will be reviewed before the exercise closes, and the residual rate recorded with the stopping decision. Exceptions: unsearchable items will be logged, resolved or accepted with reasons, and reflected in the description of the search. All limits on the search will be stated in the disclosure statement / DRD."

SUGGESTED WORDING · PROPOSING TERMS TO THE OTHER SIDE

"We enclose our proposed search terms at Schedule A, together with hit reports showing, for each term against our processed population of [n] documents: total hits including families, and unique hits. Sample reviews of terms [x] and [y] are summarised at Schedule B. We invite your agreement or reasoned counterproposals supported by equivalent hit data against your population. We are content to test any additional terms you propose and to exchange results before either side takes a position, and we suggest the parties aim to agree a consolidated list [14] days before the CMC, identifying any disputed terms, with their hit data, for determination."

§ 17 · QUICK CONTROL

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

The reasonable-search checklist

- ☐ Issues mapped to events, custodians, ranges and sources; every setting reasoned and recorded
- ☐ Core-document sample read before term design; house vocabulary captured
- ☐ Every term tested with hit reports and sample reads before agreement; iterations filed
- ☐ Per-issue date ranges anchored to events, applied at family level
- ☐ Source list split into searched / preserved / excluded, with cost evidence for accessibility calls
- ☐ Analytics defaults on (dedupe, threading, near-dupe), suppressions logged and reversible
- ☐ TAR decision made on numbers; validation and stopping plan written before review starts
- ☐ Exception log reviewed; unsearchable items resolved or accepted with reasons
- ☐ Null-set sample completed and recorded before the exercise closes
- ☐ Search protocol, hit reports, sampling records and methodology memo on the file
- ☐ Disclosure statement / DRD drafted from the protocol, limits stated

Red flags

- Any term list, yours or theirs, that has never met a hit report.
- Terms that are single common words, or that mirror the pleadings' vocabulary exactly.
- Ranges that start at the contract in a pre-contract dispute, or end at issue in a continuing-loss one.
- A DRD silent on chat and mobile sources for custodians who plainly used them.
- An opponent's "comprehensive search" with no sampling, no exception handling and no stated limits.
- Review stopped at a round number with nothing measuring the residue.
- Your own team narrowing anything silently under budget pressure.

When to involve a digital forensic expert

Search design becomes expert territory when: the methodology itself is challenged and needs independent evidence (CPR Part 35 reports on search adequacy and sampling); unsearchable material matters (encrypted stores, damaged media, legacy formats needing laboratory access before any term can run); deletion is in issue, so the reasonable search extends to recoverable data and its interpretation; or chat and mobile evidence must be extracted completely before participant-and-date scoping means anything. The companion guides *eDisclosure or Digital Forensics?* and *The EDRM for UK Litigation Teams* cover the escalation and the surrounding lifecycle.

§ 18 · COMMON QUESTIONS

Frequently asked questions

What is a reasonable electronic search?

One whose scope and method are rationally connected to the issues, proportionate under the CPR 31.7 and PD 31B factors, tested against the data (sampled terms, validated stopping points) and documented so its description in the disclosure statement or DRD is accurate and its limits are stated. It is a standard of process, not of perfection: relevant documents can be missed by a reasonable search, and unreasonable searches can fail in either direction, keyhole or landslide.

Do we have to search backups?

You have to preserve them and address them; you rarely have to restore them at the outset. PD 31B makes accessibility and retrieval cost part of the reasonableness balance, so the defensible pattern is: catalogue and preserve the snapshots, search the live sources, and restore selectively where a demonstrated need (typically suspected deletion, or gaps in live data) justifies the cost, with the analysis recorded in the DRD.

Can we rely on keywords alone?

PD 31B warns against exactly that: keyword searches alone may be inadequate, over- and under-inclusive at once. Keywords are one lever among seven, working with custodian and date scoping, analytics, review and sampling. A search description that says "we applied the agreed terms" and nothing else is describing a culling step, not a search.

Are we required to use TAR?

Not required, but in volume cases expect to explain a decision not to: the DRD asks the question, the courts have endorsed the technology since *Pyrrho*, and a party proposing expensive linear review of a large population will face the obvious proportionality point. The safe course is to make the decision on tested numbers and record it either way.

A relevant document has surfaced that our search never caught. Is the certificate wrong?

Not necessarily. If the search was reasonably designed, tested and honestly described, a missed document is the accepted cost of reasonableness: disclose it now, check whether it signals a systematic gap (a channel, a term family, a custodian), fix the gap if so, and record what was done. The certificate is wrong only if it described a search that was not in fact performed, which is precisely what the paper trail in §13 protects against. Remember also that a known adverse document is disclosable regardless of any search.

Who decides if the parties cannot agree terms?

The court, at the CMC or on application, and it decides best when handed short, numbered disputes with hit data and sampling evidence attached. Parties who arrive with tested positions usually settle the list at the door; parties who arrive with adjectives get orders neither side likes.

§ 19 · REFERENCE

Glossary

CAL
Continuous active learning: TAR that retrains on every coding decision, feeding reviewers the likeliest-relevant documents first.

Elusion / null-set sample
A random sample of the unreviewed population measuring what the method left behind; the stopping decision's evidence.

Family-level application
Applying filters so parents and attachments travel together, preventing undated or boundary items orphaning.

Hit report
Per-term counts (documents, families, unique hits) against a processed population; the currency of term negotiation.

Precision / recall
Of what a search returns, how much is relevant (precision); of what is relevant, how much the search returns (recall). They trade off.

Proximity operator
Search syntax binding words within n terms of each other, taming ambiguous pairs.

Richness
The rate at which a review stream yields relevant documents; its decline signals a defensible stopping point in ranked review.

Search protocol
The versioned design document for the search: scope, terms, analytics, sampling plan and exception handling.

Stemming / wildcards
Syntax matching word families (deliver*, judg!) so variants are caught without listing each.

Stopping decision
The recorded, evidence-based decision that review can end, resting on richness and null-set results.

TAR
Technology-assisted review: machine-learning classification of documents by likely relevance.

Unique hits
Documents a term finds that no other term does; the measure of what dropping the term would cost.

Sources and authoritative references

Search-design 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 ("cull hard, before a lawyer reads the first document"; terms iterated against samples with hit reports at each step; suppressions reversible and audited; exception reporting): e-discovery.uk/edrm/processing
- e-discovery.uk, EDRM Phase 05 · Review (hosted, audited review with QC): e-discovery.uk/edrm/review
- e-discovery.uk, Knowledge Centre practice note "Writing search terms that survive negotiation", and 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 access to encrypted, damaged and legacy sources that no search can reach until acquired; examination where deletion is in issue): cflab.uk/digital-forensics-services
- cflab.uk, mobile phone forensics (complete, metadata-intact chat extraction as the precondition of any defensible chat scoping): cflab.uk/mobile-phone-forensics
- CPR 31.7 (reasonable search) and CPR 31.10 (disclosure statement): justice.gov.uk, Part 31
- PD 31B, Disclosure of Electronic Documents (search factors; keyword guidance; pre-CMC discussion; the EDQ): justice.gov.uk, PD 31B
- PD 57AD and the Disclosure Review Document (Model D reasonable and proportionate search; Section 2 methodology): justice.gov.uk, PD 57AD

- *Pyrrho Investments Ltd v MWB Property Ltd* [2016] EWHC 256 (Ch): bailii.org
- Attorney General's Guidelines on Disclosure (rev. 2024: digital material, search strategy and Disclosure Management Documents): gov.uk, AG's Guidelines
- ICO, UK GDPR guidance including data minimisation: ico.org.uk

DISCLAIMER

This publication provides general information about electronic search design in disclosure in the United Kingdom as at August 2026. The worked examples are illustrative composites, not real matters. 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

Search design is where our e-discovery practice (**e-discovery.uk**) spends most of its negotiating life: reading core samples with counsel to capture the house vocabulary, running term-testing cycles with exchange-ready hit reports, configuring analytics and continuous active learning with the validation statistics recorded as they are generated, and producing the protocol, sampling records and exception reports that make the eventual certificate a summary of documents rather than of memory. Where the search meets material no index can reach (encrypted stores, damaged media, deleted data, handsets whose chat must be extracted completely before scoping means anything), the laboratory side of the practice acquires it under chain of custody, and where methodology itself is challenged our examiners give independent evidence under CPR Part 35 / CrimPR Part 19. Conflicts checks, confidential scoping discussions and fixed-fee estimates are routine.

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