COMPUTER FORENSICS LAB

London digital forensics & e-discovery · Established 2007

How to Design Defensible Keyword Searches

Terms, Syntax and the Method That Survives the Other Side's Scrutiny

Keyword searches remain the workhorse of UK disclosure: PD 57AD's Disclosure Review Document asks parties to propose them, courts order them, and most Extended Disclosure exercises stand on them. Yet the average agreed list is drafted in an afternoon from memory, never tested against the data, and defended months later when the opponent asks why "payment" was searched but "remittance" was not, or why a term that hit 190,000 documents was accepted without a syllable of analysis. Designing defensible searches is a method: building terms from the case theory and the custodians' actual vocabulary, using syntax deliberately, testing against the corpus before agreement (guide 85's subject), documenting choices, and knowing what keywords can and cannot do. This guide sets out that method for UK lawyers, from first draft to the DRD table to the challenge that never lands.

⏱ 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**, designs, tests and negotiates search protocols on Extended Disclosure matters: term development with legal teams, syntax engineering per platform, hit analysis and refinement cycles, and the documentation that makes an agreed list defensible at the CMC and after.

ESTABLISHED 2007 · LONDON | ISO 17025-ALIGNED PROCEDURES | SEARCH PROTOCOL DESIGN | DRD & PD 57AD PRACTICE | CPR PART 35 EXPERT REPORTS | FULL CHAIN-OF-CUSTODY DOCUMENTATION

§ CONTENTS

In this guide

§ 0 1 · ORIENTATION

Executive summary

THE HEADLINE POINT: TERMS COME FROM THE DATA'S OWN VOCABULARY, SYNTAX IS ENGINEERED PER PLATFORM, AND EVERY CHOICE IS TESTED AND DOCUMENTED BEFORE IT IS AGREED

A defensible search protocol is built, not brainstormed. **Term development:** candidates come from the case theory (issues, transactions, dates), but the decisive layer is the custodians' actual language: harvested from the pleadings' key documents, early sampling of the corpus, client interviews and the informal register (nicknames, codenames, abbreviations, typos: guide 63's messaging vocabulary especially): because people do not write "the alleged misrepresentation"; they write "the Jenkins thing". **Syntax:** operators, wildcards, proximity connectors and phrase handling are engineered against the specific platform's rules (they differ materially), with stemming, noise-word and diacritic behaviour confirmed rather than assumed. **Testing:** every list runs against the data before agreement: hit counts per term, sampled precision, null-set checks: guide 85 treats this in depth: so terms are widened, narrowed or replaced on evidence. **Documentation:** the origin, purpose and testing of each term is recorded in the protocol file, which is what converts "we searched reasonably" from assertion into exhibit. Keywords then do what they are good at: and their known blind spots (§5) are covered by other tools, stated in the DRD honestly.

- **Harvest the real vocabulary:** key documents, samples and interviews before drafting: the corpus names its own terms.
- **Engineer, don't guess, syntax:** platform rules for wildcards, proximity and phrases confirmed in writing per list.
- **Test before agreeing:** hit counts and sampling per term: acceptance of an untested list is the original sin (guide 85).
- **Cover the blind spots:** images, numbers, code-speak and multilingual content handled by OCR, analytics and TAR: declared, not hidden.
- **Write it down:** per-term provenance and test results: the protocol file is the defence.

§ 0 2 · FIRST PRINCIPLES

The problem in plain English: the list drafted from memory

Most keyword lists are written the way shopping lists are: from memory, by people who were not in the kitchen. The lawyers know the issues; they do not yet know the words: that the client's staff called the disputed product line "Project Kestrel", the counterparty "PBL" and the problem "the wobble"; that the finance team wrote "remit" where the pleadings say "payment"; that the crucial exchanges are riddled with a typo ("agreemnet") that a clean search will never meet. The resulting list is a translation of the statements of case into search boxes: fluent in law, illiterate in the data.

The consequences arrive in both directions. Terms too narrow silently miss the responsive core, and nobody knows, because absence of hits looks like absence of documents. Terms too broad hit six figures, get "agreed" under time pressure, and detonate in the review budget. And when the opponent later asks: as PD 57AD's cooperation duty entitles them to: why this term and not that one, "it seemed sensible at the time" is the answer of a party that did not run a method. The method exists, it is not expensive, and it is this guide.

§ 0 3 · THE RAW MATERIAL

Where terms come from: case theory and custodian vocabulary

- **The issues layer:** each pleaded issue decomposed into its entities, events and artefacts: parties, products, contract clauses, payment references, dates: the skeleton list, mapped issue-by-issue so coverage is demonstrable in the DRD.
- **The key-document harvest:** the twenty documents everyone already has, mined for the words actually used: names, abbreviations, project codes, phrasings: the cheapest and highest-yield source, routinely skipped.
- **Custodian interviews:** the guide 63 question set: what did you call it, who did you discuss it with, on which channels, with what nicknames: ten minutes per custodian buying terms no lawyer would invent.
- **Early corpus sampling:** a first exploratory pass on the loaded data (indexed per guide 81): frequency lists, names cropping up beside known terms, candidate codewords: the data proposing its own vocabulary.
- **The informal register:** typo variants of load-bearing words, phonetic spellings in chat, multilingual fragments where the business ran in two languages: listed deliberately, because the candid documents live in the careless register.

§ 0 4 · THE ENGINEERING

Syntax that works: operators, wildcards and proximity

- **Boolean structure:** AND to intersect concepts, OR to gather variants, NOT used sparingly and consciously (every NOT excludes documents nobody will ever see: it is a scope decision wearing syntax).
- **Wildcards and stemming:** truncation (pay* for pay, payment, payable) widens honestly but explosively (const* meets construction, constable and constant): each wildcard's expansion checked against the index, and the platform's automatic stemming behaviour confirmed so terms are not double-widened.
- **Proximity connectors:** the precision workhorse: "price w/5 fix*" finds pricing discussions that "price AND fix*" drowns in noise: connector distances chosen per concept and platform dialect (w/n, near/n) confirmed.
- **Phrases, numbers and identifiers:** exact phrases quoted; account numbers, invoice references and amounts handled with the platform's tokenisation rules in mind (hyphens, slashes and currency symbols split terms unpredictably: tested, per §6).
- **The platform dialect check:** one written confirmation from the provider per matter: operator set, wildcard rules, noise words, diacritic and case handling, tokenisation of punctuation: because identical strings return different sets on different platforms, and the agreed list must mean one thing.

§ 0 5 · THE BLIND SPOTS

What keywords miss: the known failure modes

- **Text that is not text:** scans, photographed documents and image-only PDFs are invisible until OCR runs (guide 86's subject): the search protocol states the OCR posture or it is searching a partial corpus silently.
- **The unnamed wrongdoing:** people arranging something improper rarely name it: "as discussed", "the usual", "call me": met by participant-and-date searches, communication-pattern analytics and the guide 70 chronology, not by ever-longer term lists.
- **Numbers and artefacts:** spreadsheet values, amounts and metadata fields often sit outside text indexes: structured-data questions route to guide 71's methods, and the DRD says so.
- **Language and encoding:** multilingual estates need terms per language and diacritic policy confirmed; transliterated names (three spellings of one surname) enumerated.
- **The honesty requirement:** these limits are declared in the protocol: keywords for the nameable, analytics and TAR (guide 89) for the rest: a search strategy that claims completeness for keywords alone is the challengeable one.

§ 0 6 · THE TABLE STAKES

The negotiation: DRD tables, hit counts and refinement

- **The DRD's ask:** Section 2's fields invite proposed terms, custodians, date ranges and repositories: the search protocol is a court document from birth: drafted to be read by a judge, not just a provider.
- **Exchange with evidence:** proposals accompanied by hit counts and, where agreed, sampled precision per term: the guide 85 discipline as negotiation currency: because "this term hits 190,000 documents, of which a 200-document sample shows 3% relevance" ends arguments that adjectives cannot.
- **Refinement cycles:** broad terms narrowed by proximity or intersection rather than deleted; missing-variant additions accepted gracefully; each cycle's deltas recorded: cooperation performed in the open, per PD 57AD's duty.
- **Receiving their list:** the same scrutiny in reverse: dialect confirmed, expansions checked, blind spots probed (what covers the scans? the chat vocabulary?): acceptance is a decision, tested like your own drafting.
- **Disproportionate demands:** answered with numbers: the cost model per term (hits, review rate, price) that converts "unreasonable" from complaint into arithmetic the CMC can act on.

§ 0 7 · THE PAPER TRAIL

Documentation: making the list defensible

- **The protocol file:** per term: origin (issue, document, interview, sample), syntax rationale, test results, refinement history, final status: one living table, maintained as the list evolves.
- **The dialect record:** the §4 platform confirmation filed with the protocol: the meaning of the agreed strings fixed in writing.
- **The blind-spot declaration:** what keywords were not asked to do, and what covers it: OCR posture, analytics, TAR, structured-data routes: the completeness story told across tools.
- **Version discipline:** lists exchanged, amended and re-run as numbered versions with deltas: guide 81's reprocessing discipline applied to queries: no silent edits to an agreed list, ever.
- **The disclosure statement's foundation:** the certificate's "reasonable search" assertion resting on a file that shows the reasoning: the difference, under later challenge, between producing a document and producing a shrug.

§ 0 8 · THE WIDER MAP

Source architecture: where else the evidence lives

Per this series' source-architecture discipline, adapted to method: when a keyword search returns nothing, or too little, the target evidence has not been proven absent: it may live beyond the terms, or beyond text search entirely. The responsive material may also live in: the unsearchable layer (unOCRed scans and images: rendered searchable per guide 86 before absence is pleaded); the informal register (the typo, the nickname, the code word: harvested per §3); the non-text estate (numbers in spreadsheets, metadata fields, structured systems: guides 71-74's query methods); the participant dimension (everything between the key people in the key window, retrieved by custodian-and-date rather than by term); the analytic layer (communication patterns, near-duplicate clusters, TAR-ranked documents per guide 89); and the adjacent channels where the vocabulary changes entirely (chat's dialect per guide 63; the guide 70 channel switches). A null result triggers this map, not a conclusion: the search protocol's honesty is that it says which layer answered each question.

WHEN KEYWORDS RETURN...	WIDER TERMS / VARIANTS	OCR & TEXT REPAIR	PARTICIPANT & DATE PULLS	STRUCTURED-DATA QUERIES	ANALYTICS & TAR	ADJACENT CHANNELS
Nothing on a live issue	Y: harvest again per §3	Y: check the unindexed	Y: the safety net	Numbers, not words	Pattern-led retrieval	Chat dialect differs
Too much noise	Proximity & intersection	n/a	Narrow by window	n/a	Ranking triages	n/a
Hits missing known documents	Diagnose the term	Y: often the cause	Y: retrieve directly	Y: field-level fetch	Similar-document nets	Y: wrong channel searched
Coverage unprovable	Issue-mapped lists	OCR posture declared	Custodian completeness	Source inventory	Elusion testing (guide 85)	Channel inventory
Opponent's null claims	Demand their variants	Demand their OCR posture	Demand the window pull	Demand query bundles	Demand method disclosure	Demand channel census

Fallback strategy in one line: a keyword null is a routing instruction: to better terms, repaired text, participant pulls, structured queries or analytics: never, by itself, a finding that the documents do not exist.

§ 0 9 · IN THE WILD

Worked examples

EXAMPLE 1 · THE CODENAME THE PLEADINGS NEVER KNEW

A joint-venture dispute ran its first agreed list: fifty-four terms translated faithfully from the statements of case: and returned a corpus strangely thin on the venture's collapse. The §3 harvest fixed it in a week: the key-document mine showed the client's side calling the counterparty "PBL"; a custodian interview surfaced the venture's internal name, "Project Kestrel", used precisely because the counterparty's name was radioactive on email; and frequency sampling flagged "the wobble" clustering beside known dates. Three additions multiplied the responsive set: and the candid documents were, as predicted, almost exclusively in the vocabulary no pleading had used. The revised protocol file recorded each term's origin, which mattered at the CMC when the opponent asked why the list had changed: the answer was a method, exhibited.

EXAMPLE 2 · THE WILDCARD THAT ATE THE BUDGET

An agreed list included "const*" (for "construction"): accepted by both sides in a Friday exchange, untested. It also matched "constant", "constable", "constitute", "constraint" and, across a 1.4m-document corpus, 212,000 hits at a sampled relevance under 2%. The producing party, having agreed the term, faced a review estimate north of £300,000 for it alone. The rescue was §6's refinement cycle: the term renegotiated to "construct* w/10 (defect* OR delay* OR variation*)", hits falling to 9,400 at 61% sampled relevance, the deltas documented, and the revised term recited in an amended order. The costs of the detour were shared: with the judge's observation that a two-hour test before agreement would have cost neither party anything.

EXAMPLE 3 · THE NULL RETURN THAT WAS AN OCR POSTURE

A defendant certified that searches for the disputed guarantee terminology returned nothing before 2021: pleaded as proof the guarantee was a late fabrication. The §8 routing exposed the real cause: the pre-2021 estate was a scanning archive, and the defendant's processing had never OCRed it: the terms had been searched against documents with no text to search. The claimant's application secured OCR of the archive (guide 86's disciplines), the re-run terms hit 340 documents including three drafts of the guarantee dated 2019, and the "fabrication" defence inverted into a disclosure-conduct problem. The certificate's author had honestly reported a null from a search that could never have succeeded: the protocol file's missing line: OCR posture: was the whole case.

§ 10 · WHERE IT GOES WRONG

Common mistakes and technical limitations

Common mistakes

- **Lists from memory:** the pleadings translated, the corpus never consulted: Example 1's thin first pass.
- **Agreement before testing:** Example 2's Friday exchange: the untested wildcard as a six-figure decision.
- **OCR posture unstated:** Example 3's null: searches certified over text that was never there.
- **NOT operators as quiet scope cuts:** exclusions agreed as syntax that no one priced as decisions.
- **Dialect assumed:** the same string meaning different things on the two parties' platforms: the confirmation letter never sent.
- **Chat searched in email English:** the guide 63 register ignored: formal terms run against informal writing.
- **Silent list edits:** terms tweaked after agreement without versioned deltas: cooperation breached by keyboard.
- **Null results pleaded as absence:** §8's routing skipped: the map's first row, unread.

Technical limitations

- **Keywords find words, not meaning:** synonymy and euphemism defeat them by design: the analytics and TAR layers exist for this, and the protocol says so.
- **Indexes have edges:** exception items, container quirks and field coverage bound what any term can reach: guide 81's reconciliation defines the searched universe.
- **Tokenisation surprises:** hyphens, slashes and symbols split or join terms per platform: identifier searches are tested, not trusted.
- **Precision and recall trade off:** every narrowing risks loss, every widening buys noise: the balance is evidenced by sampling, never asserted.
- **Language coverage is enumerable:** terms exist per language actually used: the estate's language census precedes the list.

§ 11 · INTERROGATORIES & DRAFTING AIDS

Questions to ask · Suggested wording

Ask your client

- What did your people actually call the key parties, projects and problems: nicknames, codenames, abbreviations, typos included?
- Which languages and channels carried the relevant discussions: and does the list speak each one?
- Which twenty documents already tell the story: for the §3 harvest before any drafting?

Ask your opponent

- Please provide, for each proposed term, its hit count against your data and any sampling results, and state the platform, operator dialect and OCR posture under which the searches will run.
- Please confirm what your search methodology provides for material keywords cannot reach: image-only documents, structured data, and messaging platforms: and identify the tools addressing each.
- Please provide versioned deltas for any change to the agreed term list, with re-run results.

Ask your eDiscovery / forensic provider

- What is the platform's exact dialect: operators, wildcards, noise words, tokenisation: in writing, for the protocol file?
- What do the per-term hit counts and samples show: and which terms need the Example 2 treatment before exchange?
- What fraction of the corpus is currently unsearchable text: and what is the OCR plan?

SUGGESTED WORDING · SEARCH METHODOLOGY PARAGRAPH FOR THE DRD / PROTOCOL

> "Keyword searches shall be run on [platform], whose operator syntax, wildcard behaviour, noise-word list and tokenisation rules are recorded at Appendix [A]. The agreed terms, their per-term hit counts and sampling results are set out at Appendix [B], together with each term's origin and refinement history. Image-only and non-text-searchable documents shall be rendered searchable by OCR prior to the application of search terms, and the proportion of the corpus so treated shall be stated. Search terms shall not be applied to structured data sources or messaging platforms except as adapted per Appendix [C], which records the methodology for those sources. Amendments to the term list shall be made only by exchanged, versioned revisions with re-run results. The parties acknowledge that keyword searching is one component of a reasonable search and record at Appendix [D] the analytical and technology-assisted methods addressing material not amenable to keyword retrieval."

§ 12 · QUICK CONTROL

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

The search-design checklist

- ☐ Issues decomposed; skeleton terms mapped issue-by-issue
- ☐ Key-document harvest and custodian interviews completed before drafting
- ☐ Informal register enumerated: nicknames, codenames, typos, languages
- ☐ Platform dialect confirmed in writing and filed
- ☐ Every term hit-counted and sampled before exchange (guide 85)
- ☐ Wildcards expansion-checked; NOTs treated as scope decisions
- ☐ OCR posture stated; unsearchable fraction known
- ☐ Blind spots declared with covering tools named
- ☐ Protocol file maintained: origin, tests, versions, deltas
- ☐ Null results routed per §8 before any absence is pleaded

Red flags

- Term lists exchanged without hit counts: Example 2's precondition.
- Certificates of null results over unstated OCR postures: Example 3's trap.
- Agreed lists that never met the corpus vocabulary: Example 1's thinness.
- Opposing "methodologies" that are keywords alone, claiming completeness.
- Silent term edits between productions.
- Six-figure hit terms accepted under exchange deadlines.
- Identifier searches never tested against tokenisation.

When to involve a digital forensic expert

At design, where the harvest, the dialect check and the first hit analysis happen before anything is exchanged; at negotiation, where refinement cycles run on evidence and disproportion is answered in numbers (Example 2's rescue); at challenge, in either direction, where protocols are audited: coverage, posture, versions: and Example 3's kind of null is diagnosed; and at reporting, where search methodology is evidenced under CPR Part 35 when it becomes an issue. Through **e-discovery.uk**, search design runs as a documented EDRM discipline with the protocol file as standard deliverable: the list that survives scrutiny because it was built to be scrutinised.

§ 13 · COMMON QUESTIONS

Frequently asked questions

How many search terms should a list have?

As many as the issues need and no more: quality of construction beats length: a tested thirty-term list with proximity engineering outperforms an untested hundred every time, and costs less to run, review and defend. The DRD conversation goes better with a short list you can justify term-by-term than a long one you cannot.

The other side proposes terms hitting hundreds of thousands of documents. Must we accept?

No: you counter with arithmetic: hit counts, sampled precision, review cost per term: and propose refinements (proximity, intersections, date limits) that keep the concept while shedding the noise: Example 2's renegotiation is the template, and courts consistently reward the party holding numbers rather than adjectives. Acceptance under deadline pressure of an untested term is the one concession that cannot be walked back cheaply.

Our searches returned nothing on a central issue. Can we certify that?

Only after the §8 routing: is the relevant estate actually text-searchable (Example 3's question first), did the terms speak the custodians' register, has a participant-and-date pull been run as the safety net? A null certified over those checks is solid; a null certified instead of them is a future application with your name on it.

Do keyword searches work on WhatsApp and Teams data?

They run, but the vocabulary and the unit change: informal spelling, fragments-across-messages, emoji and voice notes defeat email-register terms: so messaging sources get adapted terms from the guide 63 harvest, conversation-level retrieval, and honest declaration in the protocol that the methodology differs per source. Searching chat with the email list and reporting the nulls is a recognised way to miss the case.

Are keyword searches still acceptable now TAR exists?

Fully: PD 57AD contemplates both, and mature practice combines them: keywords for the nameable (parties, projects, references), TAR and analytics for the conceptual and the unnamed, with the DRD recording which tool covers what: guide 89 treats the TAR half. What is fading is not keywords but untested keywords presented as a complete methodology.

Who should draft the list: lawyers or the provider?

Both, in sequence: lawyers own the issues layer and the legal judgements (scope, NOTs, proportionality); the provider owns the dialect, the testing and the engineering; the custodians own the vocabulary nobody else has. The protocol file is where the three meet: and a list drafted by any one of them alone carries that author's characteristic blind spot.

§ 14 · REFERENCE

Glossary

Boolean operators
AND, OR, NOT: the logic connecting terms.

Dialect
A platform's specific syntax rules; confirmed per matter.

Elusion
Responsive material in the un-hit set; tested per guide 85.

Hit count
Documents a term returns; the negotiation's currency.

Noise words
Common words a platform refuses to index; checked, not assumed.

Precision
The fraction of hits actually relevant.

Proximity connector
Terms within n words (w/n); the precision workhorse.

Recall
The fraction of relevant material actually retrieved.

Tokenisation
How text splits into searchable units; the identifier trap.

Wildcard
Truncation matching variants (pay*); expansion-checked.

Sources and authoritative references

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

- e-discovery.uk, EDRM Phase 04 · Processing and Phase 05 · Review (search design, testing and refinement as practised): e-discovery.uk/edrm/processing · e-discovery.uk/edrm/review
- cflab.uk, digital forensics services (search protocol auditing, null-result diagnosis, CPR Part 35 methodology evidence): 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 and the Disclosure Review Document (Section 2's search fields): justice.gov.uk, PD 57AD
- CPR Part 31 and PD 31B (where applicable outside the Business and Property Courts): justice.gov.uk, Part 31 · CPR Part 35: justice.gov.uk, Part 35
- ICO, UK GDPR guidance: ico.org.uk · ISO/IEC 27037: iso.org, 27037

§ HOW A SPECIALIST LABORATORY CAN ASSIST

Working with Computer Forensics Lab

Search design at **e-discovery.uk** runs the whole method of this guide: the §3 harvest with the legal team and custodians, syntax engineered against the confirmed platform dialect, every term hit-counted and sampled before exchange, refinement cycles run on evidence, and the protocol file maintained from first draft to final certificate: Examples 1 and 2's corrections built in as standard practice, Example 3's posture line never missing. The laboratory side audits opposing methodologies, diagnoses suspicious nulls, and evidences search reasoning under CPR Part 35 where the protocol becomes the issue. Conflicts checks, confidential scoping discussions and fixed-fee estimates are routine; and if your term list is due for exchange and has never met the data, guide 85's testing pass is this week's work: before the agreement, not after it.

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