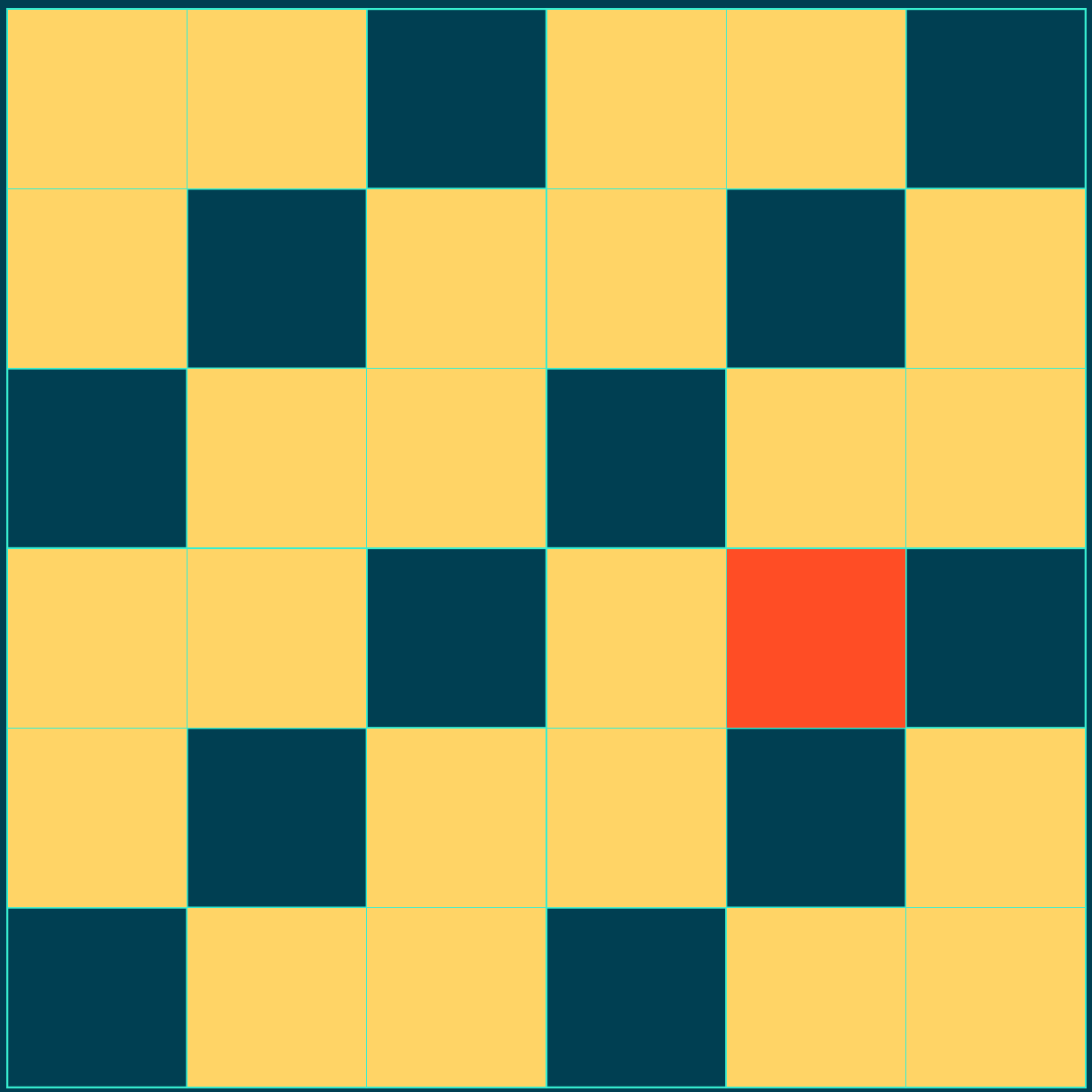
Finding needles in the data haystack



In this guide we explore insights for finding the needles in the data haystack.

Reading time	7 minutes	Page count	12 pages	Word count	1,499 words

Despite the increasingly burdensome process of data and document review, most legal matters will inevitably turn on a handful of critical documents.

Many lawyers will agree with the sentiment that despite the increasingly burdensome process of data and document review, most legal matters will inevitably turn on a handful of critical documents. The challenge in our complex and varied, technologically driven age is effectively finding those needles within an increasingly larger data haystack.

But just how quickly is that haystack growing? Statistics published in 2023 indicate that by the year 2025 there will be over 180 zettabytes of dataˆ in the global datasphere. For context, one zettabyte is equal to one trillion gigabytes. Fortunately, it is currently unlikely that any lawyers will have to deal with a zettabyte. We more commonly see cases where the starting data size is 10GB and increasingly, at the high end, the data is north of a terabyte (1,000GB). Data within this range can amount to anywhere between 40,000 to more than 4 million documents.

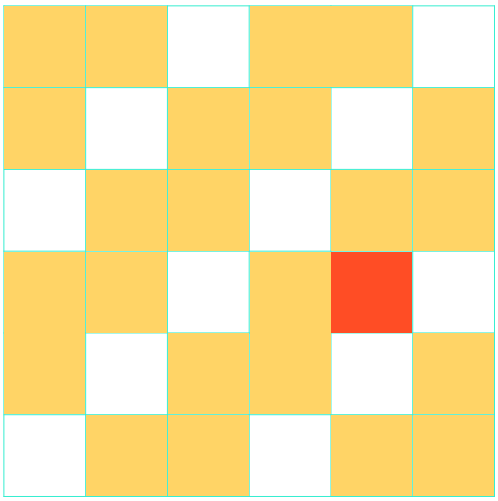
ˆSource: statista.com

Not only can the volume of data present an unfathomable challenge for lawyers but also the variety.

Data is no longer only made up of emails and Microsoft Office documents but more frequently includes custom databases, multi-media files, and new communication mediums across many devices. This variety presents new and unique challenges for collection and review.

While data collection is key, the main challenge for a lawyer's review is related to the time and budget constraints placed upon the review of documents. This applies to situations involving as few as one hundred documents into the tens of millions.

Authorities like Courts, Tribunals and Regulators also wish to avoid unnecessary delays or fees to the parties that are disproportionate to the scope and quantum of the matter. Thankfully, there are workflows available to lawyers that can reduce the volume of documents for review and increase the speed and accuracy with which those needles are found.



There are several workflows that can reduce the volume of documents for review and increase the speed and accuracy.

Target high value scores

With the dramatic increase in volume and sources of data it is important to be find a balance in the selection process. Often, the net is cast either too wide, or too narrow. Both outcomes can present different challenges for legal teams later in the matter.

You should take the time to understand the broad scope of your client's data environment. This includes an understanding of the key stakeholders involved and will help to inform calculated decisions on what information sources you should focus on. This will ensure that you are not swamped with unnecessary or irrelevant information but equally, you aren't unduly exposing yourself to the risk of missing a critical document.

Remove junk and system files

When data is extracted from its source, in particular larger data sources, there will be system files contained in the set. These computer-generated files are generally irrelevant for most reviews and can be removed. However, it is not always common practice amongst eDiscovery providers to do this, whether due to lack of capability or their standard practices. Where possible, you should look to exclude these file types from the data you intend to review. It will help to reduce the potential volume for review and your fees associated with same. This removal process should always be done using the right defensible technology and process.

Remove irrelevant information

Once you have a clean data set you can begin to do some preliminary culling throughout the sorting and processing phases. This will generally involve the removal of technical duplicates, however where appropriate data that is potentially irrelevant can also be removed from the review set. Some of the more common and effective ways to do this is to set a critical date range and exclude documents outside of that range from your review. You can also choose to focus on a particular persons or stakeholders.

Another way to identify potentially irrelevant documents is through keyword searches whereby you can set up searches that will help to identify documents that are potentially irrelevant, allowing you to focus on what is more likely to be important.

There are limitations to each of these methods, but a good eDiscovery provider will help you to manage and overcome them. You should also be mindful not to completely disregard documents that are at one point in time deemed irrelevant because the scope of matters can change over time and documents initially excluded may become relevant.

Prioritise key information

The same methods you use to remove irrelevant information can also be used to sort information into priority or critical document review sets. Dates, stakeholders, and key word searches can be used to group documents together to be batched and distributed to reviewers to be reviewed as a priority. This approach sets the initial review up for quick success and may help inform preliminary advice or ongoing collection or review strategy.

Understanding your client data

To comply with many of the common requirements throughout this guide it is important that lawyers have a good understanding of their client's data and documents. The Court's regularly call upon the parties to justify why electronic discovery and disclosure is appropriate or why the scope of the disclosure and discovery should be extended or limited. Much of the justification is based on the nature of the data the parties have within their control. With the Court now focusing on resolving some of these matters early in the proceedings it is becoming a requirement for the parties to understand their potential data and document makeup sooner. Failure to do so may involve making a decision that is not ideal or being unable to properly address the Court's requirements at all.

Court discretion

There are broad discretionary powers within many of the jurisdictional legislative provisions that will allow the Court to make Orders that place varying discovery requirements upon parties. Even in jurisdictions that do not expressly reference electronic discovery and disclosure, the parties, by operation of these powers, may be able to seek an order that the disclosure and discovery process be varied to include the use of electronic discovery. The strength of such applications will certainly be bolstered by the links you can draw between the use of technology and the advancement of the respective efficiency and proportionality provisions within the jurisdiction.

Email threading

Using this workflow you can group email conversations together in one place, no matter where they appear in the dataset. A visualisation of the email thread or chain is also available that helps you to identify missing content as well as email end points that may contain all the information contained in the preceding components of the chain. This technology can help reduce the number of documents you need to review as well as improve efficiency as all threads can be reviewed in one place.

Near duplicate identification

Using this workflow you can group email conversations together in one place, no matter where they appear in the dataset. A visualisation of the email thread or chain is also available that helps you to identify missing content as well as email end points that may contain all the information contained in the preceding components of the chain. This technology can help reduce the number of documents you need to review as well as improve efficiency as all threads can be reviewed in one place.

Helpful resources

As specialists we continually invest in R&D and best practice so we can advise our partners with confidence. These insights culminate in helpful resources and references for lawyers and decision-makers.

Data Identification Questionnaire

Our questionnaire reference aims to help you quickly and accurately identify data potentially relevant to your matter. The information captured from key stakeholders will facilitate the development of a collection plan and enable its swift and defensible execution.

Learn what to consider

Key timings for disputes

Collaborative partnership with your eDiscovery specialist can extend beyond a specific project and, if progressed towards a more integrated solution, can create significant benefits and a competitive advantage for legal teams.

Learn what to consider

Learn more

Managing review volume through reduction	↗
Developing an appropriate review strategy	↗
Leveraging technology to increase efficiency	↗

Across hundreds of matter types in all sectors and jurisdictions, we focus on solutions and impact. Here is a selection of matters that may be relevant to you.

Relevant matters

Prioritising the review order of received discovery

Dispute, Regulatory, Investigation, Collect, Process, Analyse

Read

Reducing review volumes with data processing tools

Dispute, Collect, Process, Analyse

Read

Leveraging continuous active learning in large scale document review

Dispute, Analyse, Review

Read

View all solutions ↗

You need a team with a balance of legal, eDiscovery and technology expertise, this is who we are.

Our expert team of lawyers and technologists are available to assist you with navigating all stages of your matter, from the first meeting, through scoping, to completion. We focus on technical solutions so you can focus on the law. Find out how we help.