Legility

White Paper

Create More Efficient Document Review: Five Ways to Prep Your Data

Create More Efficient Document Review: Five Ways to Prep Your Data

Contents

Five Ways to Prep Your Data

Create More Efficient Document Review

Document review software is integral to the success and effectiveness of a review. In this white paper, we'll discuss the five ways you should be preparing your data for more efficient document review:

Global Duplication

First things first - to be able to work with the smallest defensible data post-processing, you should use global deduplication. Global deduplication works by comparing all of your data against itself. Thus, if two custodians both have the same document in their possession, the post-processing data set will only contain one iteration of that document and a corresponding field indicating which other custodian(s) it also belonged to prior to deduplication. If you were to instead utilize custodian level deduplication, the data will only be compared against each custodian's own documents and the post-processing data set would contain (at least) two iterations of the same document. Depending on the size of the collected data and the number of custodians, global deduplication can reduce the number of documents in your review by anywhere from 10 percent to upwards of forty percent.

Domain Parsing

Before you blindly start reviewing documents, get to know your data set. Reviewing the domains present in your data set can prevent the waste of lawyer time on documents that are not responsive to the matter at hand. Many people use their corporate email accounts for personal business.

> “By focusing on **refining the terms to more accurately reflect the documents** that are potentially responsive to the matter, you can **save a lot of time and money** in the long run.

It is often easy to eliminate up to 20 percent of your data set by excluding junk domains from your review population. These junk domains are usually from financial newsletters, travel sites, and online stores. Always make sure that you review a statistically sound sample of the excluded documents to guarantee your domain exclusions are defensible.

Search Term Testing

In eDiscovery, deadlines usually hover around the “yesterday” timeframe. Due to this rush, legal teams generally want to jump right into a review once the search terms narrow the population to a reasonable set of documents. Reducing the processed set of data from 4,000,000 to 900,000 documents with search terms might seem “reasonable”, but without anything else to go on, it does not make sense to start the review just yet.

When individual search terms hit to a sizeable number of documents, you should review a statistical sample of documents to determine whether the terms are pulling documents responsive to your matter or whether they lead to a large population of false positives. If you identify numerous false positives, you should revise the term to eliminate the false positive pool of documents. This process can be time consuming, especially considering the rush to get started right away.

By focusing on refining the terms to more accurately reflect the documents that are potentially responsive to the matter, you can save a lot of time and money in the long run. This approach may not be logical for smaller reviews, but when dealing with 100,000 or more documents the dividends of this pre-review activity should pay off. Also, keep in mind that it is difficult to revise search terms once they are agreed upon by both parties. Therefore, if you do not spend time testing and limiting your search terms on the front-end, it will be difficult and sometimes impossible to do at later stages.

Analytics

Email Threading

When documents are processed, even with global deduplication, there are often numerous versions of the same email thread in the data set.

For example:

I. John Smith emails Jane Doe

- II. John Smith emails Jane Doe; Jane Doe replies to John Smith
- II. John Smith emails Jane Doe; Jane Doe replies to John Smith; John Smith replies to Jane Doe

With this example, your data set would include at least three versions of this same conversation. By using email threading in Relativity, you can limit your document review to only inclusive threads. Rather than your lawyers reviewing and coding three separate documents, they would only need to review one document – example III above. Email threading also considers offshoots of a conversation, so if Jane Doe forwarded example II above to someone else, that forward would be its own inclusive thread indicating that you should review it as well. Email threading saves time, money, and focuses your lawyers' attention on the conversations that matter.

Near Duplicate Identification

Near duplicate analytics in Relativity can be very helpful to your document review team. Although you should not generally use this to eliminate documents from the review itself, it should be used to quality check the review team's coding. When running near duplicate identification, you can choose the similarity percentage that the analytics tool will use. We generally recommend a percentage of 90 percent - however, you can modify this upwards or downwards, depending on your needs. For example, if you are looking to compare documents that were produced by opposing counsel to documents that you have produced in the same matter, you might want to drop the similarity percentage down to 85 percent to account for the differences in the text of the two sets of documents caused by Bates and Confidentiality stamping. One thing to keep in mind with near duplicate analytics is that the extracted text of the documents are used for the comparison, so if a document does not have extracted text then this type of analytics will not work on that document. Additionally, always be aware that this technology does not work as precisely for longer documents. For example, with 90 percent similarity the system could group a 1,000-page document as a near duplicate of another document and 100 pages of text could be completely different.

Despite these drawbacks, near duplicate analytics has many applications in the review process. It can be used to confirm that all privileged or redacted documents have been tagged and redacted appropriately. By creating a search of all documents tagged for privilege or redactions, and adding in their near duplicate groups, the review team can easily check all documents textually similar to the privileged document set and spot potential errors in the privilege coding or redactions.

Using the same logic, the review team can also check all documents tagged responsive or non-responsive against their near duplicates. These checks will ensure a consistent document review in the most efficient and comprehensive way possible, as opposed to more manual quality checks traditionally utilized in reviews (e.g., subject line comparisons). As mentioned above, using near duplicate analytics is also a very easy and comprehensive way to compare opposing party productions to your own.

Segregate Dcoument Population for Batching

We've already globally deduped the processed data, we've excluded junk domains from our review set, we've tested and refined our search terms, and we've run email threading and near duplicate analytics. Now it's time to segregate the documents that are left into separate review pools for batching. By segregating the documents, you are enabling groups of reviewers to become more familiar with certain document types, thereby increasing their efficiency and knowledge base. You should separate documents based on whether they contain privilege hits, GIF/JPEG/Video file types, and once the review is further along, based on tagged responsive and non-responsive search strings. Keep these documents segregated and never batch more than a couple of days' worth of documents for your team, so it is easy to revise the batching parameters.

You should batch documents with privilege hits based on their inclusive email thread groups, so that a sub-set of more experienced reviewers can focus on potentially privileged documents and familiarize themselves with the type of privilege found in the matter. Once these reviewers are familiar with the privileged documents, they will also easily identify additional privilege terms present in your documents that someone with less awareness would have missed.

You should also segregate and batch out loose GIF/JPEG/Video file types on their own. The reviewers focusing on these document types should be able to move very quickly through these batches and knock out large chunks of data at a much quicker speed than the rest of the team.

A week or so into the review, you should analyze the groups of documents tagged responsive and non-responsive. Discuss these tagged documents with the review team in order to determine what within the four corners of the documents made them responsive or non-responsive. If working with a more experienced review team, have them note this information in a comments box on the coding layout so you can easily consolidate the information.

Using this information, create a tagged responsive string and a tagged non-responsive string, which may or may not have anything in common with the original search terms used to cull the entire document review population. Once the responsive and tagged non-responsive strings are created, batch out the next sets of documents in three groups – by hits with families to the responsive string, then by hits with families to the non-responsive string, and a set of documents that do not hit to either string.

As the review progresses, continue to refine your responsive and non-responsive search strings based on the additional documents tagged responsive and non-responsive that did not hit to your tagged responsive and non-responsive strings. This process does require more time to set up than a traditional, linear document review, but it will improve reviewers' efficiency, force reviewers and project managers to stay engaged and analyze the patterns in your data set, and, in turn, will ensure a quicker review and a more consistent and defensible final work product.

We're here to do the best legal work of our lives alongside our innovator clients. Ready to kickoff a project?

Email us at newclients@legility.com
or visit legility.com/insights

Legility

We transform legal teams by designing best-in-class, seamless legal experiences, driven by strong values and rooted in the world-class technology, strategy, and talent required for operational excellence. Everyone and everything at our company is driven by a core mission of building strong and proactive relationships between lawyers and their clients – whether those clients are internal or external. Our lean, interdisciplinary teams work in close collaboration with our clients, helping them to in turn solve their clients' biggest problems. Our lawyer-centric, data-driven approach focuses on creating a seamless experience across all operational areas – from litigation to risk and compliance, from contract review to case strategy. This unified legal experience results in a much more consistent, impactful, and valuable relationship between legal teams and their clients and business stakeholders.