opentext™

White paper

Optimizing modern ESI investigations to find the facts swiftly

Uncovering the critical documents and key evidence for all types of matters—potential litigation, compliance, regulatory or internal investigations

Modern ESI investigations—that is, interrogating a large collection of electronic documents to quickly answer the key questions and locate the critical evidence—are intensely demanding in every sense of the word. Further complicating modern ESI investigations is the trend toward an increasingly remote workforce.

opentext™

Contents

Executive summary	3
Control data proliferation for security and access	3
Establish appropriate policies for remote operations	4
Consider seamless end-to-end cloud-based capabilities	5
Recognize the differences between an investigation and a production review	7
Preserve and collect immediately, expansively—and discreetly	7
Use communication analytics to locate additional witnesses	8
Use efficient machine learning techniques	8
Structure the investigation team with collaborative independence	9
Effectively explore the unknown	10
"Prove a negative" using TAR based on continuous active learning	10
Use statistics to scope the review	11
Initiate a review for documents that are close to responsive using analytics	12
Surface any truly responsive documents using TAR based on continuous active learning	13
Use the review and statistics to “prove a negative”	13
QC every investigation with TAR based on continuous active learning	14
Conclusion	14

opentext™

Executive summary

The simple fact that an investigation is warranted underscores the significance and importance of the exercise. Timely, cooperative self-reporting plays a central role in today's heightened global regulatory environment, and can obviate a sweeping agency investigation that would otherwise divert or exhaust valuable company time and resources. Accurate early case assessment might well mean the difference between a fully informed, favorable settlement and an overly expensive and protracted litigation. And a quiet, behind-the-scenes review of an internal candidate's emails might help ensure that the new CEO does not engender an unanticipated public relations nightmare.

In every instance, time is certainly of the essence. For example, Sarbanes-Oxley requires a company to investigate whistleblower complaints quickly or risk exposing a lack of compliance controls; the Dodd-Frank Act only increased this pressure. The failure to promptly initiate an investigation of sexual harassment allegations not only degrades the entire process, but could also give the impression that the organization is essentially discouraging legitimate complaints. And virtually every merger or acquisition carries with it a post-deal due diligence period that is measured in days, barring subsequent recovery regardless of how consequential.

Yet there is no room for error, and no excuse for being less than demandingly thorough. Even after relaxing the all-or-nothing approach to cooperation in the Yates Memo, the Department of Justice still requires a comprehensive, good faith investigation that uncovers "all relevant facts relating to the misconduct," which will then be subjected to a vigorous trust-but-verify review. Anything less than a wholly thorough investigation can be, and had been, viewed as misleading and sanctionable—particularly when inaccurate results are released in public statements.

Further complicating modern ESI investigations is the trend toward an increasingly remote workforce. While remote activities are nothing new, the rising proliferation of remote personnel and devices simply increases the need for careful, streamlined coordination of every component of the investigation, beginning at the earliest stages of the process. In turn, data management and data security considerations have become paramount, as independent, local, and often personal devices take the place of traditional centralized enterprise repositories.

Against this backdrop, this white paper outlines several key considerations to designing and implementing an efficient, effective ESI investigation that will quickly lead to the key facts and critical evidence. From planning—to ensure prompt access to the requisite ESI and streamline the process for making the data available—to focused review techniques that will locate the pertinent data quickly or reasonably establish that the data simply doesn't exist, this paper presents strategies and tips that investigators are using to optimize modern, often disperse, ESI investigations.

Control data proliferation for security and access

In a perfect world (for ESI management, at least), an organization's data would be stored in one place, and would be secure and readily accessible for collection and review.

In reality however, that has never been the case. Some fraction of the modern workforce has always been mobile, not only requiring remote access to enterprise locations and data, but also typically generating local repositories that need to be considered, and may well be implicated, in the context of an investigation.

As the remote workforce expands, however, enterprise access and local repositories become more of a direct and cognizable strain, and more of a risk, than ever before. This new reality requires a more considered approach to data management and access—one that recognizes and addresses the need for both the security and timely availability of the ESI being created, stored, transferred, and managed remotely.

While there are any number of viable solutions, a virtual desktop infrastructure provides a single approach that can address both problems. Security can be optimized by (1) controlling access to the virtual desktop, and (2) prohibiting local, endpoint storage of electronic data by maintaining ESI on enterprise servers instead. And, since user activities actually take place server-side rather than on the local machine, access is controlled by and largely restricted to the organization.

The benefit to an investigation of an approach that consolidates and centralizes ESI storage is obvious, and such an approach serves the dual objectives of timeliness and comprehensive coverage. With data primarily in one location, identification and collection can be expedited, and standardized protocols can be adopted for consistency. Likewise, the potential for missing ESI that may be directly pertinent to the investigation is minimized.

Ultimately, ESI investigations are driven by the data, so every effort to make data thoroughly and quickly available will improve efficiency and effectiveness.

Establish appropriate policies for remote operations

Even if steps are taken to maintain centralization and control of enterprise data, the reality of a remote workforce is that personal devices will be used to some extent to access and record information. It is important to recognize this reality, and implement policies and procedures that will facilitate prompt access to ESI when an investigation becomes necessary.

The scope of personal devices that are used to access organizational data is always expanding, but the current mobility profile requires consideration of at least three types of devices: smartphones, tablets, and personal computers/laptops. And a remote workforce will continue to unintentionally blur the lines between personal and workplace devices, which may counsel in favor of expanding coverage, or at least providing an explicit notification regarding organizational access to even unanticipated personal devices that are used for business purposes. It is not difficult to envision employees using the personal voice-activated virtual assistant sitting in their home office (Amazon's Alexa, for example) for scheduling, initiating phone calls, or even note taking. All of that ESI may well be fair game in an investigation, so it should be considered and encompassed by reasonable policies that advance organizational interests.

There are any number of approaches to balancing the inherent right to privacy associated with personal devices with the legitimate organizational interest in critical data—BYOD (bring your own device); CYOD (choose your own device); COPE (company owned—personally enabled); and COBO (company owned-business only) being among the more prevalent. While the particular approach should fit the organization, its operation, and its culture, every approach must establish the immediacy and breadth of the organization rights and interests, to ensure that any investigation can and will be swift and thorough.

In selecting the right approach, an organization should be cognizant of the practical realities of ESI collection from personal devices. The easiest, quickest, and most thorough technique is forensic collection of the entire device—all of the data is readily available, and organizational data can be targeted for availability in the investigation. However, when personal data becomes intertwined with organizational data on the same device, there is often some measure of resistance to a full forensic image of a personal device. That often leads to employee-directed, presumably targeted, excision of ESI, which will typically be

more expensive and time-consuming but, more importantly, might well impair the integrity of the collection, and in turn, the credibility of the investigation.

The critical point of any mobility or remote operation policy is the mandatory recognition that any organizational information belongs to the organization, regardless of the level of personal interest in the device. And that information must be wholly and readily accessible from the very moment an investigation is initiated.

One other point bears mentioning, even though it might be viewed as almost the antithesis of data management and access—ephemeral messaging. Modern mobility, especially international mobility, increasingly relies on ephemeral messaging applications—communications promptly and automatically vanish before collection and preservation are possible. Organizational policies need to address and control the use of ephemeral messaging in the investigations context, especially in light of stringent legal hold obligations, and particularly in recognition of the apprehensive view of ephemeral messaging held by many of the regulatory agencies responsible for initiating investigations.

Consider seamless end-to-end cloud-based capabilities

Every investigation brings with it a sense of urgency. The Board of Directors needs an answer before the current reporting period ends. Furtive harassment needs to be uncovered and countermanded immediately. A regulatory agency needs immediate assurance that internal procedures will suffice, otherwise a compliance investigation will

be initiated. Whether minimizing risk or maximizing reward, optimizing the "time to results" is critical for any investigation to be effective.

Optimizing time to results means looking at the entirety of the investigation process, and taking advantage of every opportunity to improve and expedite workflow. And the transition to an increasingly remote workforce provides a clear opportunity to focus on workflow components that can be leveraged in every investigation situation.

Managing remote investigations, particularly during the COVID-19 isolation period, highlighted the benefit of engaging a full-services team that could provide seamless end-to-end, coordinated services, rather than being forced to quickly cobble together a segmented cadre of independent providers. Collection is often the first physical activity undertaken during the course of any investigation. To expedite the process, there should really be a direct (and battle tested) line of communication between the investigation team and the collection expert, and a predefined, seamless pipeline to move collected data into a hosted environment to ensure prompt availability of ESI for review. Combining that pipeline with direct experience with the host analytics platform means that an investigation can begin almost as soon as the data is collected. No time is wasted in arranging for, training or coordinating with disparate providers, or coming up to speed on the most efficient and effective way to locate the documents necessary to respond to the information needs underlying the investigation.

Building on the comprehensive team approach to an investigation, taking full advantage of cloud capabilities (particularly for data transfer and access), is another way to enhance the investigation workflow. Modern collection tools can simply be pointed at the appropriate document collection, and the documents will automatically be uploaded to the cloud in the most nonintrusive and expeditious manner. From there, the documents can seamlessly be transferred to a review and analytics platform, and made available to any number of members of the investigation team for analysis.

And the availability of the cloud approach underscores the need to avoid traditional in-house investigation techniques, particularly given the obligation to be sufficiently thorough.

Typically, in-house investigations historically relied on the IT department to apply search strings to locate emails for review. Not only have studies shown that Boolean search is not an effective way to locate pertinent documents, but the possibility of missing critical acronyms, product codes and formulaic designations is obvious—not to mention overlooking an entire category of electronic documents in, for example, a network share or collaboration platform. In order to avoid this potential impediment to a thorough investigation, the scope of document collection should be designed expansively, to do little more than eliminate documents that are virtually certain to be irrelevant. This not only ensures coverage, but also avoids the need to constantly go back to the well to have the IT department collect documents with new, unanticipated search strings that were uncovered during the investigation—which, in turn, will advance the timeliness objective.

In the same vein, in-house investigations focusing on emails often relied on a single reviewer using nothing more than the search capabilities of, for example, Outlook, to locate relevant documents. In reality, given the limited analytics capabilities of Outlook (which is, after all, an email platform, not a review and analytics platform), that meant reviewing the entire document collection. The implications on the time to results are obvious, and can be completely abrogated by using a cloud-based review and analytics platform to succinctly focus the investigation.

Recognize the differences between an investigation and a production review

Most of the discussion surrounding typical investigations focuses on best-practices for planning and conducting employee interviews. However, the ESI investigation component, specifically, evaluating the evidence within a collection of electronic records, is an equally critical component of the entire investigation process—finding what some refer to as the "truth serum" for controlling those interviews and structuring much of the investigation. Indeed, the need for preparation through early assessment of the documents is becoming even more critical, as an increasingly remote workforce makes it difficult to conduct adequate personal interviews sufficiently in advance of the undesirable dissemination of information (or, even worse, misinformation) through the rumor mill.

The approach to reviewing documents for an internal or regulatory investigation differs significantly from a typical litigation production context. Recognizing this difference and the unique challenges of an investigation is the key to designing an efficient and effective document review protocol. Sometimes, however, documents themselves are the subject of the investigation, for example when responding to a civil investigative demand. Later, this white paper will cover an investigatory protocol for "proving a negative" and demonstrating to a requesting authority that, to a reasonable statistical certainty, there simply are no responsive documents.

In either situation, developing an effective ESI investigation protocol begins with recognizing the critical distinctions between a document review for an investigation and a review for production in litigation.

The objective of a typical litigation review is to proceed, from a reasonably known set of facts, to locate most of the relevant documents relating to the dispute, with the least amount of review effort. The emphasis is on document review, primarily to present the best documents for review and determine whether those documents relate to the underlying fact pattern. To that end, a litigation review is loosely designed to develop a model of positive, or relevant, documents and find most of the similar documents quickly, to the exclusion of other documents.

In an investigation, those facts are either not known or not well developed. As a result, an investigation review is crafted to quickly find pertinent documents that will establish that fact pattern or otherwise answer the critical information needs. It is not necessary to locate all, or even most, of the documents that may ultimately be relevant to the ultimate fact pattern. It is most important to be certain that the critical documents are available for review and to locate those documents quickly. An investigation is an effort to find the pieces of a puzzle and put them together to define a cohesive fact pattern.

Given this difference in objectives, there are several steps that can be taken to refine and implement a document review protocol to achieve the objectives underlying a compliance investigation.

Preserve and collect immediately, expansively—and discreetly

In an investigation, there may be little to go on and investigators likely will not know exactly who is involved or the precise circumstances. An investigation typically starts with some manner of tip or complaint, which can be written or verbal, and contains varying levels of detail. The complaint typically leads to the identification of some limited number of potential document custodians who are likely to have at least some level of knowledge of the facts surrounding the complaint. It is critical to quickly leverage the knowledge of those known custodians to expand the scope of the investigation.

Since time is of the essence, an automated legal hold application, often integrated with remote collection tools, can expedite the investigation process. Automated legal hold and forensically sound collection tools offer the opportunity to quickly and easily elicit information from those custodians and simultaneously collect documents for review. Automated legal hold tools typically include the ability to issue questionnaires to known custodians. In the investigation context, these questionnaires can be structured to quickly and efficiently elicit substantive information about the complaint from all of the known document custodians at the same time their documents are being collected. That information can then be used to scope and focus the document review even before the custodians can be interviewed. At the same time, as new information surfaces, investigators can continue to define potentially relevant data sources, work with IT to defensibly preserve those sources, recover deleted data, gain access to password-protected files and identify documents and, often, system artifacts, to piece together a chain of events. Also, when discretion is necessary, collection tools can run silently in the background without ever alerting the employee.

The critical consideration for an effective collection is to ensure that it is sufficiently expansive to encompass any documents that may be necessary to answer the specific information need underlying the investigation. In addition to making the collection effort more efficient (since there will not be any need to spin up resources more than once), an expansive collection will in fact make the investigation itself more efficient, since it will be more likely to provide the investigation team with the documents necessary to get the full fact picture, rather than leaving evidentiary holes that complicate analysis.

Use communication analytics to locate additional witnesses

The success of any investigation depends on the ability to quickly identify key witnesses and document custodians in order to unearth important details and develop the fact pattern as completely and early as possible. Including witness identification as a specific component of the document review process will provide exponential returns. The identification of more witnesses will lead to the collection of more documents, which will in turn lead to the identification of more witnesses.

With the information obtained from the legal hold questionnaires and ongoing interviews, state-of-the-art communication analytics can expedite identification through the document review process. There are several levels of communication analytics that should be used in tandem. Top-level analytics typically provides a macroscopic view of the entire social network of communications across a document population. Once critical individuals have been identified through the social network overview, the analysis can focus on their individual communication patterns. Then, using analytics to drill even deeper into the communications between specific individuals, the document review process can quickly uncover witnesses that can be integrated into the interview and document collection process. These new witnesses will similarly provide additional insight into others, ensuring a comprehensive investigation.

Use efficient machine learning techniques

Technology-assisted review (TAR), a form of machine learning also called predictive coding, is widely recognized as a valuable and effective approach to document review in the litigation context. Implemented properly, TAR can be an equally effective means of locating critical documents during the course of an investigation.

Given the differences in a litigation review and investigation review, it is important to choose an effective TAR protocol. Some TAR tools, which will be discussed later, require the entire document collection to be available at the outset and then substantial training to develop their models before review can begin in earnest. While that may be effective in a litigation review, the exigencies of a compliance investigation require review to start at the earliest possible moment—well before all of the documents have been collected.

TAR tools that use true continuous active learning (CAL) protocols avoid this initial delay and actual document review can begin with the very first document. The operation of CAL, which uses every review decision to improve the algorithm, will prioritize the best documents for the earliest review. As documents are added to the review, continuous active learning tools will incorporate them into the collection on the basis of the current training. This immediate, prioritized approach to review makes continuous active learning particularly suitable for investigations.

Another benefit of continuous active learning is the ability to initiate training with virtually anything. Since little is often known at the outset of a compliance investigation, it can be difficult to quickly locate truly pertinent documents that can be used to train a TAR tool. With CAL, training can start with a single, synthetic seed, which is a document created from whole cloth that encompasses all of the known concepts that would make a document relevant to the investigation. CAL will immediately recognize the words and phrases that underlie those concepts and prioritize similar documents for review, getting to the relevant documents quickly without even knowing where to really start.

To make the most efficient use of an appropriately sophisticated TAR tool, the document review can and should be segregated into multiple simultaneous lines of inquiry. For example, there may be several witnesses scheduled for successive interviews in a very tight window. Or there may be several discreet information needs for which evidence is being sought. To be optimally efficient in either situation, the document review should be structured to permit separate and simultaneous reviews to prepare for each interview, or research each information need, independently. With that review protocol, it is imperative that the TAR tool:

1. Permits simultaneous, independent review projects.
2. Uses all of the review decisions to train the algorithm, regardless of the project in which those decisions are made.

This type of approach can be critical, especially in multilingual investigations that utilize separate review teams for each language but require prioritization for review without regard to which language appears in the documents.

Structure the investigation team with collaborative independence

As a corollary to the efficient use of TAR, it is often best to structure the investigations team with "collaborative independence." Each team member effectively assumes independent responsibility for analysis and assessment of specific information needs. But every team member is additionally responsible for knowing and understanding the substance of all information needs, recognizing documents relevant to those needs, and constantly collaborating with other teams members to ensure the comprehensive knowledge base of the entire team.

In practice, this approach requires the investigations team to be virtually connected by a secure, collaborative instant messaging tool. In that way, team members can easily and immediately pass observations on to, or pose questions to, other team members as the investigation is progressing.

Effectively explore the unknown

When starting from scratch in an investigation, investigators may worry that a limited understanding of the situation caused them to miss a key document. A nagging concern in reviewing documents, especially in an investigation where the knowledge boundaries are blurred and ever-expanding, is how to be comfortable that there is nothing in the document population that is pertinent but unknown. When a document review focuses purely on what is perceived to be within the current scope of the inquiry, there is a very real possibility that potentially relevant documents that will help to define the full fact pattern will be missed.

Certainly, advanced analytics can be used to ferret out those unknown facts and documents. But, that can be a very painstaking and time-consuming undertaking and most compliance investigations simply do not have the luxury of time.

To solve this problem, many modern TAR tools include functionality that is directed at locating documents that are contextually diverse from everything that is known to that point in time. Contextually diverse documents obviously may or may not be relevant to the investigation, but the more contextually diverse documents that are seen over the course of the review, the less likely that the review and, in turn, the investigation, misses critical issues that are unknown at the outset.

But, what if there are no relevant documents?

Using these techniques and taking maximum advantage of appropriate technologies will ensure an efficient, effective, thorough document review in the investigation context, with commensurate results. Sometimes, however, there simply are no documents to be found. When documents are the object of the investigation, as in governmental and regulatory investigations, that conceivably means reviewing the entire document population only to come up empty-handed. The next section discusses techniques and technologies to short circuit that review process and still demonstrate that there are no documents in the collection, essentially "proving a negative" without reviewing the entire collection.

"Prove a negative" using TAR based on continuous active learning

What does it mean to "prove a negative"? The objective of an investigation is most often to quickly locate the critical documents that will establish a cohesive fact pattern and provide the materials needed to conduct effective personnel interviews. In that situation, the documents are merely a means to an end.

Occasionally, however, the documents become an end unto themselves. For example, governmental agencies often use civil investigative demands (CIDs) to investigate allegations of potential statutory liability. In that context, the documents themselves become the object of the investigation. While those documents may well have downstream utility, the emphasis of the document review in responding to the CID is purely on locating any responsive documents.

There may be situations where there simply are no responsive documents to be found. With modern electronically stored information (ESI) collections that total in the hundreds of thousands, or even millions of documents, a linear review of that magnitude can be prohibitively expensive and time-consuming.

Alternatively, it is possible to leverage advanced analytics, CAL and statistics to review only a fraction of an ESI collection, yet demonstrate that there are potentially so few responsive documents in the collection that a full-blown review would be entirely unreasonable. That is what is meant by proving a negative—undertaking an aggressive effort to locate responsive documents, finding none and using statistics to demonstrate the virtual absence of responsive documents.

What are the benefits of using TAR based on continuous active learning to "prove a negative"?

Three principal TAR protocols can be used to enhance a document review: simple passive learning, simple active learning and continuous active learning. Because of the way these different protocols train the underlying algorithms, only CAL protocols are effective in proving a negative.

As discussed in greater detail below, the objective in proving a negative is to make every possible effort to find responsive documents, and the TAR protocol should advance that objective.

The only TAR protocol that effectively seeks out responsive documents throughout the review process is CAL. A simple passive protocol trains by passing random documents to the reviewer. A simple active protocol, on the other hand, trains by a process known as uncertainty sampling, which provides the "gray" documents to the reviewer. These are the documents that are right at the border between documents that look to be responsive and those that look to be non-responsive.

By comparison, CAL primarily uses a process known as relevance feedback to pass training documents to the reviewer. Relevance feedback uses everything that is known about the documents coded to that point in time to select training documents that are most likely to be responsive.

Using a CAL protocol leverages the TAR algorithm. Every document reviewed in the process is a document that the algorithm sees as most likely to be responsive. That approach advances the objective of finding responsive documents far more efficiently than one that relies on random or gray documents and, therefore, CAL is critical to proving a negative.

Use statistics to scope the review

The first step in proving a negative is to establish the statistical parameters that will set the margins of error for the review and, in turn, the number of documents that may have to be reviewed in the process. The expectation is that no responsive documents will ever be found, regardless of how many documents are reviewed. With that assumption, statistics will control the relationship between the number of documents reviewed and the margin of error. In other words, this the number of responsive documents that might exist in the collection.

There is no hard-and-fast rule for setting the statistical boundaries. Rather, the decision depends on the relationship between the value of finding any responsive documents and the cost of obtaining these documents. In essence, the decision depends on some measure of proportionality and is likely going to be negotiated with the requesting party.

As an example, consider a collection of 500,000 documents that is not expected to contain a single responsive document. Using a binomial statistical calculator (such as the one at statpages.info/confint.html), the margins of error can be evaluated for samples of one percent, two percent, five percent and 10 percent of the collection to establish a range of alternatives.

Sample	Documents to Review	Margin of Error (CI=99%)	Potentially Responsive Documents
1%	5,000	0.0009	450
2%	10,000	0.0005	250
5%	25,000	0.0002	100
10%	50,000	0.0001	50

With a range of alternatives, the relative cost and benefit of various sample sizes can be evaluated, and the number of documents to be reviewed can be negotiated and set accordingly.

Initiate a review for documents that are close to responsive using analytics

The objective in proving a negative is to make every conceivable effort to locate the precise documents that are not expected to exist in the collection. That means truly exploiting every available analytical approach to locating responsive documents while keeping in mind that the TAR tool will eventually do the heavy lifting. The investigation team should diligently look for the subject documents using every possible technique available through the review and analytics tool.

Since no approach is likely to locate responsive documents as none are expected to exist in the collection, the investigation should focus on finding documents that are contextually close to being responsive. These "close" documents will eventually serve as the best available training examples for CAL review.

Investigators can begin the process by using keyword searches that are carefully crafted to locate any responsive documents that might exist in the collection. Be sure to solicit any reasonable keyword searches from the requesting party. Doing so will not only enhance the potential for finding truly responsive documents, but also alleviate any concern on the part of the requesting party that the scope of the review might be too narrow. If a search returns too many documents, review a reasonable random sample across the entire hit population to establish a statistical absence of responsive documents.

Then, use advanced analytics to explore specific components of the collection that are most likely to contain responsive documents. For example, keyword searches can be refined to focus on the documents held by specific key custodians. Communication analytics can be used to identify email exchange patterns that may be pertinent to the investigation. There may be certain file types, e.g., Microsoft® Excel® files or Microsoft® PowerPoint® presentations, that are more likely to be responsive. Even associated metadata, such as the original file path for a document, can be explored in a diligent effort to find responsive documents.

This review should continue until all reasonable searches have been exhausted and between 20 percent and 30 percent of the total anticipated review effort has been completed. Doing so will initially establish the absence of responsive documents and provide a reasonable starting point for training the CAL algorithm. It is important that these efforts be recorded, should it be necessary to explain and justify the process down the line.

Surface any truly responsive documents using TAR based on continuous active learning

Once the analytics review is reasonably complete, continuous active learning can structure the remainder of the review. The CAL algorithm will efficiently analyze the entire collection to locate any documents that are contextually similar to "close" documents located during the analytics review and will continuously learn from every coding decision made along the way.

Synthetic seeds can be used to optimize the CAL training regime from the analytics review. Investigators can draft an electronic document that reflects the specific content of a document that would be considered responsive if it existed within the collection. Import the document into the collection, being careful to include some designation, such as a unique Bates identifier, that makes it easy to identify and mark the synthetic seed as responsive. This will provide the continuous active learning algorithm with a very clear example of the precise language that makes a document responsive.

As with the keyword search process, a synthetic seed may be solicited from the requesting party as well. Doing so will ensure that the CAL algorithm will recognize, and elevate for review, documents that are contextually similar to specifically what the requesting party is seeking.

Make sure that some fraction of the documents reviewed during the CAL process are contextually diverse from the responsive synthetic seeds and the "close" documents identified in the analytics review. Contextual diversity functionality is critical in proving a negative, as it ensures a thorough exploration of the entire collection.

Presumably, the CAL review will not locate any responsive documents, since they are not expected to exist within the collection. As with the analytics review, documents that are close to being responsive should be coded as positive in order to continuously surface any contextually similar documents and maximize the potential for finding truly responsive documents.

Use the review and statistics to "prove a negative"

Assuming no responsive documents have been located during the review, the underlying statistics can be used to essentially prove a negative. Obviously, without reviewing the entire collection, there is no way to be certain that it contains no positive documents. What can be said, however, is that there are a very limited number of responsive documents that might exist in the collection. From the above example, a review of 25,000 documents using this process would mean that there are likely no more than 100 responsive documents in the entire collection.

Although that analysis is not based on a purely random statistical sample, this review process requires much more thorough effort to find positive documents. By using analytics and continuous active learning and including contextually diverse documents in the CAL review, this process optimizes the likelihood of finding a responsive document in the collection, if one exists. Since no responsive documents have been found in the review, the likelihood that a responsive document exists elsewhere in the collection is, for all practical purposes, even less than if the review had been random.

Altogether, this process is a reasonable way to demonstrate the absence of responsive documents in a collection without having to review the entire collection and to do so in a way that is even more stringent than a random review.

QC every investigation with TAR based on continuous active learning

An investigation that is intended to "prove a negative" obviously relies on TAR (specifically, continuous active learning) to ensure that the review was sufficiently thorough to return the intended results. But continuous active learning is equally essential to ensure the quality of an investigation that indeed does find the relevant documents and answer the underlying information needs.

Every investigation will generate a set of documents containing the evidence that was needed to answer every information need. Each set may well differ from the other, but they will be internally consistent. In other words, the documents that answer any specific information need are likely to be reasonably similar from a substantive and linguistic point of view—they will all contain similar language addressing the same point.

This internal consistency means that these documents will serve as a very reasonable training set for a continuous active learning tool. Using the documents to seed the continuous active learning algorithm will, in turn, generate a relevance ranking that elevates documents that are most like the seed documents to the top of the ranked list. The investigation team can then review the documents at the top of the ranked list to ensure that there is no new information, and certainly nothing to controvert the team's factual findings. Ensuring that the top-ranked documents are redundant of the substance of the documents already found by the team provides an added level of comfort that the review was sufficiently thorough to elucidate the current answers to the information needs, and develop the appropriate evidence.

Conclusion

When hit with an unexpected information request, with little time to react, organizations need to quickly size up the task and learn the facts. Today, every ESI is intensely demanding, with new added complexities brought on by the proliferation of a remote workforce and devices. These challenges necessitate the need for proactive, careful, streamlined coordination of every component of the investigation, beginning at the earliest state of the process.

OpenText™ Recon, a dedicated investigations service providing seamless end-to-end support, helps enterprises and outside counsel rapidly compile actionable intel, identify relevant ESI, and quickly unearth the information that will answer critical questions. Led by a team of lawyers, data scientists, linguists and technologists experienced in unstructured data interrogation, the Recon team provides rapid insight for swift decision-making and resolution:

- Litigation assessment ("…is the wind at my back, or in my face?" "Should the case proceed or settle for reasons that may or may not be known at the outset?")
- HR investigations
- Internal investigations
- Compliance investigation
- Government and regulatory investigations
- C-suite vetting ("Is this the right choice to run the company?")
- M&A due diligence ("Are we sure no one cooked the books?" "Should any terms of the deal be revised in this limited window of time?")
- Prove a negative ("Can we thoroughly and statistically demonstrate that there is no evidence to support the allegation?")

opentext™

The Recon investigations service will find and deliver the answers—even the unknown unknowns—in a fraction of time of a production review to asses risks, strengths and potential liability, while reducing costs and improving efficiency of traditional litigation reviews.

About OpenText

OpenText, The Information Company, enables organizations to gain insight through market leading information management solutions, on-premises or in the cloud. For more information about OpenText (NASDAQ: OTEX, TSX: OTEX) visit: opentext.com.

Connect with us:

- OpenText CEO Mark Barrenechea's blog
- Twitter | LinkedIn

opentext.com/contact