opentext™

White paper

When a compliance investigation hits: How to find the facts swiftly

No company wants a government or regulatory investigation, as it allows the investigating agency to open a sweeping inquiry that can encompass a company's entire document repository. This white paper discusses how to move forward swiftly and compliantly, while minimizing time, effort and cost.

opentext™

Contents

opentext™

In today's heightened global regulatory environment, regulatory compliance has never been more important, or as difficult, to attain. Unwitting employees or bad actors within and outside of even the most vigilant organizations seek to gain and use sensitive information to their advantage.

Whether potential malfeasance is triggered by a whistleblower's tip-off related to Foreign Corrupt Practices Act (FCPA) compliance, suspicions of deceptive sales and marketing practices or hints of trade secret or other intellectual property theft, no company wants an investigation.

When signs of potential violations are triggered, the investigating agency can open a sweeping investigation that can encompass a company's entire document repository—draining an incredible amount of time and resources to find facts quickly for swift action or resolution.

However, too many corporations and their law firms approach litigation and compliance investigations the same way, using the same technology, approach and people. The approach to managing electronic information in internal and regulatory compliance investigations should differ from litigation.

This white paper outlines the key differences in litigation and investigation review, and presents strategies and tips that investigators are using to design and execute an efficient and effective document review protocol. It also offers best-practice techniques for "proving a negative"—that a document responsive to a governmental or regulator investigation simply does not exist—without reviewing the entire document collection.

How an investigation review differs from a production review

Most of the discussion surrounding compliance investigations focuses on best-practices for planning and conducting employee interviews. However, document review, specifically electronic document review, is an equally critical component of the investigation process, finding what some refer to as the "truth serum" for controlling those interviews and structuring much of the investigation.

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 a compliance 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 a document review 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 document review 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. 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—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 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, a legal hold application, often integrated with 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.

Use communication analytics to locate additional witnesses

The success of a compliance 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.

opentext™

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 a compliance 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 compliance 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. To be optimally efficient, the document review should be structured to permit separate and simultaneous reviews to prepare for each interview 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.

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 a compliance 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 compliance 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.

Apply advanced analytics and continuous active learning to "prove a negative"

What does it mean to "prove a negative"? The objective of a compliance 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

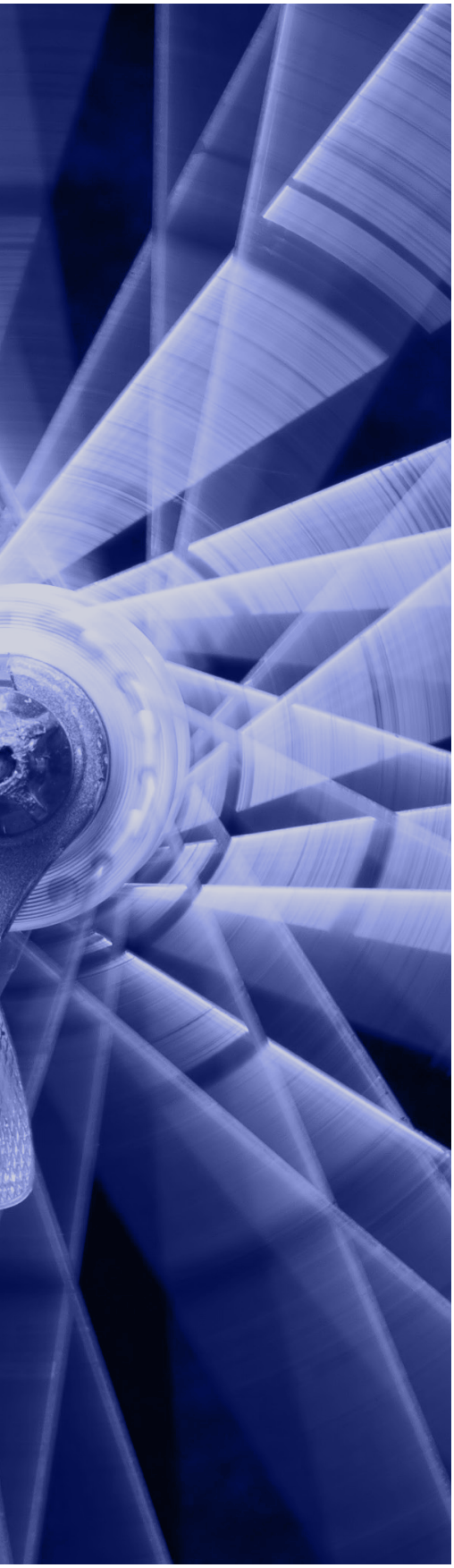
With a range of alternatives, the relative cost and benefit of various sample sizes can be evaluated, and the number of documents to be reviewed can be negotiated and set accordingly.

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.

Since no approach is likely to locate responsive documents as none are expected to exist in the collection, the review 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 continuous active learning

Once the analytics review is complete, continuous active learning can complete 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.

opentext™

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.

Conclusion

Due to increasing regulatory burdens and a rise in compliance infractions, organizations need to be able to find the facts swiftly that tell the story—or be able to prove, defensibly, that no "story" exists. By understanding critical distinctions between a document review for an investigation and a review for production in litigation, then employing best-practices technology, process and expertise designed specifically for an investigation review, organizations can maximize the value of their time while ensuring an efficient, effective and thorough document review.

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 Discovery solutions
- Twitter | LinkedIn

opentext.com/contact