Consilio Institute: Practice Guide

AN EMBARRASSMENT OF RICHES: ANALYTIC TOOLS AND TECHNIQUES

Matthew Verga
Director of Education
Xavier Diokno
Sr. Director Innovation Solutions

Consilio / ADVANCED LEARNING INSTITUTE

Consilio Institute: Practice Guide

AN EMBARRASSMENT OF RICHES: ANALYTIC TOOLS AND TECHNIQUES

CONTENTS

Disclaimers

The information provided in this publication does not, and is not intended to, constitute legal advice; instead, all information, content, and materials available in this publication are provided for general informational purposes only. While efforts to provide the most recently available information were made, information in this publication may not constitute the most up-to-date legal or other information.This publication contains links to third-party websites. Such links are only for the convenience of the reader; Consilio does not recommend or endorse the contents of the third-party sites.

Readers of this publication should contact their attorney to obtain advice with respect to any particular legal matter. No reader of this publication should act or refrain from acting on the basis of information in this book without first seeking legal advice from counsel in the relevant jurisdiction. Only your individual attorney can provide assurances that the information contained herein – and your interpretation of it – is applicable or appropriate to your particular situation.

Use of this publication, or any of the links or resources contained within, does not create an attorney-client relationship between the reader and the author or Consilio. All liability with respect to actions taken or not taken based on the contents of this publication is expressly disclaimed. The content of this publication is provided "as is." No representations are made that the content is error-free.

DIGGING FOR TREASURE

The tide of data never stops rising, and the types and sources of data never stop multiplying. Never have there been so many communication devices, apps, and services available. Never have there been so many ways to collaborate with others and generate electronically-stored information (ESI). Unfortunately, that also means there has never been more data that legal practitioners must somehow find a way to analyze and review. Finding a way that is efficient and effective requires understanding the range of tools and techniques available to you so you can pick the right tool for the right job.

Ethical Requirement

Beyond just being essential to the efficiency and efficacy of the effort, understanding how to go about sorting, filtering, and searching ESI is also a required part of attorneys' duty of technology competence for eDiscovery. In August 2012, the American Bar Association (ABA) implemented changes[1] to its Model Rules of Professional Conduct, including a change making the need to maintain technology competence explicit. Since then, that requirement or a variation on it has been implemented in forty states.[2] Understanding how to search effectively for the right data is a key part of fulfilling that requirement, and it was among the nine core skills originally identified by California's Formal Opinion No. 2015-193.[3]

ABOUT THIS PRACTICE GUIDE

This paper will review key things practitioners need to know about the range of analytic tools and techniques available to them for ESI analysis and review. We will start by reviewing use cases and goals. Then, we will review the range of tools and techniques. Finally, we will review how to put it all together.

USE CASES AND GOALS

There are three main use cases for these analytic tools and techniques and a variety of goals or priorities you may wish to pursue.

Early Case Assessment

The term early case assessment (ECA) originally referred to reducing uncertainty about the risks and costs associated with a new legal matter by quickly making a preliminary assessment of the evidence, facts, and law to inform decisions about how to proceed. Today, it also encompasses early data assessment, which is focused on evaluating the composition and completeness of collected ESI, and review preparation, which includes tasks like testing and refining searches and filters, evaluating potential workflows, and estimating needed resources. The intersection of these three connected-but-distinct activities makes the ECA phase of an eDiscovery effort one in which many different analytic tools and techniques can be useful.

[1] Debra Cassens Weiss, *Lawyers Have Duty to Stay Current on Technology's Risks and Benefits, New Model Ethics Comment Says*, ABA JOURNAL, http://www.abajournal.com/news/article/lawyers_have_duty_to_stay_current_on_technologys_risks_and_benefits/ (Aug. 6, 2012).

[2] Robert Ambrogi, *Tech Competence*, LAWSITES, https://www.lawsitesblog.com/tech-competence (last visited July 2, 2021).

[3] The State Bar of California Standing Committee On Professional Responsibility and Conduct, *Formal Opinion No. 2015-193* (June 30, 2015), available at https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL 2015-193 %5B11-0004%5D (06-30-15) - FINAL.pdf.

Document Review

After ECA, document review itself can also benefit extensively from the application of analytic tools and techniques. As noted above, such tools and techniques can be used to plan and estimate, but they are also essential for organizing and prioritizing materials for review. Moreover, full application of a technology-assisted review or continuous active learning workflow can dramatically improve the efficiency and efficacy of a document review effort.

Investigations

In investigations, whether internal or agency-initiated, organizations face discovery needs similar to those they face in litigation. ESI that is both voluminous in scale and diverse in type and source must be analyzed and reviewed in an efficient and effective way. Beyond that, investigations add the additional challenges of tighter timelines, of limited negotiability (for agency investigations), and potentially, deliberate obfuscation by bad actors. These challenges only increase the importance of leveraging the right tools to facilitate efficient analysis and nuanced review.

Goals

Depending on the use case and the specific situations, you may wish to pursue any or all of these goals:

- Assessment of a case on its merits to determine risk/how to proceed
- Rapid identification of hot documents for an internal or agency investigation
- Assessment of a collection's completeness (both within and across custodians)
- Estimation of volumes, prevalence, etc. for review and production planning
- Testing of sorting, filtering, and searching to identify relevant materials
- Efficiently conducting review (organizing, prioritizing, reduction through TAR/CAL)

Thankfully, in most document review platforms, practitioners have a powerful set of tools and techniques at their disposal for pursuing these myriad goals. The specific bells and whistles vary, but generally, there will be some kind of random sampling tools, searching and filtering tools, structured analytics tools, conceptual analytic tools, and technology-assisted review workflows. Beyond those options, there are other new and developing tools that may be available too. Let's discuss each of these tools and techniques.

SAMPLING TOOLS AND TECHNIQUES

One of the most powerful tools in your toolkit is sampling. There are a lot of ways to find materials you expect to be in a collection of ESI, but sampling is a terrific way to also find materials you didn't know to look for: the unknown unknowns. For our purposes, sampling comes in two flavors: judgmental sampling and formal sampling.

Judgmental sampling is the informal process of looking at some randomly selected materials to get an anecdotal sense of what they contain, whether that's sampling from a particular source, from a particular search's results, or from a particular time period. You're not reviewing a particular number of documents or taking a defined measurement with a particular strength; you're getting an impression and making an intuitive assessment.

Formal sampling is just the opposite: you are reviewing a specified number of randomly-selected documents with the goal of taking a defined measurement with a particular strength. Typically, that measurement is either of how much of a particular thing there is within a collection (*i.e.*, estimating prevalence) or of how effective a particular search is (*i.e.*, testing classifiers).

- **Estimating Prevalence** - Estimating prevalence is the process of reviewing a simple random sample of a given collection of materials to estimate how much of a given kind of thing is present. You might estimate the prevalence of relevant materials, of privileged materials, or of materials requiring redaction or other special steps. The size of the sample you need is dictated primarily by how precise you want your estimate to be (*i.e.*, margin of error), and how certain about it you want to be (*i.e.*, confidence level), and to a lesser extent, by how large your collection of materials is (i.e., sampling frame). Most often you will be dealing with sample sizes of a few thousand (e.g., a sample of 2,345 for a confidence level of 95% and a margin of error of +/-2% in a collection of 100,000 documents).
- **Testing Classifiers** - Testing classifiers is the process of seeing how effective and efficient a particular classifier – typically a search of some kind – actually is. Using this technique, you can estimate how much of what you're seeking a given search is likely to return (*i.e.*, recall) and how much irrelevant material is likely to get returned with it (*i.e.*, precision). These measurements are taken by running the searches against a control set, which is made be pre-reviewing and coding a sufficiently-large random sample. Comparing the search results to the already-completed coding allows for the iterative refinement of searches to increase their recall and precision before they are applied to the full collection.

SEARCH AND FILTERING TOOLS

After sampling, the next major category of tools and techniques available is search and filtering, including keyword and phrase searching, Boolean searching, fuzzy searching, conceptual searching, and more.

Searching

Searching, both on the internet and among our own emails, messages, and files, has become an inescapable part of everyday life. Almost all of this searching, like the searching you do in eDiscovery, is powered by some form of indexing. In the eDiscovery context, indexing is typically performed during the processing phase of the project.

Indexing is the process of creating the enormous databases that are used to power search features. Most common are inverted indices, which essentially make it possible to look up documents by the words within them. Inverted indices are like more elaborate versions of the indices you find in the backs of books. Decisions during processing about how indices should be generated and what common words (e.g., articles, prepositions) they should skip affect the completeness of search results you get. Searches can only find what indices show.

More sophisticated indices are created to power features like concept searching, concept clustering, and technology-assisted review, which we will discuss further below in our section on advanced analytic tools and techniques. The types of indices that are prepared and the specific features your software offers for working with them will dictate what types of searching are available to you.

- **Keyword and Phrase Searching** - Exactly as it says on the tin, keyword and phrase searching lets you search for a key word, for a phrase, or for lists of both at once. Just as with the basic internet searching we all use, if one of the desired keywords or phrases is present, the document will be returned. One key area of variation from tool to tool is whether wildcard characters can be used to find variations on words and, if so, how they can be used.
- **Boolean Searching** - Boolean search is the next step up in sophistication from basic keyword searching. It allows the use of operators such as "and," "or," and "not." These operators allow for the searcher to define specific relationships between key words and phrases to achieve higher quality results (*i.e.*, improved recall and precision). Other operators may be available, including proximity operators (*i.e.*, to find a particular word appearing within a certain number of words of another particular word).

  The range of specific operators available varies with the tools being used, as can their precise operation. Thus, it is important to understand the tools you are actually using to be sure you are searching the way you intend.
- **Fuzzy Searching** - Fuzzy searching (also sometimes referred to as approximate string matching or stemming) is another extension of basic keyword searching that may be available to you. Fuzzy searching allows a search to return variations on a word rather than just the precise word you searched (e.g., finding both invite and invitation). How much variation is allowed is typically an adjustable setting.
- **Conceptual Searching** - As noted above, conceptual searching is powered by different types of indices than traditional searching. Conceptual searching uses these indices to try to return results based on related ideas and topics rather than just based on whether the same specific words and phrases are used.
- **Other Tools and Features** - In addition to these core search functions, most review tools also offer a range of reporting and administration tools (e.g., saved searches, search history, etc.) to assist you in brainstorming, testing, and iteratively improving searches to meet your information needs. Many tools now also offer some form of word cloud or topical heat map feature to facilitate visual review of the most used words or phrases in your materials.

Filtering

In addition to your searching options, most platforms also offer you a range of options for sorting and filtering by specific properties of documents to help you surface what matters and prioritize what matters most. Most often this is based on a combination of metadata values extracted from the documents, such as file type and date, and custom-created metadata values, such as domain name or custodian.

Often, these types of sorting and filtering capabilities are now tied to visualization tools that let you see the distribution of materials (and any gaps in them) at a glance and that allow you to adjust a range of value limits to see how they narrow or expand your results. For example, many tools now offer communication maps that can show which people are communicating with each other, how often they are doing so, and other useful details.

STRUCTURAL ANALYTIC TOOLS

After sampling tools and searching and filtering tools, the next major type of tools available to aid your analysis and review are structural analytic tools that facilitate email threading, duplicate identification, repeated content identification, and textual near-duplicate identification.

Threading

Despite the rise of mobile and social sources, collaboration tools, and other alternative communication channels, email still remains a major component of most ESI collections for eDiscovery, and it tends to be voluminous. A single gigabyte of email can easily contain 5,000 to 10,000 discrete email messages, plus their attachments and embedded images and objects. Thankfully, email ESI also typically contains a significant amount of repetition and overlap that can be skipped.

For example, if you collect email from two custodians, you will have multiple copies of the email messages sent between them – a sender copy and a recipient copy for each one. Moreover, if they are engaged in a thread of replies to each other, the emails in such a thread may contain the preceding emails within themselves as quoted text, and the last one in the thread may contain the full text of the whole thread within itself. Such emails are sometimes referred to as inclusive emails, as are any standalone or offshoot emails that contain unique content or attachments.

Email threading tools typically offer some version of two functions to users: conversation threading and inclusive email identification. Additionally, as noted above, many now offer visualization features as an alternative way to explore the email threads and inclusive emails identified by the system.

- **Conversation Threading** - Conversation threading is a process in which emails are analyzed and automatically organized into thread groups, arranged chronologically. This analysis looks at existing conversation IDs, if available, and a range of email header fields and other document properties to match up replies in sequence. Such organization makes it possible to quickly identify related materials, speeding up investigation, and to quickly see the context surrounding a particular message, improving understanding. Additionally, presenting emails to reviewers as organized threads speeds up later document review.
- **Inclusive Email Identification** - Inclusive email identification is a process in which textual analysis is used to identify inclusive emails, *i.e.* those that contain a full thread within themselves or that otherwise contain unique text or attachments. Identifying the inclusive emails allows you to more quickly get the full picture, speeding up investigation, and when used as a filter, it can dramatically reduce the number of emails requiring later document review.

Duplicates

The operation of computer systems can produce a lot of duplicate files (including duplicate emails, as noted above). Although duplicates may need to be tracked and reported on in certain circumstances, they do not need to be examined. Such duplicate files are identified using a technique called hashing.

Hashing is a technique by which sufficiently unique "fingerprints" can be generated for files. Hash functions are mathematical processes that take irregular-length inputs (e.g., the data in a particular file), and use them to generate fixed-length outputs (e.g., a string of 32 numbers and letters). Hashing for duplicate identification is accomplished using a cryptographic hash function (e.g., MD5 or SHA-1), which is well-suited to matching unique inputs to particular outputs. Identical files produce identical hash values, and hash values can be easily compared by software to automatically identify matches across even a large collection of ESI.

- **Duplicate Identification** - Typically, collected ESI is hashed and deduplicated during or prior to the ECA phase of the project. But, because other rules (e.g., family group preservation) may override deduplication, some duplicates may remain. For example, if the same spreadsheet was attached to two different emails, neither copy would be removed. Most platforms provide features for identifying and managing such duplicates within your loaded collection of materials.
- **Repeated Content Identification** - In addition to identifying fully-duplicated documents, many platforms also offer some form of repeated content identification. Such features are designed to automatically identify frequently-repeated blocks of text (e.g., email signature blocks, automatic confidentiality warnings, etc.) so that they can be filtered out of search results (reducing false positives, particularly for privilege searches) and omitted from the creation of semantic/conceptual indices (improving the effectiveness of the semantic/conceptual analytic tools we will discuss below).

Near-Duplicates

In addition to true duplicates, it is common for collections of ESI to contain large numbers of near-duplicates. Near-duplicates are documents that are substantially similar to each other, but not truly identical (and therefore not removed during deduplication). There are two main types of near-duplicates that occur:

1. Superficially-identical documents that only vary in some metadata property, typically arising from their different sources or collection methods
2. Documents with some actual variation in content, like successive drafts of a contract

Finding the former reduces the number of documents to consider (and later review), while ensuring consistent treatment across duplicates. Finding the latter can provide valuable context to the development of key documents over time.

- **Near-Duplicate Identification** - Textual near-duplicate identification is somewhat more complicated behind the scenes than true-duplicate identification. Rather than comparing whole documents as single, abstracted values, the full textual content of the documents must be broken down into smaller pieces (sometimes called shingles). These small pieces can then be hashed and the sequences of those pieces compared across documents. If a sufficient number of pieces match, in the right order, the documents will be treated as near-duplicates. Typically, the threshold of similarity at which the system treats two documents as near-duplicates can be customized.

CONCEPTUAL ANALYTIC TOOLS

The next major tools available to aid your analysis and review are advanced analytic tools, powered by conceptual indexing and other advanced mathematical analyses, including concept searching, concept clustering, and categorization.

Conceptual Indexing

As we noted briefly above, there are more sophisticated types of indices than the traditional inverted indices used to power basic search functions. These conceptual indices (sometimes called semantic indices) analyze the available materials in a different way to power different kinds of features. Whether created by latent semantic analysis, probabilistic latent semantic analysis, support vector machines, or another related mathematical approach, these indices are designed to go beyond just listing all of the words in a document to reveal their conceptual content.

This analysis is accomplished mainly by analyzing the co-occurrence of unique terms across the collection of documents (e.g., how often does the term "fire" appear with the term "employee" and how often does it appear with the term "extinguisher"). This analysis of co-occurrences is used to create an n-dimensional map (like a traditional map of Cartesian coordinates, but with many more dimensions than just x, y, and z). The more frequently unique terms co-occur together, then the stronger the relationship between them, and the more co-occurring terms in two documents, then the closer to each other they will appear on the map. Dense clusters of such documents suggest key topical areas in the document collection (e.g., employee termination discussions in one area of the map and fire safety discussions in another).

Features Powered by Conceptual Indexing

Conceptual indices are used to power a variety of branded features in different eDiscovery software platforms, but regardless of name or variation, there are three key functions that are generally available:

- **Concept Searching** - Searching against a conceptual index does not require an exact match in the way that searching against an inverted index does. Instead, the terms or phrases you search are mapped onto the existing index and documents that are close enough to those search terms on the map will be returned as results – even if none of the exact terms you searched appear. Some concept searching features are referred to as natural language search features, and some also offer an option to search for more documents like a given example document, which may be a real sample or a synthetic one, created for the purpose.

  Another advantage of searches against these indices is that they can reveal more than just a binary, yes-or-no result. Because of the nuanced multidimensionality of the index, you can get results scored on how responsive or not responsive they are to your search (*i.e.*, how close or far away the result was on the map).

- **Concept Clustering** - Concept clustering is an automated, unsupervised process in which software analyzes the conceptual index that has been created. Rather than looking for the

closest matches to a user-provided search, the software looks for the densest clusters of related materials it has identified and groups those results together into clusters defined by their most frequently occurring terms. How dense a cluster must be to qualify is typically a customizable property. Those clusters can then provide an alternative way to explore a collection of documents, to learn about the scope of topics and range of materials it contains, and to identify areas for further exploration.

- **Categorization** - Categorization is akin to a hybrid between concept searching and concept clustering. It is a process in which a user selects a set of example documents to define a cluster for the software, and then the software attempts to find all the other documents that should go in that cluster with the examples provided. This is the functionality that powers some kinds of technology-assisted review.

TECHNOLOGY-ASSISTED REVIEW WORKFLOWS

Technology-assisted review is used to refer to a family of workflows that leverage categorization (or similar functions), in combination with sampling, to achieve a reliable document review process that requires significantly fewer hours of manual, human review than traditional all-manual approaches. Since its initial rise to prominence in 2011, the available array of TAR tools has expanded and evolved, and eDiscovery service providers have continued to develop new workflows to leverage them in useful ways for the diverse range of projects their clients face.

Although full deployment of a TAR workflow is typically part of the review phase of an eDiscovery project, these workflows – or limited versions of them – may also be leveraged to explore a collection during ECA, to organize and prioritize it for a more traditional review process, or to create a yardstick against which to measure a more traditional review process.

TAR approaches come in two major varieties, which we will refer to as TAR 1.0 and TAR 2.0:

- **TAR 1.0 – Predictive Coding** - TAR 1.0 refers to the initial, categorization-based workflows offered in eDiscovery – many of which were, and are, referred to as predictive coding. Broadly speaking, these workflows involve leveraging a sampling process to create a training set or seed set (*i.e.*, a user-defined cluster or clusters), which the chosen software than uses to find other similar documents. These results are then reviewed and coded, and that coding is used to improve the software's results. This training cycle is iterated multiple times until an acceptable quality of results is achieved. The effectiveness of the whole process is measured using either a previously prepared control set or an additional random sampling effort.

- **TAR 2.0 – Continuous Active Learning** - TAR 2.0 refers to more recent workflows developed to leverage new tools based on different mathematical approaches. Rather than being based on identifying the similarities in a large, prepared training set like categorization and TAR 1.0, these workflows are characterized by continuous active learning that updates relevance scoring and prioritization for all documents dynamically as each additional document is coded by a reviewer.

  This is accomplished by focusing on a single, binary classification (*i.e.*, relevant to topic X and not relevant to topic X) and analyzing the differences in language between successive, single example documents to identify the hyperplane that best divides the relevant examples from the non-relevant examples on a multidimensional map. Each additional example the software analyzes and maps can lead the software to identify a more efficient hyperplane between the two groups, improving its classifications.

These workflows emphasize speed over structure, and so, they work best in situations where there is a clear, binary classification decision to make and where family groups and other contextual factors are less important than overall speed.

NEW AND DEVELOPING TOOLS

The above tools and techniques are all well-established and widely-available. Software and service providers, however, are continually innovating new tools and techniques to address new and developing challenges. Some that you should be aware of are tools for PII analytics, entity extraction, image analysis, and generative AI.

PII Analytics

Personally identifiable information (PII) is defined by the United States Department of Labor as:

> Any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means.[4]

Common examples of this kind of information include: name, social security number, passport number, driver's license number, taxpayer identification number, patient identification number, financial account number, credit card number, VIN number, title number, street address, email address, telephone number, and biometric data. As more privacy laws and regulations are being promulgated in more jurisdictions, protection of PII is becoming ever more important, and PII frequently appears in the kinds of ESI collected for investigations and litigation.

PII analytic tools are designed to automatically identify certain kinds of PII wherever it appears in your collected ESI. There are two general ways the tools can work. They can operate based on pattern matching (e.g., finding text strings formatted like social security numbers), or they can operate based on algorithmic or AI analysis (*i.e.*, context aware; can identify more than just text string formatting). The more sophisticated tools can provide additional help, such as finding more types of PII in more types of data, finding custom defined types of PII, automatically associating identified PII with identified custodians or entities, and providing the basis for programmatic redaction of specified data strings.

Entity Extraction

Like PII identification, analytics tools can also identify and extract a broad array of entities. Entities often include people's names, addresses, email addresses,

[4] "Guidance on the Protection of Personal Identifiable Information," U.S. Department of Labor, available at https://www.dol.gov/general/ppii.

phone numbers, as well as locations, organization names, and product names. Entity extraction applications, which are also known as Named Entity Recognizers (NERs), apply artificial intelligence to classify various types of entities, such as a location, a person, or an organization.

For example, in the sentence "Fashion designer Ralph Lauren founded Ralph Lauren in 1967" the name "Ralph Lauren" is used to refer to both a corporation and a person. To differentiate these two uses, entity extraction applications analyze the context of the name, including the surrounding words and grammatical structure. In this case, the word "founded" enables the application to determine that Ralph Lauren was both referred to as person ("Ralph Lauren founded") and referred to as an organization ("founded Ralph Lauren").

Image Analysis

Image analysis tools enable users to find objects of interest within images or videos. Searching images for objects of interest (e.g., a person, a car, a passport photo, inappropriate content) often occurs when working with mobile phone data or security video (e.g., video of employees enters a building). Image analysis applications attempt to identify objects by examining the image's pixels. When a group of pixels form a pattern, the application can then isolate distinct objects within the image. For example, when analyzing images of people entering a building, the application uses the image's pixels to identify a person versus a delivery cart or a person's face and their purse or bag.

Image analysis applications do this by training on a very large library of objects that appear in our everyday environment (e.g., people, animals, trees, cars, etc.). This training allows the application to classify the objects that are detected in the image (e.g., a delivery cart, a person's face, a bag, or purse). When a user submits a search, such as an employee's face, the application can then compare the employee's face against all other faces that were extracted from the video.

Generative AI

Recent advances in AI have led to the creation of generative AI applications that can generate new content based on their understanding of existing content. Generative AI can respond to natural language queries, draft natural language responses, generate images, generate audio, and generate video, and numerous other generative applications are still in development or testing. The most widely known of these tools so far is ChatGPT, which attempts to understand natural language queries and draft natural language responses. It accomplishes this using something called a Large Language Model (LLM).

Large language models are tools trained on enormous collections of text – often, text scraped from the Internet (e.g., Wikipedia, Reddit, news articles, blogs, case law databases, etc.). LLMs allow generative AI applications to understand grammar and sentence structure, enabling them to predict the best next word in a sentence with a very high degree of accuracy. For a very simple example, a user could ask it to complete the phrase "peanut butter and," and based on its training across Internet data, the application would complete the phrase with "jelly."

In the context of discovery and investigations, tools based on generative AI may soon provide a way to explore your collected ESI for relevant information and materials using natural language queries, or an "assistant" to aid in deposition or motion preparations. Expanding from the simple example above, imagine submitting a more complex request, such as: "Generate a draft legal complaint for an Illinois state court by John Doe against Jane Doe for property damage to a fence arising out of a car accident on May 1, 2023, at 1060 W. Addison St. Chicago, IL." A properly trained LLM could make fulfillment of such a request possible.

Generative AI applications also have limitations, including providing inaccurate or biased results. LLMs can only be as good as the data sets on which they are trained, and data scraped from across the Internet contains all the inaccuracies and biases that are common online. Also, because LLMs simulate language rather than cataloging information, generative AI applications may also be prone to "hallucination" of information that sounds plausible but is not real. For example, one attorney using ChatGPT was provided with fake case citations, which they used, resulting in sanctions.[5] Due to these limitations, generative AI applications should still be used with caution during their nascency, and users should always validate and QC responses.

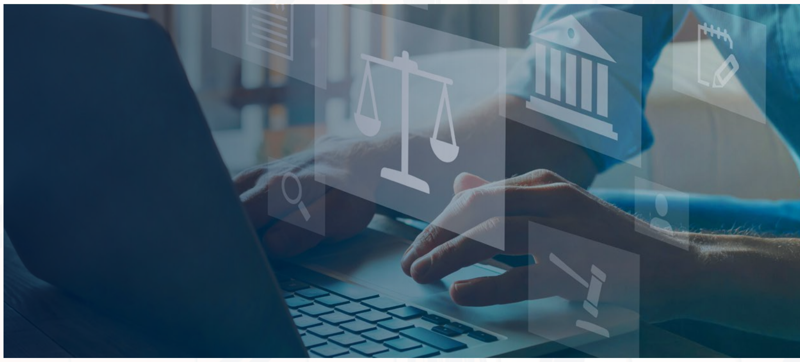
PUTTING IT ALL TOGETHER

Our survey of analytic tools and techniques has revealed a wide range of available options, each with different strengths and use cases, but achieving effective analysis and review is not a question of applying as many of these tools and techniques as you can. Rather, it is a question of selecting the right ones to best serve your current goal – whether that's traditional ECA, data assessment, investigation, or review preparation – and then building on those steps in a rational way to eventually achieve all your goals for the project.

Much like a telescope or microscope, the best result does not come from lining up as many lenses as possible. You must align the right ones on the right order to bring what you seek into sharp focus. Let's review some examples of how to think through these decisions.

For Pursuing Traditional ECA

When your top priority is pursuing traditional ECA, the first question to ask yourself is how much knowledge you have of what you expect to find. If you know a lot about what you're looking for in your ESI (e.g., from thorough custodian interviews, from overlap with prior legal matters, etc.), you may be able to jump right to searching for it. If you don't know a lot about the materials you're seeking, which is more common, you will want to start with one or more of the tools and techniques best suited to revealing unknown unknowns:



- Formal random sampling to estimate prevalence, which lets you see a cross-section of everything you have and some of all the different terms and phrases your custodians use to help you better plan your next ECA steps
- Visualization tools (e.g., communication maps, word clouds, etc.), which can reveal patterns of communication and behavior and assist with completing the picture of what happened in other ways
- Conceptual indexing features, which let you use concept searching to find relevant materials without knowing the best search terms, concept clustering to explore a cross-section of topics, and categorization to use a few relevant examples to find more
- Continuous active learning (TAR 2.0) workflows, which can rapidly surface relevant materials in certain circumstances

Once you start to get a handle on what you are really seeking (or if you already knew), you can transition from these initial, exploratory efforts to more targeted search and filtering efforts, which can quickly find relevant materials and hot documents. And, as you find relevant materials to review, thread and duplicate

[5] Jane Wester, "Judge Imposes $5K Fine on Lawyers Who Submitted ChatGPT-Generated Fake Case Citations," N.Y.L.J. (June 22, 2023), available at https://www.law.com/newyorklawjournal/2023/06/22/judge-imposes-5k-fine-on-lawyers-who-submitted-chatgpt-generated-fake-case-citations/.

management tools can be used to find related materials to review for context as needed (e.g., related emails, alternate drafts, etc.).

For Assessing the Completeness of Your ESI

If your top priority is assessing your collected ESI, finding individual documents and facts is less important than ensuring a sufficiently complete collection has taken place and that any filtering applied during processing has not been excessive. In such situations, your focus should be on tools and techniques that help you see the big picture of your ESI collection and reveal the gaps within it:

- Metadata filtering and visualization tools, which help you assess the completeness of your collection by revealing ranges of values and gaps in those ranges, as well as potentially revealing important date ranges and sources, the connections between custodians, and more
- Concept clustering, which can provide a valuable overview of the content types and topics within your materials, including revealing an absence of things you expected or the presence of things you don't need
- Visualization tools (e.g., communication maps, word clouds, etc.), which can reveal collection gaps, including missing date ranges, missing custodians, and more
- Thread and duplicate management tools, which can provide another way to map conversation threads to reveal gaps requiring further collection, or which can reveal the presence excessive near-duplicates suggesting a collection or processing issue

Formal random sampling can also be useful, particularly if there are disputes over the appropriate scope of preservation and collection that need to be resolved. Sampling to estimate prevalence can be used to apply relative value determinations to different sources and tranches and to estimate costs and benefits associated with specific proposed work.

For Pursuing Review Preparation

When your top priority is review planning and preparation, you are concerned with learning about what happened, but only insofar as that informs what must be reviewed later and how it should be prioritized. And, you are concerned with understanding the properties and the big picture of the ESI you've collected, but only insofar as that informs what tools and techniques for culling you should choose and what review methodologies are likely to be effective. All of the tools and techniques discussed so far can be leveraged to assist in this effort:

- Formal random sampling to estimate prevalence, which allows you to accurately estimate what you have, to evaluate the suitability of potential review workflows (including assessing the viability of a TAR or CAL solution or the need for additional objective culling), and to create a yardstick against which to measure future review work
- Formal random sampling to test classifiers, which allows you to iteratively improve any searches you plan to apply for culling, to ensure that they minimize unnecessary downstream review work and that they avoid missing any important materials
- Searching and metadata filtering, which can both be leveraged to eliminate as much of the chaff as possible without losing an unreasonable amount of the wheat, thereby reducing all downstream review and production costs
- Thread and duplicate management tools, which can dramatically speed up later review work, both by eliminating materials not requiring review and by providing superior organization to what remains
- Semantic indexing features, which can let you use concept clustering to help organize and prioritize subsequent review activity or let you leverage TAR workflows

KEY TAKEAWAYS

There are nine key takeaways from this practice guide to remember:

1. There are a variety of use cases in which analytic tools and techniques will be valuable, and a variety of goals you may pursue with them.

2. To pursue these goals, document review platforms include a range of useful features, including: random sampling tools; searching and filtering tools; structured analytics tools; conceptual analytics tools; technology-assisted review workflows; and potentially, PII tools, entity extraction tools, image analysis tools, and generative AI tools.

3. Sampling tools and techniques – particularly formal sampling – are good at revealing unknown unknowns, at giving an accurate overview of your materials, and at providing a reliable basis for improving searches and planning subsequent steps.

4. Search and filtering tools, including newer visualization tools like communication maps, are good at finding specific materials, at identifying gaps in your collection, and at eliminating irrelevant materials prior to review.

5. Structured analytics tools for managing threads and duplicates are good at placing documents in context, at prioritizing and organizing materials for review, and at avoiding duplicative review.

6. Conceptual analytics tools are good at exploring a collection of materials without foreknowledge of the contents and key terms and rapidly surfacing relevant materials.

7. Technology-assisted review workflows are good at prioritizing and organizing materials for review, reducing the volume of materials to be reviewed, and ensuring review quality.

8. Sometimes you may be able to leverage other new and developing analytic tools, such as PII tools, entity extraction tools, image analysis tools, and generative AI tools.

9. Successfully achieving your goals requires leveraging the right combination of these tools and techniques based on what you're trying to find out and how much you already know.

ABOUT THE AUTHOR

Prior to becoming an attorney, Xavier worked in the information technology industry for several years in database administration, telecom and VoIP implementation, and software development. As a senior director of Consilio's Data Analytics group, Xavier oversees projects that involve, Technology-Assisted Review, Immediate Case Assessments™, and analytics research.

Xavier Diokno has a bachelor's degree in computer science from Southern Illinois University, a master's degree in computer science from the University of Illinois at Chicago and a juris doctor degree from DePaul University College of Law. He is licensed to practice in the state of Illinois and the United States Patent and Trademark Office. Xavier oversees Consilio's Data Analytics team, where he advises clients on how to leverage technology in supporting their projects.

Xavier Diokno

Sr. Director Innovation Solutions

m +1.312.638.3130
e xdiokno@consilio.com

consilio.com

ABOUT THE AUTHOR

Matthew Verga is an attorney, consultant, and eDiscovery expert proficient at leveraging his legal experience, his technical knowledge, and his communication skills to make complex eDiscovery topics accessible to diverse audiences. A fifteen-year industry veteran, Matthew has worked across every phase of the EDRM and at every level, from the project trenches to enterprise program design. As Director of Education for Consilio, he leverages this background to produce engaging educational content to empower practitioners at all levels with knowledge they can use to improve their projects, their careers, and their organizations.

Matthew Verga

Director of Education
m +1.704.582.2192
e matthew.verga@consilio.com

consilio.com