LAW360®

Portfolio Media. Inc. | 230 Park Avenue, 7th Floor | New York, NY 10169 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

AI Will Soon Transform The E-Discovery Industrial Complex

By **Todd Itami** (February 6, 2025, 3:10 PM EST)

Todd Itami

The e-discovery industrial complex will be transformed by generative artificial intelligence sometime soon.[1] Our current paradigm of data handling will vanish and be replaced by fully integrated enterprise solutions.

This article examines the current state of the industry and the anticipated outcomes of this shift, while acknowledging the tough journey ahead.

Currently, the challenging intersection of legal and technical domains, combined with limited technical expertise among lawyers, has led to an e-discovery data market dominated by specialist vendors commanding premium rates for data services.[2] These vendors employ some of the most knowledgeable legal tech professionals in the world. This high skill bar contributes to high fees because of the significant complexity required for many e-discovery tasks.

But artificial intelligence is set to revolutionize this landscape, making processes easier, faster and more accurate. This transformation means less reliance on cumbersome workflows and more focus on what truly matters: delivering value to clients. There are many steps between where we are today and where we will end up.

Of course, every case requires fact-specific workflow design, so this article focuses on general ideas.

E-Discovery Gets Harder

Identifying, collecting and reviewing relevant information is hugely burdensome and technically challenging. Truly, books could be written on why the task is so hard and expensive.

The entire scheme of enterprise information technology was not created with e-discovery in mind. Enterprise information systems are designed to execute their primary functions and interact with other systems synergistic or adjacent to those primary functions — not to make lawyers happy.

Identifying exactly the right data for discovery, let alone exporting it in a forensically sound fashion, was not top of mind during enterprise development of the past two decades. Indeed, the very nature of certain very effective data structures does not lend itself to collection and review in any sort of targeted fashion.[3]

Some platforms are better than others, but for the most part, the platforms have focused on delivering value to clients outside of the e-discovery context.

Before generative AI, data was already growing exponentially in size, complexity and diversity. Consider the following points:

- Data growth rates in terms of absolute size are increasing year over year. Much has been written on this volume-based point.

- Data is becoming more complex: Productivity platforms and various types of content management tools have become increasingly cross-compatible. And it's more than hyperlink cross-pollination — most major platforms now provide plugins or modules for direct integration with other programs, including competing components.

- Data is more diverse: New data sources present new challenges for collecting data for e-discovery. Novel features create novel metadata and data structures that can be technically difficult to capture or reproduce in a way that allows for relevance review. For example, short messaging is constantly a pain point for review platforms as message complexity, encryption, ephemerality, and multimodal integrations present problems light years beyond viewing cool emojis. 😎

These data novelties create interesting second-order problems in e-discovery aside from simple extraction and review. For example, new collaborative software features and less department-based organization of people could blur the lines as to what counts as custodial data in a particular case. This could be problematic for jurisdictions that have default rules with a set numbers of custodians per side as a means of defining discovery boundaries.

In sum, e-discovery is getting technically and legally harder, not easier.

Generative AI will, at first, just throw gas on this e-discovery fire.[4] Speaking broadly, the current unprecedented wave of investment in generative AI will complicate the situation by:

- Creating even more data, because high-quality content can now be generated automatically;

- Creating novel forms of data, e.g., inputs, outputs, data structures, etc.; and

- Boosting the diversity of both AI and non-AI apps by significantly lowering the cost of, and barriers to, software development.

And here you thought it was getting crazy in 2024!

2025 E-Discovery: A Blunt Instrument

The legal industry grapples with the technical challenges of e-discovery through brute force. Even the most basic document type in corporate America — email — must usually be overcollected and processed into a discovery database to ensure that nothing potentially relevant is missed.[5]

And most other data sources follow this same workflow: very broad collection, filtering, processing and normalization into a database, and then culling, usually with elaborate sets of search terms, before attorneys lay eyes on document No. 1.

To give you an idea of scope, large firms usually have hundreds of cases with millions of files collected, tens of cases with tens of millions collected, and a few cases where collected files are counted in the hundreds of millions. I wish I were exaggerating.

And while every case is different, the armchair consensus in the industry is that less than 5% of collected data is ultimately produced to the opposing party or used in an internal investigation.

Initial broad culling — deduplication, filtration of system files and substantive search term narrowing — usually results in at least an 80% reduction in volume. After that, the documents are reviewed for relevance, which usually results in an additional 75% reduction in document count. But wait, it gets better.

Again, this is a gross generalization, but the punchline is that maybe 5% of what is produced — i.e., 0.25% of the gross — is actually generally relevant to the case, with only a fraction of that data being important to the outcome of the case.

Let me put this another way: If you told me that 1% of data collected for a major case was truly outcome-determinative, I would be more surprised than the time I found out that I had been unintentionally dating a particularly convincing AI for five months.

Not only is the system inefficient from a "data in, data out" perspective, it has also been monetized at every step of the process. Some vendors are better than others, but over the last decade, someone somewhere in the industry has managed to charge per-gigabyte, often monthly-recurring, fees for:

- Initial collection — device-level, enterprise-level, etc.;
- Processing — staging database-in;
- Hosting — staging database;
- More processing — staging data promoted to review;
- More hosting — review database;
- More processing — production imaging; and
- More processing — loading other parties' productions.

With some vendors, all of these data moves gets a charge. And we are ignoring both the per-document analytics fees for specialized tools, and the hourly rates for technicians that many vendors charge on top of these compute and hosting fees. Vendors even sometimes charge a final fee — in addition to hourly work — to shut the whole thing down.

To be clear, I don't think this billing schema is some sort of salesperson gambit. This is just the way that industry pricing evolved — in part because of the extremely low margin for error in each of these steps.[6]

Nevertheless, this pricing culture is well above the cost for equivalent secure, fast and reliable data services in other industries. And that fact makes the e-discovery industrial complex particularly vulnerable to outside forces, courtesy of the coming AI wave.

The AI Endgame

In my view, this current discovery data paradigm will be rendered mostly obsolete by generative AI.

The relics of old — data identification primarily through interviews, collecting extraordinary quantities of data, myriad technical collection difficulties, search terms, human document review workflows, pricing culture, etc. — will be displaced by outputs directly from the systems managing the data.

Sometime soon-ish, parties will agree on a written scope of discovery, give the scope to the enterprise systems holding the data, and the system will provide a set of documents that will require little, if any, attorney review.[7]

Not only will the platforms themselves identify and package the relevant documents, but the data delivery will be accompanied by automatically generated legal work product and analysis. That, my friends, is the endgame.

But this endgame will not be realized overnight — this journey from industrial complex to turnkey will happen in stages, outlined below.

1. Early Days

In the first stage, AI will simply be an efficiency enhancer, replacing human reviewers and drastically improving quality. This is happening right now.

2. Nascent New Value

In the second stage, review platforms will integrate AI in more sophisticated ways. Advanced integrators will go beyond relevancy and privilege document review, and will truly engage the variety of structured and unstructured outputs.

This will include the integration of human review tasks that technology-assisted review could never accomplish: logging, data extraction, reporting, sophisticated quality control, identification of recalcitrant personally identifiable information and automatic relation of disparate data sources, to name a few.

But fantastic new value will start to emerge that would be unimaginably cost-prohibitive in 2024. AI will be able to accomplish tasks that would have previously required thousands of hours of review by lawyers with comprehensive institutional knowledge and case background.

3. A New Era

In the end, generative AI will collapse the stages of the e-discovery life cycle. This is where the assembly-line view of the e-discovery cycle will no longer make sense.

Right now, it is unclear how this will play out, how much segmentation there will be in the industry and who the players will be. What is clear is that current state of play will become a mere memory.

Collateral Development

Large language model development will fuel other types of much-needed software development important to the legal industry. Some of the stickiest problems described above will be ameliorated, not because chatbots achieve sentience, but because other applications obviate the need for certain steps in the e-discovery workflow.

There are many examples here that I would love to discuss, but maybe the best one is content indexing.[8] For AI to efficiently "see" the data on your systems, it must be indexed. On the e-discovery side, we process huge amounts of data into staging databases precisely to solve the indexing and search problems.

As mentioned above, obscure file types; optical text recognition, or OCR; and data unitization issues create complications. But AI developers want to give the models access to this data.

When enterprise systems improve on-platform indexing — either through native capabilities or integration of indexing-specific applications — this will greatly benefit the legal tech community. Suddenly, data will be accessible and reliably searchable in-place using legacy technologies and AI alike.

This is just one example of how the rising tide of AI will float many legal tech boats (and bots).

And don't forget that the deep neural network breakthrough will produce its own set of non-LLM tools that will help with the discovery process. This huge potential should not be overlooked. For example, new handwriting analysis models have proven to significantly outperform traditional OCR programs in many use cases. It's not a chatbot, but it will be an impactful technology for lawyers in many cases.

Conclusion

The blunt instrument of 2025 discovery will soon become a laser scalpel. The industrial complex will still have its place.[9] But more importantly, the real winners will be the clients.[10]

Todd Itami is of counsel and director of artificial intelligence and e-discovery solutions at Covington & Burling LLP.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of their employer, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] The e-discovery industrial complex includes a set of services and software employed during large litigations and investigations. Legal professionals rely on these services to accomplish a primary task: production of large quantities of information responsive to discovery requests.

[2] In keeping with my commitment to a software-agnostic approach, any perceived references to specific products are not intended to endorse or highlight any specific solution, product, or company.

[3] A simplified example: many messaging apps, despite beautiful and intuitive user interfaces, store individual messages in combined tables on the user's device with every new message sent or received placed on a new row in a big table. You can imagine that it could be challenging to parse such tables at scale into discrete conversational threads for attorneys to review, all while preserving the esoteric metadata associated with each message. That is, assuming you have a way to pull the table out of the application in the first place!

[4] Side bar: It is wholly unhelpful that many have treated GenAI applications as an alien form of life requiring special rules for every aspect of use. I see this phenomenon as unnecessarily complicating the situation in a space where—spoiler alert—existing frameworks are perfectly capable of handling these new tools from governance, confidentiality, privacy, and information security perspectives. Don't get me wrong, proper handling of artificial intelligence is both hugely important and not easy; but we certainly don't need to completely reinvent the governance and acceptable-use wheels.

[5] In some cases, the burden of conducting a full-blown forensic collection may not be appropriate for the needs of the case.

[6] Hourly rates for professionals at vendors are also underpriced, in my opinion.

[7] I hope to unpack this in future articles. Streamlining this process will eventually upset the current balance of proportionality and cause us to reconsider the entire flow of civil discovery. The speed of the e-discovery investigation phase alone will greatly change the pace and focus of a litigation or investigation. I am conscious of other critical factors like attorney-client privilege review, but I am excluding them here to keep things simple.

[8] Content indexing reads the content of files and builds an "index" of the content. This index allows programs and other tools, like LLMs, to quickly and efficiently "see" the substance of the file to avoid repeatedly digging through each file, line by line, with every search or interaction. This indexing process can be as simple as a text-based inverted index that functions very similarly to a book index, or as complicated as converting letters, words, or phrases into numbers (embeddings/vectors). These processes are getting smarter in exciting ways that will help lawyers grapple with traditionally difficult data types like very long documents, very short documents, and documents with repeated or dirty data.

[9] In fact, I believe individuals currently on the inside of the complex are some of the best-equipped people to take us on the incredible journey that awaits.

[10] Clients will receive higher quality work for less money, value and legal analysis that was never before possible, and increased time and energy to focus on the most important and challenging aspects of legal disputes, rather than focusing on how to get the data from point A to point B.