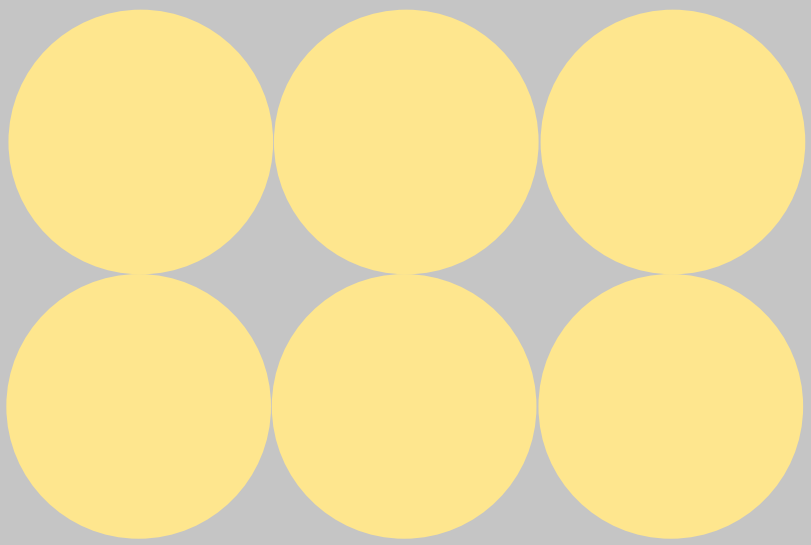
Legal Data Intelligence™

From Assessment to Intelligence

Operationalizing Early Data Strategy (EDS) in the Age of GenAI

By: Kevin Clark, Melina Efstathiou, Ross Gotler, Matt Hamilton, Tristan Jenkinson, Glenn Melcher, Daniel Miller, Linda Sheehan

July 2026



Introduction: How Did We Get Here?

In October 2025, LDI published Expanding From Early Case to Early Data Assessment: Leveraging Legal Data Intelligence for Dispute Avoidance and Strategic Response, arguing that law firms and legal departments are sitting on a gold mine of data that need only be organized and analyzed to provide bespoke, actionable, and strategic insights for their businesses. While the core concept, that Early Case Assessment (ECA) should lead to Early Data Assessment (EDA) may be compelling, as with everything when dealing with multiple sets of unstructured data, the devil is in the details. Over the intervening months, generative artificial intelligence (GenAI) has come roaring into our lives, making it accessible to all types of sophisticated data analytics that once required teams of analysts and engineers.

The pace of GenAI adoption is unprecedented. The American Bar Association's Legal Technology Survey Report reveals a rapidly accelerating adoption, with 30 percent of respondents reporting that they are using AI technology in 2024, up from just 11 percent in 2023. In 2026, the 8am Legal Industry Report indicated that 46 percent of law firms (and 58 percent of firms with 20 or more lawyers) have implemented general purpose AI tools. Adoption may be even more rapid overseas, with approximately 90 of the top 100 UK law firms having implemented some form of AI.

Now that some of the excitement around due diligence, onboarding, and configuration of AI tools has abated, organizations must figure out how to integrate them into their workflows in a way that can take full advantage of their capabilities. This enterprise, however, is not without its challenges.

GenAI has become pervasive so quickly that many law firms and legal departments have not yet had the opportunity to consider how best to integrate it into their data landscape and workflows, leaving their information governance regimes ill-equipped to feed effectively into analytical workflows. Early Data Strategy (EDS) provides a framework rooted in the application of Legal Data Intelligence to help legal practitioners navigate this novel challenge—responsibly harnessing the influx of available data to surface actionable insights, strengthen case strategy, and make better decisions, earlier.

Early Data Strategy

Early Case Assessment led us to Early Data Assessment, but now we can leverage our bespoke data into Early Data Strategy.

GenAI's Impact on Information Governance

Often attributed to mathematician Charles Babbage[i] and popularized by IBM programmer George Fuechsel in the 1950s, the principle "Garbage In, Garbage Out" is a strong fundamental concept regarding content used by AI. If you are using your own content as a knowledge base for AI (for example within an AI RAG system[ii]), it is important to ensure that the data set is curated using best practice information governance methodologies, so that the ingested content is suitable for its intended use.

Redundant, Obsolete, or Trivial (ROT) data could have a major impact on the results that AI may provide. For example, if there are copies of out-of-date information on a particular topic, this information could be relied upon and provided back to you in a response from the AI system. Ideally, this information should be removed from the knowledge base as part of those information governance procedures. It may make sense to retain this content within a separate knowledge base designed for prior content which can then be interrogated separately as and when needed.

This is far from the only reason why understanding your data is key when it comes to AI usage. Fundamentally, you need to be able to control what is fed into AI systems that you are using. The first steps to that control are:

- Understanding what data is stored
- Mapping out where it is stored
- Ensuring that those structures make sense from an information governance perspective

Once these are understood and mapped out, it is easier to identify and log which data sets are acceptable for different AI systems to access.

Potential risks of neglecting to curate a data set include:

- **Privilege contamination**: Where an AI tool could reveal privileged information to an individual or process who should not have access, or in such a manner as to waive the privilege.
- **Model training exposure**: Where protected intellectual property or other key information such as legal strategy documents, settlement ranges, and internal investigation files are ingested into a vendor's model, which then has some level of knowledge of their content.
- **GDPR and data protection**: GenAI tools processing personal data embedded in legal files without a lawful basis.
- **Contract and intellectual property violations**: Data subject to contractual use restrictions, confidentiality obligations, licensing limits, or intellectual property protections may be processed in a manner inconsistent with those obligations.
- **Data privacy compliance**: Inability to respond sufficiently to a data subject access request because of a lack of accurate information as to what data AI has processed.

Data mapping exercises can be a key part of information governance activities. One aspect of data mapping is logging which AI systems have access to which data sets. As part of this, it is also helpful to ensure that you understand and log key details of that AI system, including how and where the system processes data. Some AI systems use data from prompts and responses to improve the model itself. This could potentially put privileged or sensitive information at risk and should be taken into account when deciding which data sets, if any, that AI system is able to access.

Where data is processed is also an important consideration, especially for those processing data covered by the GDPR or other data privacy, secrecy, or localization laws and frameworks. Although the underlying data may reside in a particular jurisdiction, the relevant "processing" from an AI perspective could potentially take place in another jurisdiction. This is something to be very aware of as recently several AI systems have updated their approaches on where they process data. For example, Microsoft introduced "flex routing" for Copilot. For customers in the European Union, this allowed for "large language model (LLM) inferencing to occur outside the EU Data Boundary during periods of peak demand to help maintain a consistent Copilot experience" (where Inferencing is "the processing step when an AI model executes the prompt to produce an output or response").

Such actions could have severe implications under GDPR, for example, as it could be deemed that "processing" is occurring outside of allowed restrictions. Care should be taken to ensure that such changes in approach are also monitored as part of information governance procedures for AI tools.

AI and Machine Unlearning

Information governance is a continued effort, not a single process. In addition, the information governance systems may need to extend to AI-based content.

As discussed above, where your own data is being used as a knowledge base for an AI system, it needs to be kept up to date – not just when the AI system is set up, but on a continuous basis. If not, the same risks of ROT data being included in responses to user requests will occur.

This means that to some extent you may need those AI systems in use to "unlearn" content that they have been using.

For some systems (for example RAG-based systems) this may be more straightforward because the content of data in the knowledge base is not used to build the actual model itself. When removing data from a knowledge base, you may need to delete the content and then ensure that the search index is updated (typically the content of the document would be indexed and included for searching) and that any vector embeddings and cached content, including responses, summaries, or embeddings, are removed as deletion from a source may not cascade automatically.

For systems that have used the content as actual training material, this can make "unlearning" far more complex. Models fine-tuned on proprietary data, including email and internal documents, present a middle ground case that is more difficult to remediate, but does not require rebuilding the model from scratch.

Additionally, "unlearning" may be important to data privacy requirements. For example, under GDPR, data subjects can request that their data be deleted. If an AI system is trained on personally identifiable information, removal of such information may potentially be required in some circumstances. Further, while the regulatory framework is still catching up, the EU AI Act; the UK's Information Commissioner's Office (ICO) updated guidance on AI and data protection (2024); and the National Institute of Standards and Technology (NIST) AI Risk Management Framework all touch upon unlearning and erasure obligations.

AI Vendor Contracting as a Control Point

Vendor negotiations are a critical control point for how organizations operationalize Early Data Strategy and manage AI risk at scale. As AI-enabled platforms become embedded across the enterprise, the commercial terms governing data use, model interaction, and confidentiality directly shape both value realization and risk exposure.

A clear understanding of your data and intended AI use cases translates to appropriate data-related protections being put in place as you move towards Early Data Strategy. This is particularly important where legal privilege is at stake, as maintaining confidentiality remains a prerequisite across jurisdictions. However, different AI tools treat client data in different ways and even across models within the same platform.

One of the most important, but inconsistently treated, clauses in AI contracting is whether client data may be used to train, fine-tune, or otherwise improve the vendor's models. With the plethora of AI tooling options, organizations must navigate a complex mix of licensing structures, enterprise carve-outs, product-specific commitments, and opt-in or opt-out mechanisms, across an evolving tech stack.

This challenge is further compounded by potential gaps in exposure across the tech stack. Existing vendor contracts may not even address unlearning upon request, and legacy agreements signed before the explosion of generative AI may not contain provisions addressing model training. While contracts with new vendors may trigger detailed review of all terms, including those relating to AI, renewal processes for established platforms may not receive the same level of contractual scrutiny. Organizations should ensure that all contracts, for both new and existing vendors, follow a process that ensures detailed review.

GenAI Governance is a Live Document

The release of Anthropic's Fable 5 model in June 2026 demonstrates how data protection can vary dynamically depending on model selection, often with limited transparency. Privileged documents could become inconsistently protected depending on what model is used, as auditing which model was used becomes harder. Anthropic released its latest Fable 5 model without an enterprise zero data retention policy alongside automatic re-routing to earlier models. In this same example, the US government export control directive resulted in a global suspension days after release of Fable 5, reportedly due to the vendor's inability to reliably segment access by user nationality. This underscores the heightened risk of AI governance resulting in vendor practices that change rapidly in response to external pressures, including regulatory intervention.

In a GenAI environment, it is important to understand what data exists, where it sits, which systems can reach it, under what permissions, for what purposes, and on what contractual terms. Without that consolidated view, legal teams cannot answer basic but critical questions with the confidence required in disputes and investigations. Data scoping is no longer confined to an email archive and a project folder. EDS doesn't neatly scope, extract, and ingest data into a contained environment. Depending on the details of the specific matter, the scope may now account for collaboration platform versioning, the storage and retention of prompts, iterations, and AI outputs, whether client data has been collected by or disclosed to third parties through AI integrations, what data a model was training on at any given time, and which model version was running at each relevant point. The legal question extends beyond whether a single document or prompt retains privilege. It now extends to whether the model itself and the data flows that feed it are compliant with established

requirements, such as privacy and privilege rules, and fast-evolving AI governance guidelines. A model may be trained on data that your organization does not own, does not control, or has no rights to use for that purpose (such as personal data and third-party intellectual property). This complexity multiplies where a firm deploys a third-party AI model or extends AI platform access to clients through collaborative portals.

To prepare for EDS, you need to know what data flows through which integration points, and for what purposes. Data maps that track both internal data and the contractual and technical boundaries governing shared environments can help answer questions relating to what terms each AI vendor processes client data, and whether the terms are complying with your obligations and risk appetite. Opposing counsel and courts are increasingly asking pointed questions about vendor contracting and data flows. Regulators are doing the same. Legal Data Intelligence practitioners, leveraging their legal backgrounds and technology fluency, are uniquely positioned to bridge the legal, technical, and contractual dimensions of EDS that no single function can manage alone.

AI System and Decision Governance

Then there is governance of the AI systems themselves. This means knowing which tools are approved for which activities, maintaining an inventory of what is in use, and mapping each tool's data access against your data map. Without this visibility, legal teams cannot meaningfully assess or manage AI-related risk. You cannot answer the regulator's question, the court's question, or the client's question about what was done with and by which system.

Equally important is how AI outputs inform human decisions. Such decision governance requires that material AI-assisted outputs receive appropriate human review and that the decision-making process is documented, with certain jurisdictions looking to implement certifications for the use of AI in drafting court papers. However, not all AI use cases carry the same risk, as seen in the EU AI Act's risk-based approach, and controls should reflect that approach.

Connecting Governance Layers

Information governance failures create AI system governance risks. AI system governance failures create decision governance failures. The absence of any one layer means the other two cannot function reliably. The data map is the common foundation. It connects the data to the system, the system to the output, and the output to the decision. Our preceding described information governance as "a strategic enabler that transforms how legal data is leveraged to anticipate risk and drive smarter decisions." AI governance is now the necessary evolution of that enabler. Without it, the strategic value that information governance was designed to deliver cannot be realized.

War Stories: Costly Landmines and Lessons Learned

The following scenarios are drawn from practice. Details have been anonymized. Each story illustrates the same underlying dynamic: the organizations that struggled did not fail because of bad lawyers or inadequate technology. They failed because they did not know their data before they needed it.

The Inquiry That Started Three Weeks Late

A midsize financial services and investment firm received a formal SEC inquiry. The underlying conduct was not the problem. The response was.

Legal and IT had never aligned on where custodian data lived across the firm's environment. There was no data map, no preservation protocol, and no documented inventory of which systems held what. When the inquiry landed, the first three weeks were consumed not by collection or review, but by the threshold question of who the relevant custodians were and where their data resided. By the time the firm had answers to those questions, it was already behind its response deadline.

The penalty the firm ultimately paid was not driven by the substance of the SEC's findings. It was driven by the delay. The regulator's position was straightforward: an organization of this size and sophistication, operating in a regulated industry, should have known where its data was. The absence of a data map was not treated as an oversight. It was treated as an indicator of broader governance failure.

Contrast this with the EDA-ready organization. Within 48 hours, the company suspends deletion, preserves likely relevant material, and engages with government from a position of cooperation and remediation planning. Legal can confirm timely preservation, collection, and disclosure of relevant documents, extending to a clear list of custodians and data locations and proactive identification of overseas documents, while providing defensible reasoning for certain data sources restricted due to data privacy and blocking statutes, demonstrating genuine cooperation with the authorities. The underlying conduct is addressed promptly providing evidence of swift corporate action. The difference between those outcomes can turn on whether legal was able to say yes to that first question, "Has your client preserved the documents, records, and communications relevant to the conduct described?" That readiness cannot be built after the knock on the door.

2 Million Documents, Three Relevant Custodians

A corporate client under federal investigation had no scoping protocol and no data map. When outside counsel issued preservation and collection instructions, the internal team did what organizations without a framework tend to do: they collected everything. Data was pulled from more than 40 custodians and loaded into a review platform. 2 million documents entered the pipeline.

The matter ultimately turned on three custodians. The relevant document population, once properly scoped, was under 50,000. The client paid for the difference, including hosting, processing, and review costs associated with nearly 2 million documents that should never have been collected. The proportionality argument that Federal Rules of Civil Procedure 26(b)(1) was designed to support was available to them in theory. Without a data map and a scoping protocol, they had no practical mechanism to invoke it.

This is not an unusual outcome. It is a predictable one when organizations approach federal investigations without the data infrastructure to scope their response. The cost is rarely just financial. Time spent managing an over-collected dataset is time not spent on strategy.

Containing the Investigation Before it Became the Crisis

A board of directors retained outside counsel and engaged an LDI team to support an internal investigation into a senior executive. The allegations involved potential criminal conduct and, at a minimum, serious leadership failures and unprofessional behavior affecting multiple parts of the

organization. There were several potential targets, and it was not initially clear how far the conduct extended or who else might be implicated.

The team was engaged to work directly with IT to move quickly and quietly. The objective was to get into the data, understand what had happened, identify the individuals involved, and do so before the situation could spread further. Working from a structured collection and analysis approach, the team moved through the relevant data rapidly, isolated the primary actor's conduct, and identified the colleagues who had been drawn in, most of them inadvertently, without awareness of the full picture.

Within a short timeframe, the board had what it needed: a clear account of what had occurred, who was responsible, and who was collateral. They were able to act decisively, remove the relevant individuals, and protect the organization from a position of documented, defensible evidence rather than rumor and allegation. The investigation did not leak. The organization did not lose control of the narrative.

That outcome was only possible because the data work was done right and done fast. The tools and discipline that make EDS work in a regulatory inquiry or litigation context are the same ones that make an internal investigation defensible, contained, and conclusive.

The Under-Collection Trap

Self-collection of data can cause problems down the line for a variety of reasons. Even with data from Microsoft 365, one of the most common data sources, mistakes can be easily made on the collection and preservation of data. With no oversight, logging, or QA procedures, such mistakes may not be picked up until too late.

In a case in the UK Patents Court, an internal team at a manufacturing company missed around 800,000 documents in their initial collection (roughly 40 percent of the data that should have been collected). They were ordered to pay £578,444 to the opposing party as a result, and the trial had to be delayed by two years.

Sadly, even a second collection by the internal team was deemed deficient, and a third collection was ordered to be performed in conjunction with an independent vendor.

EDS: The Next Frontier

Early Data Strategy is not a product. No vendor sells it, and no platform delivers it upon deployment. It is a capability, and like most legal capabilities of consequence, it is built over time through disciplined practice, organizational commitment, and a willingness to treat data as a strategic asset rather than a litigation byproduct.

What EDS Looks Like in Practice

One practical marker of mature EDS would be the capability of the General Counsel to query the organization's legal portfolio in natural language and receive a reliable answer. Not a dashboard prepared by an analyst. Not a report commissioned for a specific purpose. A direct, real-time response: the three highest-frequency claim types in the last 24 months, the cost drivers in each, the outside counsel performance metrics across the portfolio, the matters that share a risk profile with an emerging dispute.

That capability requires GenAI. It also requires everything GenAI needs to function reliably in a legal context: governed data, accurate taxonomy, consistent intake, maintained access controls, and a clear ownership structure. GenAI applied to an ungoverned data estate does not produce the GC's natural language query tool. It produces a sophisticated mechanism for surfacing unreliable information with apparent authority.

GenAI's Role in EDS

Within a governed EDS framework, GenAI contributes three distinct capabilities that were not practically available before.

1. **Pattern detection at scale**: identifying claim profiles, settlement drivers, and litigation risk factors across matter portfolios too large for human review.
2. **Anomaly flagging**: automatically surfacing matters or data patterns that deviate from established baselines and warrant early attention.
3. **Natural language querying**: allowing legal and business stakeholders to interrogate the legal data estate without requiring technical intermediaries, making the intelligence genuinely accessible to the people who need to act on it.

The Maturity Roadmap

Most organizations are realistically 12 to 24 months from being in a position to run EDS safely. That is not a discouraging assessment; it is an accurate one, and it argues for starting now. The roadmap runs from data map completion through taxonomy governance, consistent intake practice, platform selection, and GenAI deployment against a governed environment. Each stage has value independently. The organization that completes stage one has already materially reduced its litigation response time and improved its ability to meet-and-confer obligations, even before a single AI tool is deployed.

Implementing and Leveraging the Benefits of EDS

As the preceding paper made clear, companies and law firms are sitting on a gold mine of relevant, bespoke data, they just need to know how to leverage it for strategic guidance. ECA gave us a snapshot, EDA gave us a roadmap, and now EDS gives us a compass.

Organizations that use GenAI most effectively will not necessarily be the ones that implemented it first, but rather those that thoughtfully governed their data first. To be clear, information and data governance requires effort and diligence, which may take months or years to implement, but the benefits will be more than worth it.

What concrete steps can a company or firm take right away to get them started along this path?

- **Audit your data map:** Does it exist? Is it current? Does legal own it and know what GenAI tools can see?
- **Review your GenAI vendor contracts:** Check for training data clauses, data residency terms, and access governance provisions.

- **Assign an LDI owner:** If no one owns Legal Data Intelligence in your organization, that gap is now a GenAI governance risk.

---

[i] Charles Babbage is also known as the person who first had the notion of the Analytical Engine, the world's first conceptual general purpose mechanical computer.
[ii] IBM defines RAG as "an architecture for optimizing the performance of an artificial intelligence (AI) model by connecting it with external knowledge bases. RAG helps large language models (LLMs) deliver more relevant responses at a higher quality."