Building the Future of Law: A Legal Data Infrastructure

Executive Summary

Law firms frequently contend with extensive, unstructured data—including discovery materials, email communications, internal notes, transcripts, and more. This creates what we refer to as “Fact Chaos.” In this state, essential information is scattered across multiple sources and labelled inconsistently, undermining efficiency and increasing the risk of overlooking critical details. Recent findings underscore the severity of this issue: retrieval-augmented generation (RAG) systems—a common method in legal AI—misstate facts in 17–33% of complex legal responses (Magesh et al., 2024). Meanwhile, expanding an LLM’s context window to 100K or even 1M tokens has demonstrated diminishing returns, as it compels models to sift through volumes of extraneous text (Li et al., 2024).

This whitepaper introduces the **Legal Fact Layer** as a practical way to address these issues. Instead of repeatedly searching through lengthy files, a Legal Fact Layer extracts and centralises key facts—parties, events, clauses, obligations—into one continuously updated repository. Crucially, each fact is captured and stored with contextual information about its source, reliability, and legal relevance, clearly distinguishing between mere allegations or claims and verified statements of truth. This method lowers the risk of missed information, enforces consistent data labelling, and leads to more dependable AI outputs. Small discrepancies can have serious consequences in legal work, making high data accuracy essential.

Key topics include:

- Why “Fact Chaos” persists and how it affects both human and AI performance.
- Limitations of document-centric methods and why RAG alone can overlook vital facts.
- How a Legal Fact Layer functions—from fact extraction and tagging to real-time verification.
- Anticipated benefits, such as faster matter comprehension, fewer inaccuracies, more reliable AI drafting, and greater client confidence.

In a field where near-100% accuracy is critical, relying on scattered repositories is increasingly untenable. By exploring this paper, readers will see how a shift to fact-

focused management—grounded firmly in contextual relevance and clarity about factual provenance—can reshape daily legal processes, establish more dependable AI applications, and offer measurable advantages for clients.

1. Introduction: The Legal Industry's Transformation

1.1 The Rise of AI and the Document-Centric Legacy

Law firms today generate large volumes of data of all kinds, from an ever-growing number of sources. Despite the availability of powerful AI technologies such as Large Language Models (LLMs), this data remains scattered in multiple repositories, hindering visibility and therefore usefulness. Inconsistent labelling, "handmade" folder structures, and frequent duplication compound the difficulty of quickly extracting and verifying even the simplest detail.

Many legal technology providers rely on retrieval-augmented generation (RAG) to address these challenges. While RAG undoubtedly improves LLM performance in the legal domain versus foundation models, real-world evaluations show that even advanced RAG systems often overlook crucial facts (Vals AI, 2025). When subtle or decisive data continues to slip through the cracks, lawyers remain understandably sceptical.

This reliance on document-centric methods exacts a high price:

- **Repetitive Searches**: To confirm a critical date or clause, lawyers must revisit and re-read lengthy PDFs.
- **Inaccuracies and Oversight**: When facts are buried or mislabelled, the risk of missing key information rises sharply.
- **High Onboarding Costs**: New team members must sift through the same unstructured documents repeatedly, compounding inefficiency.

1.2 Why "Fact Chaos" Inhibits AI

AI systems—especially LLMs—thrive on structured, consistent data. When forced to parse random text fragments or outdated document versions, they often produce incomplete or incorrect outputs. AI is awfully good at producing plausible but false information, or contradictory statements that are detached from other parts of its work, that requires painstaking manual correction.

Without a well-organised data foundation, advanced AI tools cannot reliably scale. Contradictory sources prompt the model to guess details, references to the same entity that differ in name can fragment into multiple versions, and lawyers end up re-checking results manually—cancelling out any time saved in the process (Dahl et al., 2024).

1.3 Moving to a Fact-Centric Approach

A Legal Fact Layer turns the standard model on its head. Instead of details being permanently buried in static documents, it extracts and organises the discrete facts that truly matter—parties, timelines, obligations, communications—proactively, into a dynamic repository that:

- Captures each fact in a standardised format.
- Verifies the data against source documents.
- Updates automatically whenever new information, versions, or filings emerge.
- Makes these facts available for queries, analysis, or AI-driven drafting.

By shifting to a proactive fact-centric system, law firms can achieve consistent, real-time alignment across legal teams and practice areas. The same repository that supports day-to-day legal work also informs AI-driven insights, leading to iterative improvements in accuracy and efficiency.

To appreciate why a fact-centric system is transformative, we must understand how legal facts differ from everyday facts. Traditional business data may remain static across contexts, but in law, a single clause or date can gain or lose significance depending on the stage of litigation, applicable legal standards, or how the narrative of the dispute evolves over time. This dynamic nature of legal facts reveals the limitations of purely document-based workflows and underscores the need for a specialised solution—one that can seamlessly extract and tag critical parties, events, and obligations. By doing so, it keeps pace with shifting legal contexts, ensuring no material fact slips through the cracks.

2. Why Legal Facts Differ from Everyday Facts

Legal facts are not simply statements or claims. They carry specific contextual nuances and depend heavily on the circumstances in which they're presented. The same fact can be irrelevant or highly significant, truthful or simply alleged, depending entirely on its origin and how it was introduced in a legal context.

For example, consider the following statement:

“On 26 June, Rowan stole a computer from the workroom.”

Taken at face value, a non-legal system (or standard AI tool) creating a chronology might mistakenly record this as a definitive fact. If asked, such a system may answer unequivocally, "Rowan stole the computer."

However, in a legal matter, this statement's significance changes dramatically based on context:

- **Who said it?** Was it stated by an eyewitness, an aggrieved colleague, or someone with ulterior motives?
- **Where was it said?** Was it presented under oath in court, recorded formally in an affidavit, or casually mentioned in an unverified text message?
- **Why was it said?** Is it part of sworn testimony (a **statement of truth**) or merely an allegation made during preliminary stages of an investigation (a **claim**)?

These distinctions directly impact the legal weight and admissibility of the statement. A claim without clear contextual grounding cannot be taken as a legally established fact. A legitimate **statement of truth**, supported by credible evidence and made under appropriate circumstances, however, is legally compelling and central to proving liability or innocence.

In short: Legal facts are fundamentally different from everyday facts because they are always embedded within context. It's never enough to record the existence of a statement; you must understand and verify the context, provenance, and credibility surrounding it before considering it a legally relevant fact.

How a **Legal Fact Layer** Helps: A **Legal Fact Layer** tags and tracks each fact (including clauses or dates) in real time, flagging which one's matter for each case. Lawyers and AI systems can instantly see contested details, withdrawn statements, or newly added evidence—all tied back to the original source. This ensures no critical fact is overlooked or misapplied when circumstances shift.

3. The Challenge: Unstructured Data and “Fact Chaos”

3.1 Fragmented Data and Delayed Decisions

Law firms often store matter data in multiple, disconnected locations. Practice management tool for matter overviews and sometimes servers for PDFs and Word documents as well, email systems for correspondence, and so forth. As a result:

- Checking a single date or name can require wading through hundreds of pages.
- Minor but crucial facts can be buried or mislabelled.
- Each new participant—whether an associate, partner, or co-counsel—must “rediscover” the same facts and context from scratch.

Although retrieval-augmented generation (RAG) improves upon purely generative approaches by fetching relevant text snippets, it still relies on the assumption that the right passages are found, while crucial data may never appear in the AI’s final output (Vals AI, 2025). Lawyers remain understandably sceptical when subtle or decisive data continues to slip through the cracks.

3.1.1 A Day in the Life with the Legal Fact Layer

Consider a scenario where you are preparing for litigation involving thousands of pages of exhibits and revised details. With a fact-centric platform such as Mary Technology’s Fact Management System (FMS), your daily workflow transforms:

- **Instant Fact Retrieval**: Enter a query (e.g., “Which witness statements mention the incident date that changed from 15 March to 20 March?”) and immediately see only the relevant references—rather than scrolling through endless PDFs.
- **Automated Contextual Linking**: Each document is automatically connected to related timelines, communications, or conflicting details, helping you spot what really matters.
- **Real-Time Updates**: When new evidence emerges or a file version changes, the system flags all affected facts, preventing reliance on outdated data.
- **Seamless Collaboration**: Colleagues can annotate or confirm facts in a shared repository, ensuring everyone—across offices or practice groups—accesses the same current record.

This simplified approach frees valuable time for strategy and client interaction, while minimising the risk of missed details or contradictory information. By easing the burden

of continuous fact-finding, a Legal Fact Layer places lawyers back in control, allowing for more proactive, data-driven decision-making.

3.2 Why Document-Centric Systems Fall Short

The legal industry, quite sensibly, has long centred its operations around documents. Document Management Systems (DMS) defined the initial transition from paper-based workflows to digital repositories, upending how law firms store, index, and secure large volumes of files. Practice Management Systems (PMS) later expanded these capabilities, handling billing, time recording, and other operational tasks crucial to running a firm—yet these platforms never needed to systematically manage “facts.” Until recently, the legal field did not rely on advanced AI tools that demand granular, context-aware data.

Most established systems continue to focus on entire documents instead of extracting and maintaining discrete factual information. Because these documents often lack comprehensive metadata, practitioners must re-read the same pages repeatedly to verify a single point. Any changes to one file may not propagate to related materials, and as repositories swell, so does the volume of unindexed or untagged content—limiting deeper insight generation and compounding costs. Even PMS solutions that advertise AI capabilities (RAG) remain chiefly optimised for document storage and security rather than real-time fact management. Feeding a large context to a language model can easily miss subtle updates or discrepancies if the underlying data is not systematically structured (Vals AI, 2025).

A Fact Management System (FMS) addresses this gap by shifting from document storage to fact-level oversight. Instead of relying on static PDFs or Word files, an FMS captures each key fact—whether it is a clause, date, communication, or event—tags it appropriately, and continuously updates it when new evidence appears. This approach ensures advanced AI solutions work from a verified, coherent factual baseline, minimising duplicated reviews, preventing overlooked evidence, and eliminating the version confusion common in purely document-centric workflows.

4. The Data-First Imperative

4.1 Lessons from Knowledge Graphs

Knowledge graphs—used by major technology companies—are an effective way to organise information about the world. They map entities (people, places, products) and

the relationships between them (works at, purchased from, located in). This structure makes searching and linking vast amounts of data straightforward, providing clear relationships that can be navigated with ease. Essentially, the knowledge graph acts as a **retrieval booster and consistency check**, preventing the LLM from ignoring related information.

Adopting a similar approach in the legal domain means identifying and storing not just entire documents, but the key facts inside them:

- **Parties** (who is involved?)
- **Events** (what happened, and when?)
- **Clauses** (what obligations or requirements exist?)
- **Communications** (who said what, and to whom?)

A knowledge-graph-inspired framework therefore enables lawyers and AI tools to quickly traverse these facts and relationships directly—rather than hunting for them in text-based documents. Recent legal AI experiments confirm that combining RAG with a structured knowledge layer reduces hallucinations and improves multi-hop reasoning (Barron et al., 2024; Nguyen & Satoh, 2024). However, these early hybrids remain far from flawless. Layering a prototype knowledge graph on top of RAG does not fully eliminate missed facts or ambiguous references—but it does illustrate how deeper, structure-driven methods can significantly advance AI reliability.

5. The Vision: A Legal Fact Layer

5.1 From Document Dumps to a Single Source of Truth

A Legal Fact Layer is not just another database or file storage system. It is a central repository that:

- **Extracts Key Entities**: Identifies parties, contracts, obligations, events, and communications—similar to a knowledge graph identifying nodes and edges.
- **Maintains a Clear Schema**: Consistently labels and organises facts so that practitioners, support staff, and AI all reference data the same way.
- **Tracks Versions and Disputes**: Marks facts as contested or updated when new evidence appears, creating a real-time reflection of the matter.
- **Feeds AI Tools with Verified Data**: Large Language Models or other AI solutions draw from curated facts rather than rummaging through unstructured text.

Significantly, a well-built **Legal Fact Layer** greatly enhances the construction of accurate timelines and comprehensive matter understanding. Studies show that unstructured LLM summaries can misorder events or, more concerningly, skip subtle but decisive facts (Wang et al., 2024). By capturing each detail with a verified source and context, a **Legal Fact Layer** can present an accurate, unified picture—ensuring nothing is over-summarised or omitted in the process.

When a lawyer queries, “Which transcripts or witness statements discussed the incident date originally recorded as 15 March but later corrected to 20 March—and who referenced the incorrect date?”, a traditional system might only capture some references or overlook the change entirely. By contrast, a **Legal Fact Layer** unifies every mention of both dates, highlighting precisely when the error was introduced, who repeated it, and how it was resolved—rather than forcing a time-consuming search through scattered PDFs and email threads.

This more direct approach ensures no update is missed, enabling practitioners to react swiftly to shifting facts. Instead of wondering whether every reference to an outdated date has been discovered, lawyers can verify changes in real time. Ultimately, the **Legal Fact Layer** lets teams focus on more useful tasks rather than perpetually chasing version histories.

5.2 Fact Management Systems (FMS) as the Building Block

Mary Technology is introducing to the world what is known as a Fact Management System. You will find more comprehensive explanations below, however implementing a Fact Management System (FMS) is a practical first step toward a Legal Fact Layer. An FMS:

- **Automates Ingestion**: Splits PDFs, extracts relevant text, de-duplicates files, and identifies unique facts (like nodes in a knowledge graph).
- **Structures and Tags**: Applies uniform labelling conventions (dates, names, events) to facilitate advanced searches.
- **Facilitates Continuous Verification**: Teams can review extracted data, correct mistakes, and confirm or dispute facts in real time.
- **Dynamic Chronologies:** In seconds, you automatically have a comprehensive timeline ready for client advisories, negotiations, or trial prep—saving up to 60% of your usual chronology time.
- **Interactive Dashboards:** Mary’s dashboards highlight missing information, track case progress in real time, and keep everyone aligned with a single source of truth.

Verified facts can be reused across multiple queries and issues within legal matters, saving lawyers from repeating the same data extraction for every new query. Ultimately, the **Legal Fact Layer** is the base that supports both lawyers and new AI features—like automated summaries, strategic insights, and analysis—without drifting into error-prone territory.

6. How the Legal Fact Layer Empowers AI

6.1 High-Quality Inputs for Large Language Models

Without a Fact Layer, LLMs often produce contradictory or superficial results, may "hallucinate" plausible yet incorrect details, or conflate multiple versions of the same document or party. By contrast, an AI tool integrated with a Legal Fact Layer can reference deterministic, validated facts, reducing guesswork and inconsistencies. The model "knows" exactly which timeline, parties, or provisions to focus on, yielding:

- **Repeatable, Deterministic Outputs**: The same query against the same data set produces the same result every time.
- **Reduced Confusion Among Entities**: Clear tags distinguish parties or obligations that might otherwise appear identical.
- **Greater Transparency**: Lawyers can readily trace an AI-generated statement to its underlying source fact.

The Vals AI Benchmark Report (2025) found that even the top-grade legal technologies employing RAG missed at least three critical facts from a set of just ten documents. In the realm of law—where trust is paramount—this can be the difference between adopting AI at scale versus discarding it as too risky.

6.2 Beyond Generic Document Processing

Many 'AI for law' solutions rely on simple text ingestion and keyword search, hoping to surface relevant information. Some suggest that simply expanding LLM context windows will address the shortfalls of this current technical methodology. In practice, longer context alone does not guarantee better organisation; it can amplify contradictions if data is not properly structured (Li et al., 2024).

Since legal matters evolve, continuous data ingestion and version control are essential. Relying solely on re-indexing large text blocks remains inefficient and prone to missing

critical updates (Hoffmann et al., 2022). A Fact Layer, however, akin to a specialised legal knowledge graph, makes more complex reasoning possible:

- **Targeted Summaries**: Summarise only the data relevant to a specific query, preserving crucial detail.
- **Logic-Driven Drafting**: Draft documents that precisely reference verified, consistently labelled facts.
- **Enhanced Reliability**: Errors within the training corpus become far less detrimental when the system has a structured "source of truth" to check against.

Recent prototypes combine RAG with a knowledge-graph-like approach, letting AI systematically traverse known relationships instead of guessing (Barron et al., 2024). Nguyen & Satoh (2024) exemplify this with their Knowledge Representation Augmented Generation (KRAG) framework, which uses inference graphs to help LLMs reason more effectively about domain-specific content—particularly in complex legal contexts.

By guiding models to incorporate key entities, relationships, and structured argumentation, KRAG ensures outputs are more consistent and better explained. This hybrid strategy yields more consistent legal answers, especially when references span multiple documents or conflicting versions.

7. Contrarian Perspectives: Challenging the Status Quo

7.1 "Document-Centric Workflows Are Good Enough"

Many legal professionals feel that storing PDFs in searchable folders suffices—after all, that is how legal work has "always" been done. Yet field tests of LLM-based summarisation highlight 10–20% error rates in reassembling key facts (Vals AI, 2025). For high-stakes matters, these omissions or misorderings create unacceptable legal risk. A single, central Fact Layer drastically cuts the time spent retrieving and verifying information.

7.2 "We Trust Our Lawyers, Not Automation"

Human expertise remains indispensable in law. But as matters expand, even the most diligent professional can overlook or mislabel crucial details. A structured repository

does not replace practitioners; it augments them by systematically managing routine tasks.

7.3 “LLMs Alone Will Fix Data Problems”

Generative AI is powerful but still depends on data quality. No matter how advanced, an LLM cannot reliably reconcile thousands of conflicting or duplicative references on its own. Garbage in, garbage out, as they say.

In law, hoping that AI is good enough to deal with your data problems is not an acceptable attitude. A Fact Layer grounds AI tools in verifiable data, reducing random guesswork.

7.4 “We Already Have a Practice Management Tool”

Practice management platforms are the centrepiece of law firms, and we only expect their impact on law firms to grow as AI enables them to offer a wider variety of tools to help lawyers in their day-to-day work. A Fact Management System is very much a compliment to a Practice Management System that makes its impact to law firms great, not a replacement in any sense. PMS’, however, do not typically provide a fact-centric architecture that systematically identifies, updates, and centralises all key matter details. A Fact Management System (and indeed Legal Fact Layer) complements (and elevates) existing systems by ensuring the underlying data is organised and accessible for sophisticated AI usage.

8. Conclusion

Modern AI tools, especially Large Language Models, offer remarkable speed and sophistication in legal drafting, research, and analysis. Yet without structured, accurate data, these technologies have demonstrated so far that they cannot reliably scale, even as the foundation models improve at a stunning rate of progression in recent years. The current state—reliance on fragmented repositories and repeated manual reviews—stifles efficiency and fosters risk.

A Legal Fact Layer provides the antidote to “Fact Chaos.” By capturing, verifying, and continuously updating crucial matter details in one authoritative repository, law firms:

- Provide AI with clear, unambiguous information, significantly reducing hallucinations and errors.

- Improve efficiency through quick fact retrieval and consistent, data-driven insights.
- Strengthen client Trust by delivering transparent, verifiable outputs and demonstrating streamlined workflows.
- **Meet evolving client demands** in an era where accountability, speed, and cost savings are increasingly important.

Just as knowledge graphs found their place in the way large technology companies organise complex information, a Legal Fact Layer can fundamentally elevate legal practice management, client service, and AI-driven innovation.

Even as AI models evolve, experts caution that bigger context windows or purely text-based methods fail to eliminate the need for a stable, well-maintained fact repository (Li et al., 2024; Hoffmann et al., 2022). Embracing a fact-centric model ensures that today's law firms—facing ever-larger datasets, rising client expectations, and rapid AI evolution—remain well-positioned in the data-driven evolution of legal practice (Wang et al., 2024).

9. What's Next?

Whilst this whitepaper presents a broader thought-leadership perspective, law firms or practitioners interested in seeing this future today can speak to us about what we are building. Our Fact Management System (FMS) merges natural language processing pipelines, knowledge-graph methodologies, and secure, scalable infrastructure to bring the Legal Fact Layer concept into everyday legal work.

For more information on implementing a structured "source of truth" to reduce errors, manage costs, and expedite decision-making, please visit www.marytechnology.com **or contact us at** dan@marytechnology.com.

By **adopting a fact-centric approach**, law firms can mitigate the risks of "Fact Chaos," better serve their clients, and **remain at the forefront of modern legal practice.**

Citations

Below are the key sources referenced or consulted in this whitepaper, drawn from the research materials:

1. **Barron et al., "Bridging Legal Knowledge and AI: RAG with Vector Stores & Knowledge Graphs," ICAIL 2024**
https://arxiv.org/html/2502.20364v1#:~:text=receive%201%2C%20and%20responses%20entirely,and%20analyzing%20the%20decomposed%20hierarchical
2. Li, Z., Li, C., Zhang, M., Mei, Q., & Bendersky, M. (2024). *Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach*. (2024).
https://aclanthology.org/2024.emnlp-industry.66.pdf
3. **Wang et al., "Legal Evaluations and Challenges of LLMs," arXiv 2024**
https://arxiv.org/html/2411.10137v1#:~:text=Similarly%2C%20%5B34%5D%20assessed%20GPT,case%20summarization%20and%20legislative%20interpretation
4. **Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E.,** "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models," arXiv 2024
https://arxiv.org/abs/2401.01301
5. Vals AI Benchmarking Report (Feb 2025)
https://www.vals.ai/vlair
6. **Nguyen & Satoh, "KRAG: Knowledge Representation Augmented Generation," arXiv 2024**
https://arxiv.org/html/2410.07551v1#:~:text=This%20paper%20introduces%20Knowledge%20Representation,framework%20or%20in%20tandem%20with
7. Magesh et al., "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," 2024
https://arxiv.org/pdf/2405.20362
8. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de las, Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., & Sifre, L. (2022). *Training Compute-Optimal Large Language Models*. https://arxiv.org/abs/2203.15556.