Study #2 — Normalization Under Structural Entropy

File-Level Media Normalization at Scale with Adaptive Geocoding Caching

1. Executive Summary

What began as a normalization study on managed photo libraries evolved into a much harder operational benchmark: large-scale normalization under structural entropy. Study #2 isolates the most operationally difficult segment of the archive — 247,130 heterogeneous files distributed across 8,861 folders, deep directory trees, fragmented historical layers, recovered media, exports, duplicates, metadata-light segments, and large video workloads.

Unlike managed photo libraries, folder-based archives do not preserve structural coherence naturally. They accumulate operational disorder over time through device migrations, backup replication, exports, recovery workflows, and years of uncontrolled filesystem growth. This benchmark evaluates how deterministic file-level normalization behaves when archive structure collapses.

Study #2 highlights five operational findings.

- **Structural entropy dominates cost —** Processing behavior varied more with workload composition than with dataset size itself. NoGPS density, duplicate movement, directory locality, file size, and video concentration became the dominant operational variables.
- **Normalization remained operationally stable —** Despite heterogeneous workload conditions and multi-day execution, global operating cost converged toward a stable regime of approximately 4.33 sec/file.
- **Duplicate handling became a primary I/O workload —** Folders introduced physical duplicate movement instead of logical duplicate suppression. Across the benchmark, 85,630 duplicated files were physically handled, transforming duplication into a real storage and I/O cost component.
- **Metadata availability reshaped throughput —** Metadata-light segments containing large NoGPS concentrations consistently reduced unit cost, revealing that geolocation resolution is not the dominant bottleneck under fragmented workloads.
- **Execution state influenced performance materially —** Long-running IO-bound execution proved sensitive to operating system interaction state. Screensaver and idle-state conditions introduced measurable throughput degradation even under controlled power settings.

The strongest result of Study #2 is not raw throughput. It is deterministic convergence under operational stress. Even when archive structure becomes fragmented, heterogeneous, and operationally difficult, normalization behavior remains measurable, predictable, and structurally consistent over time.

Study #1 demonstrated that structure improves normalization efficiency. Study #2 demonstrates that normalization remains viable even when structure deteriorates.

Messy archives are where normalization matters most.

The files remain.

Normalize first. Catalog later

2. Managed Libraries vs Unmanaged Archives

Study #1 focused on managed photo libraries — structured environments where media already exists inside curated catalog systems. Study #2 moves into a fundamentally different operational reality: unmanaged archives accumulated directly at filesystem level.

This distinction matters because libraries and folders do not behave the same way structurally. Managed libraries naturally preserve locality, metadata continuity, and organizational consistency. Folder-based archives accumulate entropy progressively through years of migrations, exports, backups, device replacement cycles, recovery workflows, and uncontrolled duplication.

Managed libraries and unmanaged archives operate under different structural conditions.

Managed Libraries

- High metadata consistency
- Strong geolocation continuity
- Natural chronological locality
- Logical duplicate handling
- Predictable directory structure
- Cache-friendly workload behavior

Unmanaged Archives

- Fragmented directory trees
- Mixed provenance and historical layers
- Recovered and metadata-degraded media
- Physical duplicate propagation
- Large NoGPS concentrations
- Randomized I/O access patterns

The operational difference between these two environments became immediately visible during normalization. Managed libraries behaved predominantly as cache-dominant workloads, where accumulated location reuse and structural locality progressively reduced operational cost. Folders behaved as heterogeneous operational regimes where throughput constantly shifted according to metadata availability, file size distribution, duplicate movement, and filesystem locality.

This benchmark demonstrates that archive quality is not binary. Archives exist across a spectrum ranging from highly structured catalog systems to fragmented filesystem accumulations. Normalization must remain operationally stable across that entire spectrum.

Libraries revealed the advantages of structure.
Folders revealed the operational cost of losing it.

That distinction defines the transition from Study #1 to Study #2. The goal is no longer simply to demonstrate normalization efficiency under organized conditions, but to understand how deterministic normalization behaves under structural entropy.

3. Structural Entropy at Scale

Large folder-based archives do not become structurally complex all at once. They accumulate entropy gradually through years of perfectly normal use. Every migration, backup, export, synchronization workflow, recovery process, and device replacement cycle introduces new structural layers into the filesystem.

Unlike managed libraries, unmanaged archives preserve historical accumulation directly at directory level. The filesystem becomes a visible record of operational history: duplicated folders, nested exports, recovered media, temporary copies, fragmented chronology, inconsistent naming conventions, and metadata degradation.

Study #2 exposed several forms of structural entropy simultaneously.

- **Deep directory fragmentation —** The benchmark processed 247,130 files distributed across 8,861 folders, creating highly fragmented traversal patterns and unstable filesystem locality.
- **Mixed provenance accumulation —** The archive combined media originating from smartphones, cameras, exports, backups, recovered datasets, migrated libraries, and derivative workflows accumulated across 25 years.
- **Metadata degradation —** Large segments contained incomplete or missing geolocation metadata, producing NoGPS-heavy workloads and reducing semantic continuity across files.
- **Duplicate propagation —** Unlike managed libraries, folders preserved physical duplication directly in the filesystem. Duplicate handling became a real operational workload involving storage movement and additional I/O.
- **Recovered and derived media —** Recovery workflows introduced metadata-light files, thumbnails, exports, and structurally inconsistent assets with highly irregular naming and temporal continuity.

One of the most important observations from Study #2 is that structural entropy does not simply increase complexity — it changes the operational behavior of normalization itself. As workload composition shifted, the system transitioned between distinct processing regimes influenced by locality, metadata density, duplicate movement, file size, and media composition.

This produced periods of highly stable throughput followed by abrupt operational transitions, particularly when the execution moved between homogeneous segments and heterogeneous historical layers. The result was not random instability, but structured operational variability.

Structural entropy changes workload behavior more than workload volume.

That distinction became one of the defining findings of Study #2. The archive did not behave like a uniform dataset. It behaved like a sequence of operational regimes shaped by historical accumulation.

4. Operational Conditions

Study #2 was not executed inside a controlled laboratory environment or synthetic benchmarking framework. The benchmark was performed under real-world operational conditions using consumer hardware, long-running workloads, and a historically accumulated archive with heterogeneous structural characteristics.

This distinction is important because normalization behavior at scale is shaped not only by algorithms, but also by storage topology, filesystem locality, workload continuity, operating system interaction state, and long-duration execution dynamics.

Execution Environment

- Machine — MacBook Pro M2 Pro / 32 GB unified memory
- Source — encrypted WD My Passport HDD (4 TB)
- Destination — encrypted Samsung Portable SSD T7 Shield (4 TB)
- Execution model — fully local processing
- Workload duration — multi-day continuous execution

Unlike the managed photo library benchmark from Study #1, folder normalization introduced a significantly more demanding operational profile. The workload combined random HDD reads, SSD writes, duplicate movement, metadata extraction, large video handling, and deep directory traversal simultaneously.

Operational Stress Factors

- **HDD read locality —** Highly fragmented directory structures forced unstable read locality and seek-heavy access patterns, particularly during globally ordered traversal.
- **SSD write amplification —** Folders normalization physically moved organized and duplicated files, transforming normalization into a sustained storage write workload.
- **Large video segments —** Video-heavy directories produced measurable increases in unit processing cost due to file size and sustained I/O pressure.
- **NoGPS-heavy segments —** Metadata-light workloads reduced geolocation overhead while simultaneously exposing the lower operational bound of the pipeline.
- **Long-running execution state —** Multi-day execution exposed behavioral variations associated with operating system interaction state, idle conditions, and screensaver activity.

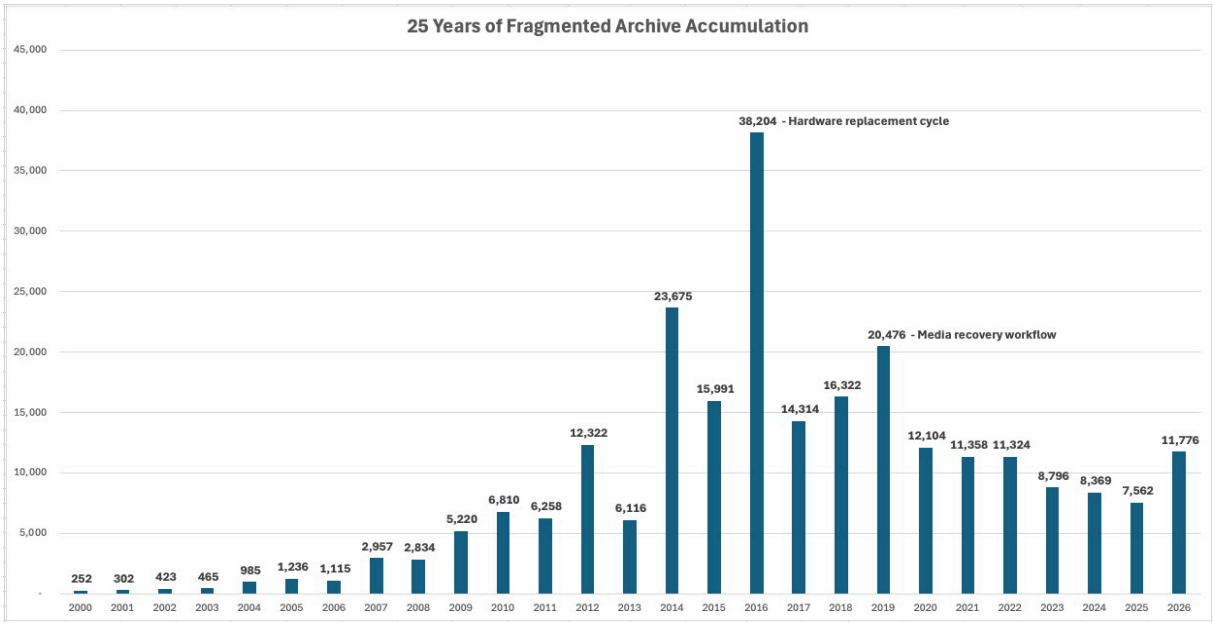

Figure 1 — Historical accumulation of heterogeneous media across 25 years of unmanaged archive growth.

One of the most important operational findings of Study #2 is that workload composition consistently dominated throughput behavior more than dataset size itself. Periods containing large videos, duplicate-heavy directories, or fragmented locality produced higher operational cost than equally large but homogeneous metadata-light segments.

This benchmark therefore revealed that normalization cost at scale is not defined by a single baseline throughput. Instead, the system continuously transitions between operational regimes shaped by archive composition.

Operational behavior emerged from the interaction between archive structure, workload composition, and execution conditions.

Metric	Value
Files Processed	247,130
Folders Traversed	8,861
Multi-Day Runtime	392h 50m
Converged Global Cost	~4.33 sec/file
NoGPS Files	123,169
Duplicated Files	85,630
Videos Processed	31,044
Total Data Volume	~1.88 TB
Errors	8

Figure 2 — Consolidated operational metrics across 247,130 files and 8,861 folders.

5. Operational Regimes

One of the central findings of Study #2 is that large-scale normalization does not operate under a single continuous cost model. As archive composition changed throughout execution, the system repeatedly transitioned between distinct operational regimes with stable local behavior but different throughput characteristics.

These transitions were not random fluctuations. They emerged from structural differences inside the archive itself — metadata density, filesystem locality, duplicate concentration, file size distribution, media composition, and execution-state conditions.

Operational behavior clustered into recognizable regimes.

- **Cache-dominant regime —** Segments with strong location continuity and accumulated reuse produced highly stable throughput and reduced external dependency.
- **NoGPS-heavy regime —** Metadata-light workloads containing large concentrations of files without geolocation consistently reduced unit processing cost by minimizing location resolution overhead.
- **Video-heavy regime —** Directories dominated by large video files produced sustained increases in operational cost due to prolonged I/O activity and larger storage movement.
- **Duplicate-heavy regime —** Folders containing large duplicate concentrations transformed normalization into a physical movement workload involving sustained write amplification.
- **Seek-bound regime —** Highly fragmented directory traversal reduced filesystem locality and forced unstable HDD access patterns, particularly during globally ordered execution.

- **Execution-state regime —** Long-running IO-bound workloads exhibited measurable throughput degradation under sustained idle-state and screensaver conditions, despite stable archive composition.
- **Mixed regime —** Most long-running segments operated under mixed conditions where workload composition continuously shifted between metadata density, file size, locality, duplication patterns, and execution-state transitions.

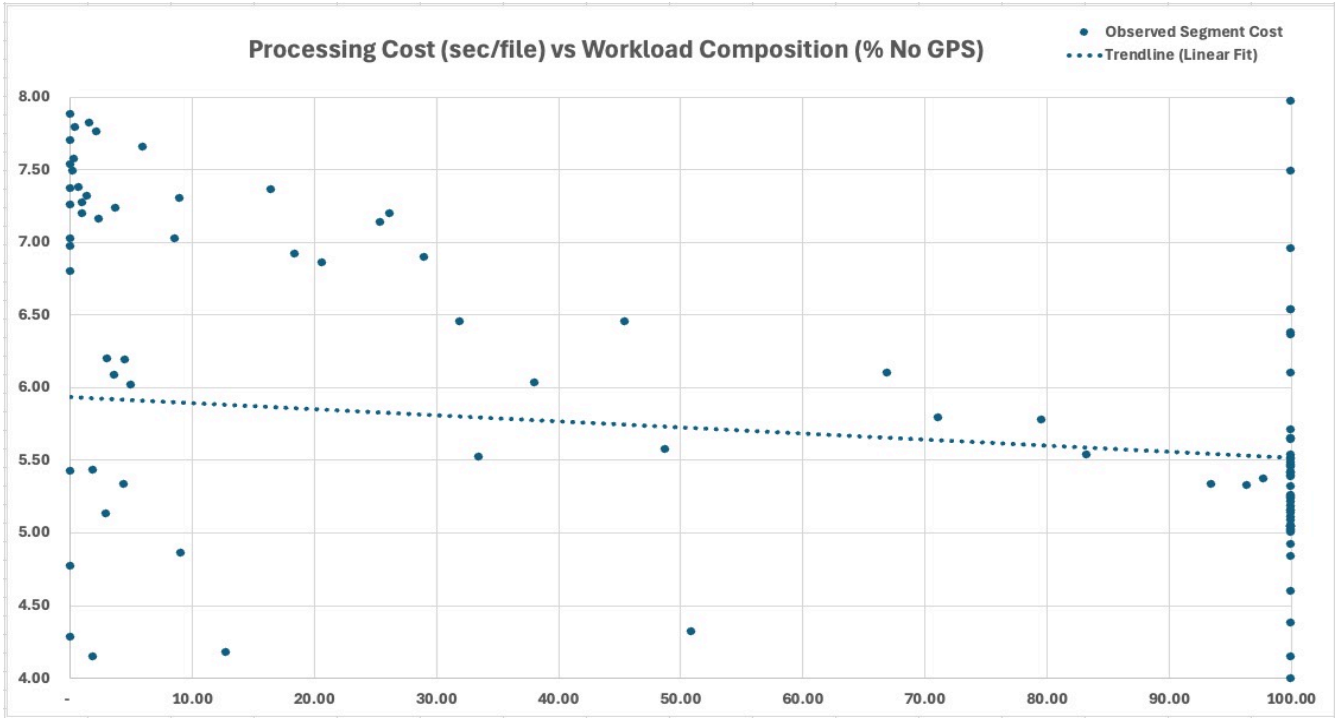

Figure 3 — Relationship between workload composition and observed operational cost across heterogeneous processing segments.

One particularly important observation emerged from NoGPS-heavy workloads. Contrary to intuitive expectations, metadata-light segments often produced lower processing cost than metadata-rich segments because they avoided repeated geolocation resolution and reduced operational branching.

This revealed that normalization cost was not primarily driven by dataset size, but by workload composition. Two equally large segments could exhibit radically different operational behavior depending on metadata availability, duplicate concentration, filesystem locality, and file size distribution.

Performance variability emerged from regime transitions rather than random operational noise.

Execution-State Regime

One of the most unexpected findings of Study #2 was that operational throughput depended not only on archive composition, but also on the interaction state of the operating system itself.

During long-running IO-bound execution, segments processed under prolonged screensaver and idle-state conditions consistently exhibited significantly higher unit cost, despite identical workload composition and controlled power-management settings.

In several observed segments, throughput degradation approached 50% under sustained idle-state execution. The workload remained stable. The archive composition remained stable. Only the execution state changed.

This revealed that operational variability could originate not only from archive structure, but also from OS-level scheduling behavior associated with user inactivity.

Execution state became an operational variable.

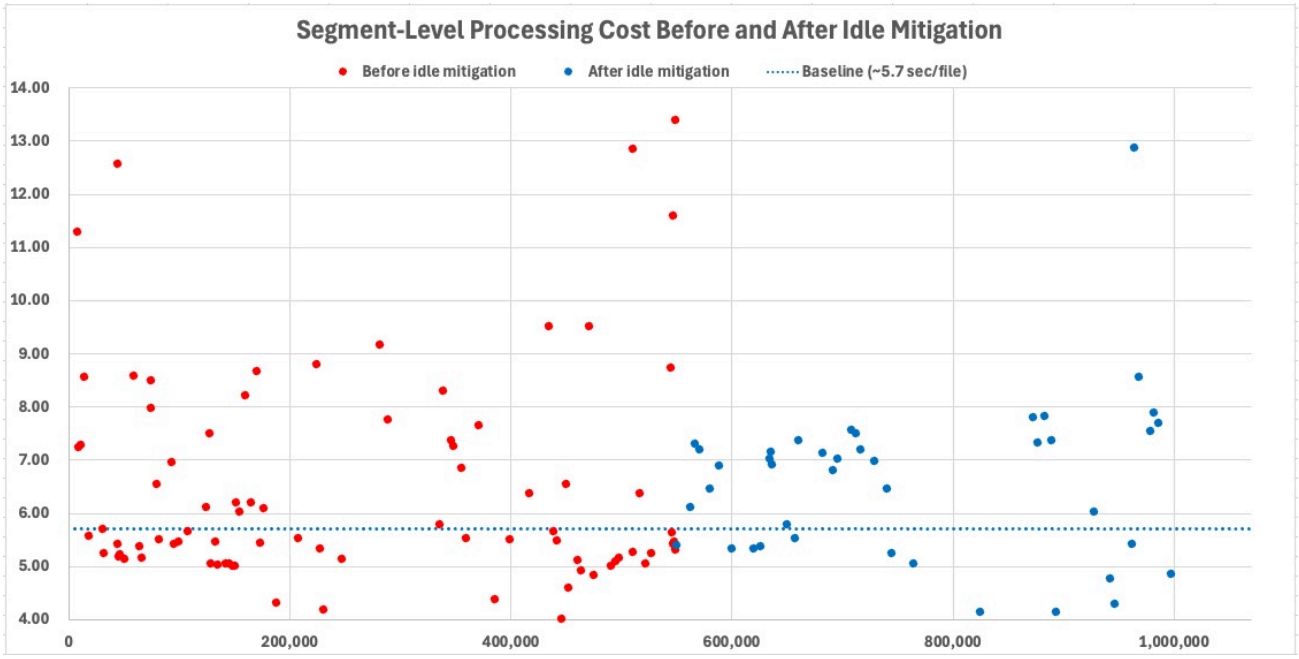

Figure 4 — Segment-level operational clustering before and after idle-state mitigation.

Another critical finding was that local operational variability did not prevent global predictability. Individual segments could oscillate between low-cost and high-cost regimes, yet the cumulative operating behavior progressively converged toward a stable global throughput profile.

The archive therefore behaved less like a monolithic dataset and more like a sequence of historically accumulated operational states.

Despite extreme local variability, the cumulative operating profile remained structurally stable over time.

Local variability. Global predictability.

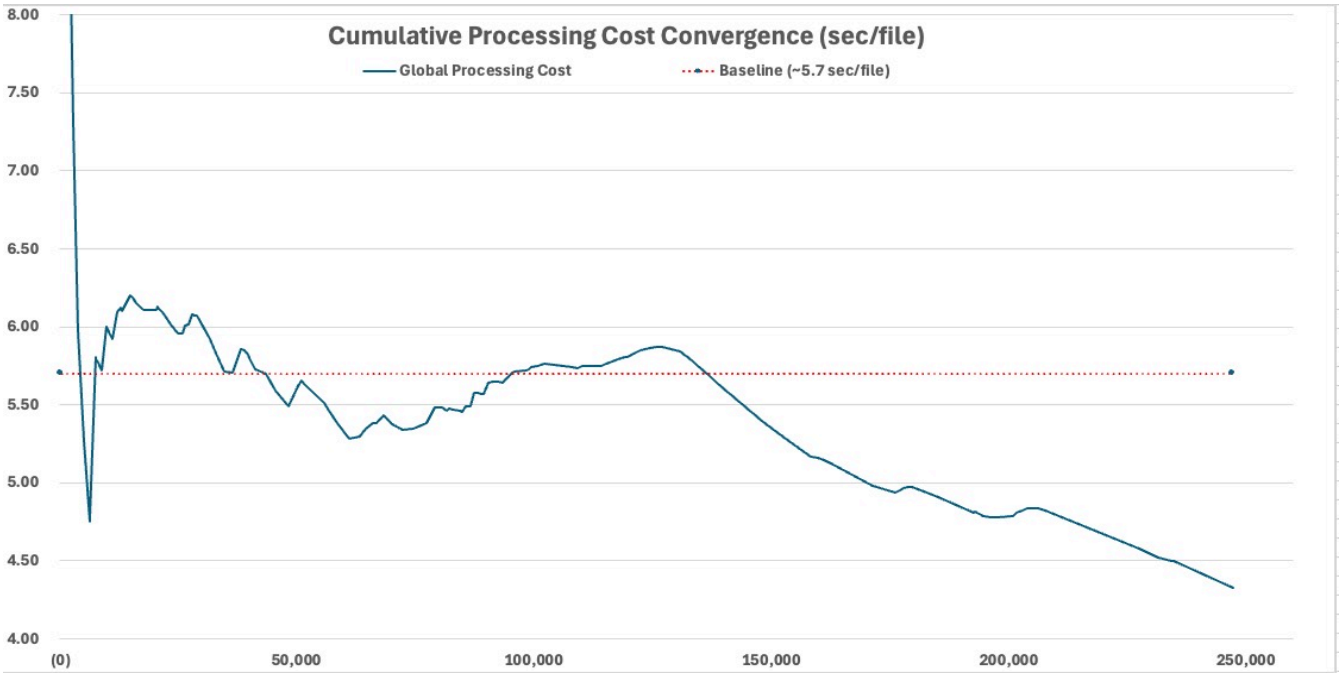

Figure 5 — Global processing cost convergence across 247,130 heterogeneous files processed under shifting operational regimes.

6. Unexpected Operational Findings

Study #2 revealed several operational effects that only became visible under long-running, heterogeneous, real-world execution. These findings did not emerge from normalization logic alone, but from the interaction between workload composition, filesystem structure, storage behavior, and execution state.

Several unexpected operational behaviors emerged at scale.

- **Execution-state sensitivity —** Idle-state and screensaver conditions produced measurable throughput degradation during sustained IO-bound execution, despite stable workload composition.
- **Directory locality restoration —** Highly fragmented traversal periods frequently recovered throughput abruptly once execution entered more homogeneous directory segments.
- **NoGPS lower-bound exposure —** Metadata-light segments consistently exposed the lower operational cost boundary of the pipeline by minimizing location-resolution work.
- **Video-driven amplification —** Large video concentrations significantly increased operational cost through prolonged read/write activity and storage pressure.
- **Duplicate movement as physical workload —** Folders transformed duplicate handling into measurable storage movement and sustained write amplification.
- **Stable global convergence —** Despite severe local variability, cumulative operating behavior progressively converged toward a stable throughput profile.

One of the most important implications of these findings is that normalization at scale cannot be fully understood through synthetic throughput benchmarks alone. Real-world archives accumulate filesystem entropy, metadata degradation, duplicate propagation, recovery-origin media, and execution-state variability simultaneously.

Study #2 demonstrated that operational behavior emerged from the interaction between data, storage, and execution conditions rather than from any single workload characteristic in isolation.

Operational variability remained explainable even under structural entropy.

Rather than invalidating predictability, these findings made the benchmark more operationally meaningful. The archive did not behave randomly. It behaved as a sequence of observable operational states.

Normalization remained operationally intelligible even under fragmented conditions.

7. What This Means

Study #1 demonstrated that structure improves normalization efficiency. Study #2 demonstrated something more important: normalization remains operationally viable even when structure deteriorates.

This distinction matters because most real-world archives are not curated systems. They are historically accumulated operational environments shaped by migrations, exports, backups, device replacement cycles, recovery workflows, and years of unmanaged filesystem growth.

The benchmark revealed several broader implications.

- **Normalization is fundamentally operational —** At scale, normalization behaves less like a metadata utility and more like a long-running operational system interacting continuously with storage, filesystem structure, and workload composition.
- **Archive quality exists on a spectrum —** Managed libraries and fragmented folders are not separate categories. They represent different points along a continuum of structural entropy.
- **Workload composition matters more than raw volume —** Operational cost was shaped primarily by metadata density, duplicate propagation, locality, media composition, and execution state rather than by file count alone.
- **Determinism survives structural entropy —** Even under fragmented workloads, heterogeneous regimes, and multi-day execution, normalization behavior remained measurable, explainable, and globally predictable.
- **Normalization becomes more valuable as archives degrade —** The more fragmented and historically accumulated the archive became, the more important deterministic normalization proved to be.

One of the most significant conclusions from Study #2 is that structural entropy does not eliminate operational structure. Instead, entropy reveals hidden operational regimes that remain observable through throughput behavior, locality shifts, metadata density, and execution-state transitions.

This changes how normalization should be understood. Normalization is not simply a preparation step before cataloging. It is the process that restores operational intelligibility to historically fragmented archives.

Messy archives are where normalization matters most.

The files remained deterministic even when the archive itself became operationally chaotic. That distinction became the defining result of Study #2.

Managed libraries revealed the advantages of structure. Folders revealed the operational cost of losing it.

The files remain.
Normalize first. Catalog later.

8. What Remains

Study #1 revealed how structure improves normalization behavior. Study #2 revealed what happens when that structure deteriorates through years of operational accumulation.

The benchmark demonstrated that fragmented archives are not exceptional cases. They are the natural outcome of long-term digital history: migrations, backups, exports, recovery workflows, device replacement cycles, duplicated storage layers, and decades of unmanaged filesystem growth.

What appears operationally chaotic at archive level still contains deterministic structure at file level. That distinction became the central insight of Study #2.

Normalization did not eliminate entropy. It made entropy operationally intelligible.

Throughout the benchmark, workload composition continuously shifted between metadata-rich and metadata-light segments, video-heavy and duplicate-heavy regimes, fragmented traversal

patterns, recovery-origin media, and execution-state transitions. Yet despite that variability, cumulative behavior remained explainable, measurable, and globally predictable.

The system did not hide operational complexity. It exposed it.

That exposure matters because normalization is ultimately about restoring structural trust in long-lived archives. Not by removing historical accumulation, but by making it observable, deterministic, and operationally manageable.

Structural entropy is inevitable.
Operational opacity is not.

Large archives will continue evolving. Platforms will change. Storage systems will change. New layers of accumulation will emerge over time.

But the underlying problem remains the same: archives become operationally harder to trust when structural drift remains invisible.

Study #2 demonstrated that deterministic file-level normalization remains viable even under fragmented, heterogeneous, and operationally difficult conditions.

The files remain.
Normalize first. Catalog later.

9. From Benchmark to Operational Model

Study #2 began as a benchmark of fragmented folders, but its implications extend beyond throughput analysis. The benchmark revealed that large-scale normalization behaves as an operational system shaped by workload regimes, filesystem structure, metadata continuity, and execution-state conditions.

What initially appeared as archive disorder gradually exposed a deeper operational pattern: fragmented archives still preserve deterministic structure at file level, even when global organization collapses.

Study #2 revealed several operational principles.

- **Deterministic file identity —** Each file preserves intrinsic structural truth independent of catalog state or historical fragmentation.
- **Operational locality matters —** Filesystem continuity, directory structure, and execution order materially influence throughput behavior under large-scale workloads.
- **Metadata quality shapes operational cost —** Metadata-rich and metadata-light segments produce fundamentally different throughput regimes.
- **Accumulated knowledge reduces dependency —** Location reuse progressively transforms normalization into increasingly knowledge-driven execution.
- **Structural visibility restores trust —** Normalization exposes duplication, metadata degradation, unresolved media, and operational complexity directly at file level.

One of the most important outcomes of Study #2 is that operational predictability survived structural entropy. Even under fragmented workloads and heterogeneous regimes, normalization behavior remained explainable and globally stable over time.

This suggests that normalization should not be understood merely as a cleanup utility applied after archive degradation occurs. It functions as an operational layer that preserves structural intelligibility as archives evolve.

Normalization is not only about organization.
It is about preserving operational trust in long-lived archives.

Together, Study #1 and Study #2 establish a broader conclusion: managed libraries reveal the advantages of structure, while fragmented folders reveal the operational consequences of losing it.

The files remain.
Normalize first. Catalog later.