

The AI Data Foundation Blueprint

Every AI initiative on the roadmap reads from this foundation.
None of them work without it — and none of it requires a data lake.

PREPARED FOR

Meridian & Frost, PLLC
Dana Whitfield · Marcus Bell

DATE

May 28, 2026
v1.0 — approved scope

PREPARED BY

Brad Taylor
Fractional Chief AI Officer



The honest framing

"AI-ready data" is where enterprise vendors make their money and mid-size firms lose theirs. Before this blueprint commits you to anything, here is the version of this project you should refuse — and the version you actually need.

⊗ What the enterprise pitch sells

A centralized data lake, an 18-month program, seven figures, and a data-engineering team the firm doesn't have and shouldn't hire.

At 118 people this fails before it finishes: the migration outlives everyone's attention span, and the lake becomes a second K: drive with a better logo.

☑ What a 118-person firm needs

Governed document libraries in the SharePoint you already pay for. A metadata catalog: owner, tier, version, retention. Tier tagging enforced at ingestion. Disciplined ownership.

Cost: mostly internal time and a modest Azure bill. The expensive part is discipline, not infrastructure.

6

SOURCE SYSTEMS

Inventory, page 3

~1.9M

FILES ON THE K: DRIVE

630 GB · since ~2007

~40%

ESTIMATED
DUPLICATES

From the March
assessment

3

MIGRATION WAVES

Most files retire, page 5

The blueprint's one job

Make the knowledge layer, the intake automation, and every future workflow read from **one governed place** — instead of six systems and a 19-year-old shared drive that nobody owns.

In this blueprint

- 3 **Source inventory** — the six systems, with an honest AI-readiness verdict on each
- 4 **Target architecture** — governed libraries, the metadata catalog, PII and retention
- 5 **The K: drive migration** — three waves; most files retire, not migrate
- 6 **Phases, owners, cost** — who does what, when, and what it unblocks

Source inventory — six systems, one verdict each

From the March assessment's workflow and data map. Every system keeps its day job; the question is only how its data reaches the governed layer — and whether it should.

SYSTEM	WHAT LIVES THERE	TIER	AI-READINESS VERDICT	INTEGRATION
CCH Axcess	Tax returns, workpapers, e-file records	RED	Ready within limits. Stays the system of record; only extracts move.	API sync, read-only
QuickBooks Online	~140 CAAS client books	RED	Ready. Structured, well-owned; feeds the Q3 close workflows.	API sync
Karbon	Jobs, tasks, client correspondence	YELLOW	Ready. The intake automation's connector (due May 30) does the work once.	API + webhook
SafeSend	Delivery and e-sign records	RED	Partial. Vendor AI features under review; exports only until the data-handling terms are settled.	Read-only export
M365 / SharePoint	Email, Teams, current working documents	1-3	Destination, not just source — the curated libraries live here.	Native destination
The K: drive	~1.9M files, 630 GB, everything since ~2007	UNTAGGED	Not ready, and never will be as-is. Retire it — never index it.	Migrate-and-retire

The K: drive gets the honest verdict

Indexing 1.9M ungoverned files doesn't create knowledge — it creates liability with a search box. The K: drive's usable content moves through the three waves on page 5; the rest retires under the retention schedule. No AI system reads it in the meantime.

Integration patterns, defined

API sync	Scheduled read-only pull into structured extracts. The source stays the system of record; nothing writes back.
Webhook	The source pushes changes as they happen — used where freshness matters, like Karbon job status for the intake automation.
Read-only export	Periodic file export, no live connection. The pattern for vendors whose AI and data terms aren't settled yet.
Migrate-and-retire	One-time move of the files worth keeping, then the source goes read-only and is decommissioned. Reserved for the K: drive.



Target architecture



Source systems

CCH Access · QuickBooks Online · Karbon · SafeSend · M365 · K: drive (until retired). Each keeps its job; nothing gets ripped out.



Sync & extract

API sync for CCH and QBO, webhook for Karbon, read-only exports for SafeSend, a one-time three-wave migration for the K: drive.



Governed storage

Curated SharePoint document libraries (owned, versioned) for documents; structured extracts in Azure storage. These two places are the only ones AI systems ever read from.



Metadata catalog

Owner · tier · version · retention on every entry. Tier tagging happens here, at ingestion — a document without an owner and a tier doesn't enter. No exceptions, no backfill promises.



Consumers

The knowledge layer (backlog #3), the intake automation (#2), CAAS close workflows (#4), and future evaluation runs — all reading from the same governed layer, never from sources directly.



PII handling patterns

Tier 3 data stays in its system of record wherever possible. Extracts that must carry client identity are masked at extraction (client name → client ID) and live only in access-controlled Azure containers. Tax-return information (§7216) never enters a Tier 1 or Tier 2 store — the catalog's tier field makes that a checkable rule, not a hope.



Retention schedule sketch

Returns & workpapers	7 years, per firm policy and peer review
Client correspondence	3 years, then dispose
Templates & policies	Current version + supersession chain, indefinitely
Everything else	Retention field set at ingestion; disposal becomes a query, not an archaeology project



The K: drive migration — three waves

Nineteen years of everything, taken in three passes. The order is deliberate: files people need this week, then files the AI roadmap needs this quarter, then the archive. Most files never migrate — they retire.

Wave 1 · Active-client files — June–July

Files for clients with open engagements: dedupe, apply the naming convention, assign an owner, land in the client libraries. If nobody claims a file, it doesn't move — it waits for wave 3's retention check.

Wave 2 · Templates & precedents — underway now

This wave is the knowledge-layer corpus cleanup — one effort, not two. Nicole Tran's team is working through the ~290 duplicate tax templates the assessment found; on the current pace the cleanup passes 40% by end of June, with the corpus gate set at 90%.

Wave 3 · Archive — Q4

Retention rules applied, legal-hold check with counsel, survivors to cold storage. Everything past retention with no hold is disposed of, with a log. This is where most of the 1.9M files end their run.

Dedup & authority rules

One current version per document — the library owner names the survivor when duplicates collide. The supersedes chain is recorded in the catalog, so history stays queryable without staying retrievable. A superseded file can be consulted; it can no longer be mistaken for current.

What "done" looks like

The K: drive goes read-only at the end of wave 2, so new files can't accumulate behind the migration. After wave 3 plus one quiet quarter — no access requests that the governed libraries couldn't serve — Tom Osei decommissions it.

Why waves, not a big bang

A single all-at-once migration optimizes for the announcement, not the outcome. Waves let wave 1 unblock daily work, wave 2 unblock the knowledge layer, and wave 3 happen calmly in Q4 — and if priorities shift, the roadmap loses a wave, not the whole project.



Phases, owners, and what it costs

PHASE	OWNER	TIMING	COST	STATUS
Inventory — six systems mapped, K: drive sized	Brad Taylor	March	Covered by the assessment	● Done
Pilot domain — tax templates (= the corpus cleanup)	Nicole Tran	May–Aug	Internal time	● In progress
CAAS + policies — libraries, catalog entries, QBO extracts	Nicole Tran · Elena Ruiz	Q3	Internal time + Azure storage, est. <\$400/mo	● Scheduled
Archive wave — retention, legal hold, cold storage, decommission	Tom Osei	Q4	Internal time + cold storage	● Scheduled

No seven-figure line item, on purpose. The infrastructure is SharePoint and modest Azure storage; the investment is ownership discipline, and that is paid in attention, not dollars.

What this unblocks, by backlog item

- #3 · Knowledge assistant

Reads the wave-2 corpus. Blocked until the cleanup gate passes — the build waits on this blueprint, not the other way around. (1,800 h/yr at stake)
- #2 · Intake automation

Files classified documents into governed libraries by naming convention, with owner and tier set on arrival. (2,100 admin h/yr)
- #4 · CAAS close workflows

Q3 build reads QBO extracts and the CAAS library instead of screen-scraping six places. (950 h/yr)
- Evaluation plan

Golden-dataset answers pin to cataloged source-of-truth versions — without the catalog, "correct version" isn't testable.

About this sample

This is the data blueprint an advisory client receives before any AI system gets built — right-sized for a 118-person firm, honest about what retires, and priced in discipline instead of infrastructure. Meridian & Frost is a fictional composite; the method is what you'd get.

Related samples

Governed Knowledge Layer Design Spec (first consumer)

Workflow Automation Spec (second consumer)

AI Executive Assessment Report (where the inventory came from)

bradtaylorai.com/advisory · brad@bradtaylorai.com