

The Governed Knowledge Layer

Your documents, retrieval-ready and permission-aware — with citations enforced and versions that tell the truth. The design an implementation team builds from, once the corpus gate passes.

PREPARED FOR

Meridian & Frost, PLLC
Leadership & implementation team

STATUS

v0.9 — draft for build approval
June 30, 2026 · build gate: corpus cleanup $\geq 90\%$

PREPARED BY

Brad Taylor
Fractional Chief AI Officer

What this is — and what it isn't

A governed retrieval layer over the firm's own documents, feeding Claude Enterprise. A staff member asks a question; the system retrieves relevant passages from a curated corpus — filtered to what that person is already allowed to open — and answers with citations to the source document and version. If the corpus can't support an answer, the system refuses and says so. That refusal is a feature, not a failure.

✓ What it is

A retrieval layer. It reads curated SharePoint libraries and answers from them — nothing else.

Permission-aware. Access rights travel with every chunk; retrieval enforces them inside the query.

Grounded. Every answer cites document and version, or the system stays silent.

✗ What it isn't

Not a chatbot on the raw K: drive. The K: drive never gets indexed — see rule 4, page 4.

Not a new place documents live. SharePoint stays the system of record; the layer only reads.

Not a replacement for CCH Axcess or Karbon. Tax prep and practice management stay where they are.

~707

DOCS IN PHASE-1
CORPUS

172 templates · 450
memos · 85 policies

1,800 h

ANNUAL SEARCH
BURDEN

Backlog #3 · score 81

100%

ANSWERS CITED — OR
REFUSED

No source, no answer

180 d

STALE-DOCUMENT
WARNING

Age flagged in every
answer

Why v0.9 and not v1.0

This design is complete but the build is gated: corpus cleanup must pass 90% (it is at 40% as of this draft). Approving the design now lets the builder start the moment the gate opens, instead of losing a month to procurement. Gates are on page 6.

In this specification

- 3 **Architecture** — six stages from curated library to cited answer
- 4 **The four rules** — permission, version truth, citations, curation. The heart of the design
- 5 **Corpus plan & ingestion** — the ~707 documents, the metadata envelope, the refresh cadence
- 6 **Launch gates & phasing** — three gates, rollout order, the 8–10 week build
- 7 **Risks & open decisions** — what can slip, and the two calls still unmade



Architecture — six stages, one direction

Documents flow one way: from curated SharePoint libraries to a cited answer. Every stage below runs inside the firm's Microsoft tenant except the model call, which goes to Claude Enterprise under the no-training addendum signed in April.



SharePoint — curated libraries

Three libraries: tax templates, research memos, policies. Every document has a named owner and a current version before it enters. This is the corpus — nothing outside it is visible to the system.



Ingestion

Parse → normalize → chunk at 400–800 tokens, split on heading boundaries. Version and permission metadata travel with every chunk (envelope on page 5).



Hybrid index

Semantic embeddings plus a keyword index. Both are queried on every question; results are merged and re-ranked. Keyword search catches the citations and form numbers that embeddings miss.



Permission-aware retrieval

The asker's Entra ID groups are applied as a filter inside the retrieval query — never as a screen after the fact. Content a user can't open never enters the model's context. Tested with permission traps.



Claude Enterprise API

Answers are drafted from retrieved passages only, under the signed no-training addendum. The model sees the chunks and the question — not the libraries, not the tenant.



Cited answer — or refusal

Every answer carries clickable citations to source document and version, plus stale or supersession warnings. Below the confidence floor, the system refuses and says what it searched.



Out of scope in phase 1

CCH Axxess workpapers, client PII stores, and the raw K: drive. None of it gets ingested, indexed, or retrieved. Phase-2 candidates go through the same gates.

Nothing new leaves the tenant

The knowledge layer reads; it never becomes the system of record. Documents stay in SharePoint, permissions stay in Entra ID, and the only external call is the governed model API.



The four rules that make it governed

Everything else in this document is plumbing. These four rules are the specification — if a build decision conflicts with one of them, the build decision loses.

1



Permission-aware retrieval

You can only retrieve what you could already open. The Entra ID group filter is part of the retrieval query itself, not a screen applied afterward — a post-filter can leak restricted content into the model's context even when the final answer hides it.

Tested with permission traps: six golden-dataset questions whose only correct answer is a refusal. The threshold is zero leaks, and it is a hard gate.

2



Version truth

Every document carries **valid_from** and **supersedes** metadata. Superseded documents stay in the index — staff still ask about them — but any answer built on one opens with a supersession warning naming the current version.

Documents untouched for 180 days get a stale flag in the answer. The 2023 mileage-rate memo cannot impersonate the 2026 one.

3



Citations or silence

Every answer carries clickable citations to the source document and version. If retrieval returns nothing above the confidence floor, the system refuses and reports what it searched. An uncited answer is a guess wearing a suit.

Refusals are logged and reviewed monthly — they are the map of what the corpus is missing, and they feed the phase-2 backlog.

4



Curated corpus only

A document enters the corpus when it has an owner and a current version — not before. The K: drive does not get indexed, in this phase or any other.

Pointing an assistant at an ungoverned drive is what killed the last generation of knowledge assistants: confident answers drawn from five versions of the truth. Rule 4 is the reason the gates on page 6 exist.

The acceptance test, in one sentence

A tax senior asks a question, gets an answer citing the current template with its version and date — and when she asks about a client folder she can't open, the system refuses without revealing that the folder exists.

Corpus plan & ingestion

LIBRARY (PHASE 1)	DOCUMENTS	OWNER	CLEANUP STATUS
Tax templates YELLOW	172 post-cleanup	Nicole Tran	40% complete · gate at 90% — 118 of ~290 duplicates retired so far
Research memos YELLOW	~450	Priya Raman	Owner assigned; current-version pass follows the template cleanup
Policies & procedures YELLOW	~85	Marcus Bell	Clean — versioned since the March policy work (v1.0 → v2.1)
Phase-1 total	~707		Corpus gate: ≥90% of templates deduplicated with owners attested

Chunk metadata — the envelope

```
// every chunk carries this envelope
{
  "doc_id": "memo-2026-0142",
  "title": "S-corp reasonable compensation",
  "owner": "P. Raman — tax research library",
  "tier": "2-yellow",
  "version": "2026.1",
  "valid_from": "2026-02-11",
  "supersedes": "memo-2023-0891 (2023.2)",
  "permissions_group": "MF-Tax-Staff",
  "chunk_heading_path": "Analysis > Safe-harbor"
}
```

Ingestion pipeline

- Parse** PDF, DOCX, XLSX → text with heading structure preserved
- Normalize** Tables flattened to rows; boilerplate and footers stripped
- Chunk** 400–800 tokens, split on heading boundaries, heading path retained
- Index** Embedding per chunk + keyword index; metadata stored alongside

Refresh cadence

- On change** SharePoint webhook re-ingests on create, update, or permission change
- Nightly** Full reconcile sweeps for anything the webhook missed
- Deletes** Removed documents leave the index within 24 hours, verified nightly

Why 400–800 tokens

Big enough to hold a complete thought (a safe-harbor test, a policy clause), small enough that a citation points at a passage a human can verify in seconds. Chunks never cross document boundaries, so a citation is always one document, one version.



Launch gates & phased rollout

Three gates, in order. No gate, no launch — the schedule bends, the gates don't.

1 **Corpus cleanup ≥90%**

Tax templates deduplicated to one current version each, owners attested. At 40% as of June 30; Nicole Tran tracks weekly in the monthly report.

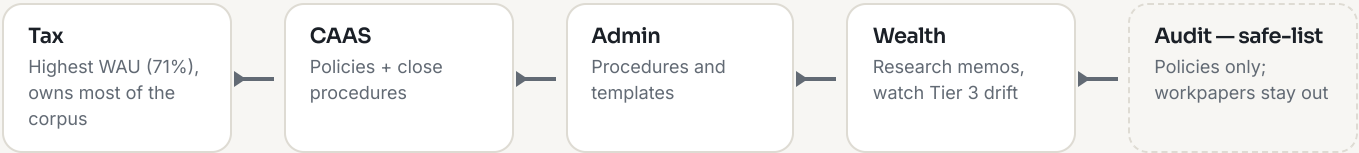
2 **Golden-dataset evals pass**

150 Q&A pairs incl. 20 traps (see the Evaluation Plan, Jun 20).
Thresholds: groundedness ≥95% · correct-version citation ≥98% · permission leaks 0.

3 **Champion pilot**

Two weeks, the 6 champions, live corpus and real questions. Exit: thresholds hold in the wild and champions sign off in writing.

Rollout order after the pilot



Build plan — 8–10 weeks after gate 1

- Weeks 1–2 · Ingestion & index**
Pipeline against the three libraries; metadata envelope verified on every chunk.
- Weeks 3–5 · Retrieval & enforcement**
Hybrid retrieval, in-query permission filter, citation and refusal behavior.
- Weeks 6–7 · Evaluation runs**
Golden dataset against gate-2 thresholds; fix, re-run, repeat until green.
- Weeks 8–10 · Pilot → rollout**
Champion pilot, then departments in the order above.

Build facts

- Window** 8–10 weeks, clock starts when gate 1 passes
- Budget** \$45–65K (held at HOLD in the June fund/hold/kill memo until gate 1)
- Builder** TBD under a separate SOW — open decision, page 7
- Direction** Brad Taylor directs; builder implements to this spec
- Calendar** No dates on purpose. The gate sets the date, not the other way around

Why the build waits

A knowledge assistant over five versions of the truth answers confidently with the wrong one. Cleanup first is slower by a quarter and better by years.

Risks & open decisions

RISK	HOW THIS DESIGN HANDLES IT
Corpus cleanup slips — 40% today, gate at 90%.	Then launch slips. That is the design working, not failing. Weekly tracking in the monthly report keeps the slip visible instead of quiet.
Version metadata discipline decays after launch, and version truth decays with it.	The owner rule: no owner, no entry. Owners attest their library quarterly, and the 180-day stale flags surface neglect inside the answers themselves — where it embarrasses someone into fixing it.
Permission model inherits SharePoint group hygiene — the filter is only as clean as the groups.	Tom Osei's pre-work, due before ingestion starts: audit Entra ID groups on the three libraries, replace direct-user grants with group grants, remove ex-employee and orphaned accounts. Permission traps in the golden dataset catch regressions after launch.

Open: builder selection

Last Rev built backlog #1 and #2 and knows the tenant; a second bid keeps the pricing honest. Decision needed roughly four weeks before gate 1 passes so the SOW is signed when the clock starts. Marcus owns the call; I write the recommendation.

Open: CAAS client folders in phase 2

Elena Ruiz wants client-specific folders in the phase-2 corpus. That is Tier 3 client data — it waits on the \$7216 consent refresh already in flight for the intake automation, and on counsel's opinion. Elena scopes it in Q3; it enters through the same four rules or not at all.

About this sample

This is the design specification an advisory client receives before a build is approved — the document that lets an implementation team start without a discovery phase, and lets leadership see exactly what they're buying. Meridian & Frost is a fictional composite; the rigor is what you'd get.

Related samples

Evaluation Plan — Golden Datasets & Benchmarks (the launch gate)

AI Data Foundation Blueprint (the corpus this reads from)

Why knowledge assistants fail (the essay behind rule 4)

bradtaylorai.com/advisory · brad@bradtaylorai.com