Xtensyon.

WHITE PAPER

# Data Classification for RAG: What to Index, What to Redact, What to Leave Alone

May 7, 2025 • Xtensyon Labs • 9 min read

> *RAG projects run into trouble when teams index everything “just in case.” This paper proposes a simple classification scheme that decides which content can be embedded, which needs redaction, and which should stay out of the index.*

## TL;DR

- Classify before indexing; fix policy after the fact is expensive.
- Separate content sensitivity from access control; you need both.
- Redact at ingestion and at query time for defense in depth.
- Keep an allowlist for fields that may enter prompts.

## Executive Summary

Teams often treat the vector index like a dump of internal knowledge. That approach creates security risk and retrieval noise. We outline a classification scheme with four tiers that maps directly to index rules, retention, and prompt policy. It is designed to be applied by content owners, not only by security teams, and it fits common enterprise systems like SharePoint and Confluence.

## Why It Matters

Once sensitive text is embedded, it becomes harder to reason about where it can surface. Even with ACLs, mistakes happen: wrong group sync, mis-tagged folder, or cached prompts. A clear classification plan reduces the chance of incidents and improves answer quality by keeping the index focused.

## What We Built

- A tiered policy that defines what may be embedded, what must be redacted, and what must be excluded.
- An ingestion pipeline that applies redaction and stores provenance in metadata.
- A prompt policy layer that blocks unsafe fields and adds warnings for borderline content.
- An audit report that summarizes indexed content by tier, owner, and source system.

## Observed Outcomes

- Cleaner retrieval after excluding duplicates and low-value content from indexing.
- Faster security review because rules were documented with examples and owners.
- Fewer “mystery answers” after enforcing provenance and versioned citations.

## Implementation Notes

- Start small: pick one business domain and apply the scheme before scaling.
- Use deterministic redaction for known patterns, then review edge cases manually.
- Do not embed raw emails without a policy. They contain more sensitive data than teams expect.
- Build a deletion workflow that removes both text and embeddings when sources are deleted.

## Governance & Risk

- Avoid security theater. A tag is useless without enforcement at ingestion and retrieval.
- Make classification part of content publishing, not a separate project.
- Log policy decisions so audits do not require guesswork.

## Release Checklist

- Do content owners understand the tiers and examples?
- Is redaction applied before embedding?
- Do prompts include only allowlisted fields?
- Can we delete embeddings when original sources are deleted?
- Do we have a monthly report for what is indexed and why?

## Conclusion

A good RAG index is curated. Classification is the simplest way to keep it safe and useful without turning governance into a heavyweight process.

## Keywords

data classification RAG PII redaction security governance