RORI

Research On Regulatory for Industry(s)

Macro Development Plan

Version 1.1
Created: February 13, 2026
Authors: Frank & Candi
Status: Approved

Changelog

v1.1 — Added Collection & Research swim lane (collect/*) with web scraping, file system ingestion, AI-assisted research discovery, source corroboration, and FTHB seed manifest.
v1.0 — Initial macro development plan.

Purpose

This document defines the macro development plan for RORI — a curated regulatory knowledge platform with agent-based retrieval. It is organized into phases with discrete components, each intended to be developed as a feature branch off of `main` in GitHub. Contributors pick up a branch, develop it, and submit a pull request back to main.

The plan follows the staged delivery approach defined in the RORI Project Proposal and is designed to produce manageable, well-scoped units of work.

Branch Naming Convention

All branches follow the pattern: **`phase{N}/{component-name}`** or **`collect/{component-name}`**

Swim Lane Overview

Phased Development (phase0–phase3): The core platform — research, engine, retrieval, agent, API, verticals, and scale. Components are sequenced with dependencies.

Collection & Research (collect/*): The data acquisition machinery — web scraping, file system ingestion, AI-assisted research discovery, and source corroboration. This swim lane runs in parallel with Phase 1 and feeds directly into the ingestion pipeline.

These tracks converge at `phase1/ingestion-pipeline` — the collection swim lane produces staged documents, and the ingestion pipeline processes them into the indexed repository.

Collection & Research Swim Lane (collect/*)

This swim lane builds the active data collection layer — the machinery that finds, fetches, corroborates, and stages regulatory source material. Without this, the platform has nothing to index.

The collection system is guided by a **Research Manifest** — a versioned, human-reviewed document that defines what sources to target, where they live, and how to access them. The manifest is seeded with known primary sources and expanded through AI-assisted deep research. No source enters the scraping or ingestion pipeline without appearing on an approved manifest.

Dependency: Requires `phase0/first-vertical-corpus-acquisition` (the initial cataloging effort). Can begin development in parallel with Phase 0 research branches and must be operational before Phase 1 ingestion pipeline testing begins.

`collect/research-manifest-system`

The research manifest is the steering mechanism for all data collection. It is a versioned, structured document (or database) that catalogs every regulatory source RORI targets — including URL or access method, format, jurisdiction, update frequency, and approval status. Workflow: Seed with known primary sources → AI-assisted expansion proposes new sources → Human reviews and approves → Approved sources enter scraping pipeline.

Deliverables:

- Manifest data model (source URL, format, jurisdiction, regulatory domain, update frequency, approval status, last fetched, last verified)
- CLI or UI for adding, reviewing, and approving manifest entries
- Version history for manifest changes
- Export format for downstream consumption by scraping and ingestion systems
- FTHB seed manifest populated from K4X-REG-001 reference (see Appendix A)

`collect/web-scraping-infra`

Automated, scheduled web scraping from authorized targets. Uses a managed scraping service (e.g., ScrapingAnt) to handle rotation, rate limiting, and JavaScript rendering. Only scrapes sources that appear on an approved research manifest. Target categories: Federal regulatory sources (.gov), GSE guides (Fannie Mae, Freddie Mac), state regulatory sources, authorized .org sources.

Deliverables:

- Scraping service integration (ScrapingAnt or equivalent)
- Target management — pull active targets from research manifest
- Scheduling engine — configurable per-source scrape frequency
- Change detection — compare fetched content against last-known version, flag changes
- Output staging — scraped documents land in staging area for ingestion pipeline pickup
- Rate limiting and politeness controls (respect robots.txt, configurable delays)
- Error handling, retry logic, and scrape audit log

`collect/filesystem-ingestion`

Watched directory / file system mount where documents can be dropped for automatic ingestion. Supports bulk import of regulatory documents obtained outside of web scraping — PDFs downloaded manually, FOIA responses, documents from partners, GSE guide archives.

Deliverables:

- Watched directory configuration (mount points, polling frequency)
- File type detection and validation
- Metadata extraction from filename conventions or sidecar files
- Deduplication against existing staged documents
- Staging area integration (same staging area as web scraping output)
- Bulk import CLI for one-time large corpus loads
- Import audit log

collect/ai-research-discovery

AI-assisted research to expand the research manifest beyond known sources. Uses web search and LLMs to proactively discover regulatory sources the seed manifest doesn't cover — finding what we don't know we're missing. Workflow: Prompt-driven search per regulatory domain → Citation chasing from known regulations → Research manifest proposal with confidence scores → Human review gate (nothing enters without approval).

Deliverables:

- Research prompt templates per regulatory domain
- Web search integration for source discovery
- LLM-based citation chain analysis
- Research manifest proposal format (structured output with confidence scores)
- Human review workflow (approve, reject, flag for further research)
- Discovery audit trail

collect/source-corroboration

Deep research to validate and cross-reference sources before they enter the repository. Verifies currency, checks for supersession, cross-references with authoritative secondary sources, and flags conflicts between sources (e.g., state law extending federal protections).

Deliverables:

- Currency check automation (compare against Federal Register, agency update logs)
- Cross-reference search for corroborating sources
- Supersession detection logic
- Conflict detection and flagging for human review
- Corroboration report per source (verified/unverified/conflicting, with evidence)
- Integration with research manifest (update source status based on corroboration)

collect/fthb-corpus-targeting

The FTHB-specific execution of the collection swim lane. Applies the manifest, scraping infrastructure, and corroboration pipeline to the First-Time Homebuyer regulatory domain. Covers: HUD Housing Counseling (24 C.F.R. Part 214), Fair Housing Act, RESPA, ECOA, GLBA, FCRA, GSE Guides, state/local requirements, and CFPB AI guidance.

Deliverables:

- FTHB seed manifest fully populated with primary source URLs and access methods
- Scraping schedules configured per source (daily for CFPB, weekly for GSE guides, monthly for CFR)
- Corroboration runs completed for all seed sources
- Gap analysis: regulatory areas covered vs. missing
- Staged corpus ready for Phase 1 ingestion pipeline

Phase 0: Foundation & Research

This phase resolves the open questions from the project proposal before application code is written. Branches produce decision documents, evaluation frameworks, and the project skeleton — not application code.

`phase0/project-scaffold`

Monorepo structure, CI/CD pipeline (GitHub Actions), linting, formatting, branch protection rules, contributing guide.

Deliverables:

- Repository layout and workspace configuration
- CI/CD pipelines (lint, test, build)
- Branch protection rules on main
- CONTRIBUTING.md with branch workflow documentation
- Pre-commit hooks

`phase0/research-retrieval-algorithms`

Deep research into state-of-the-art retrieval for legal and regulatory text. Hybrid retrieval, re-ranking models, deterministic retrieval modes for audit repeatability.

Deliverables:

- ADR with findings and recommendation
- Comparative analysis of retrieval approaches
- Benchmark references and prior art

`phase0/research-chunking-strategies`

Research into chunking approaches that preserve semantic integrity of regulatory text. Hierarchical, section-aware, cross-reference-preserving strategies.

Deliverables:

- ADR with findings and recommendation
- Sample chunking outputs against representative documents
- Evaluation criteria for chunk quality

`phase0/research-repository-architecture`

Evaluate the storage layer: vector DB, knowledge graph, hybrid. What serves regulatory data with complex hierarchical relationships — supersession chains, applicability hierarchies, jurisdiction trees.

Deliverables:

- ADR with findings and recommendation
- Storage architecture diagram
- Evaluation matrix of candidate technologies

`phase0/research-agent-framework`

Evaluate agentic orchestration patterns for tunable-depth regulatory reasoning. Tool-use patterns, planning, self-correction, citation threading.

Deliverables:

- ADR with findings and recommendation
- Framework comparison matrix
- Prototype interaction patterns

`phase0/evaluation-framework`

Define how RORI rigorously measures accuracy, completeness, consistency, and repeatability. Test harness and benchmark suite.

Deliverables:

- Evaluation methodology document
- Test harness scaffold
- Benchmark dataset definition
- Metrics definitions and measurement approach

`phase0/first-vertical-corpus-acquisition`

Identify, catalog, and begin acquiring the Mortgage/First-Time Homebuyer regulatory corpus. Seeds the research manifest for the collection swim lane.

Deliverables:

- Source catalog with URLs, formats, and access methods
- Licensing and usage rights assessment
- Sample documents staged for ingestion testing
- Corpus coverage map
- Initial research manifest seed entries

Phase 1: Core Engine — Single Vertical (Mortgage/Homebuyer)

Build the working system end-to-end against the first vertical. The collection swim lane feeds staged documents into the ingestion pipeline.

`phase1/ingestion-pipeline`

Handles disparate document types. Accepts staged documents from both the collection swim lane and manual imports. Must be robust and extensible.

Deliverables:

- Document type detection and routing
- Text extraction for each supported format
- Structure preservation (headings, sections, tables, lists)
- Staging area integration (collection swim lane output)
- Error handling and ingestion reporting
- Extensibility pattern for new document types

`phase1/curation-enrichment`

Metadata extraction and enrichment: jurisdiction tagging, effective dates, applicability scope, supersession chains. Quality gates and validation checks.

Deliverables:

- Metadata extraction pipeline
- Deduplication logic
- Quality gate definitions and validation rules
- Enrichment audit trail

`phase1/semantic-chunking`

Implements the chunking strategy from Phase 0. Section-aware, hierarchy-preserving, cross-reference-maintaining.

Deliverables:

- Chunking implementation per ADR
- Cross-reference link preservation
- Hierarchy metadata per chunk
- Evaluation framework benchmark validation

`phase1/indexing-layer`

Hybrid index per Phase 0 architecture decision. Dense vector embeddings, sparse/lexical index, structured metadata index, optional graph index.

Deliverables:

- Index creation pipeline
- Embedding generation
- Metadata index for structured queries
- Graph index for regulatory relationships (if applicable)

`phase1/retrieval-engine`

Core retrieval: hybrid search, re-ranking, confidence scoring, coverage estimation. Deterministic retrieval modes for audit repeatability.

Deliverables:

- Hybrid search implementation
- Re-ranking pipeline
- Confidence scoring per result
- Coverage estimation
- Deterministic retrieval mode
- Evaluation framework benchmark results

`phase1/agent-core`

Agent layer: orchestrates retrieval, reasoning, synthesis. Tunable response depth. Citation threading. Hallucination guardrails.

Deliverables:

- Agent orchestration per ADR
- Tunable depth parameter
- Multi-step reasoning
- Hallucination guardrails
- Context window management

`phase1/citation-provenance`

End-to-end citation provenance: source document → section → version → ingestion timestamp → agent response. Non-negotiable.

Deliverables:

- Provenance data model
- Citation attachment at every stage
- Provenance query API
- Audit report generation

`phase1/version-control-regulatory`

Change tracking for living regulatory documents. Version history, supersession detection, effective date management.

Deliverables:

- Document version tracking
- Supersession chain detection
- Effective date management
- Re-ingestion workflow
- Citation stability after updates

`phase1/developer-api-alpha`

Alpha API surface for agent invocation with tunable parameters. Enough for third-party AI pipelines.

Deliverables:

- REST API endpoints
- Request/response schema with tunable parameters
- Authentication and access control
- API documentation
- Example integration patterns

Phase 2: Multi-Vertical & Advanced Capabilities

`phase2/vertical-insurance`

Insurance vertical: state-by-state regulation, broker/agent compliance. Highly jurisdictional (50 states).

Deliverables:

- Insurance corpus acquisition and cataloging
- Ingestion pipeline validation
- Jurisdiction mapping (state-by-state)
- Evaluation framework benchmarks

`phase2/vertical-medical-gig`

Medical/Gig Platform vertical: labor law, medical licensing, telehealth, platform compliance.

Deliverables:

- Corpus acquisition and cataloging
- Multi-domain intersection handling
- Evaluation framework benchmarks

`phase2/cross-corpus-comparison`

Document-to-document comparison engine. Structured diff: alignments, gaps, conflicts.

Deliverables:

- Comparison engine
- Structured diff output format
- Multi-corpus query support
- Comparison result provenance

`phase2/gap-analysis-engine`

Requirements corpus vs. compliance posture. Structured gap reports with prioritized findings.

Deliverables:

- Gap detection logic
- Coverage scoring
- Prioritized gap report generation
- Remediation guidance linkage

`phase2/rubric-quality-assessment`

Score work product against regulatory rubrics. Flag deficiencies, produce remediation guidance.

Deliverables:

- Rubric definition format

- Scoring engine
- Deficiency flagging with source references
- Remediation output

`phase2/synthesis-output`

Transform analysis into structured deliverables: compliance plans, gap reports, requirement matrices.

Deliverables:

- Output template system
- Compliance plan generation
- Requirement matrix generation
- Export formats

`phase2/multilingual-support`

Multilingual ingestion, cross-language retrieval, translation-aware semantic matching.

Deliverables:

- Multilingual ingestion pipeline
- Cross-language retrieval
- Translation-aware matching
- Jurisdiction-language mapping

`phase2/packaged-application`

Front-end paired with agent for specific use cases. First end-user experience.

Deliverables:

- Front-end application
- Pre-configured use case workflows
- User authentication and session management

`phase2/api-hardened`

Production-grade API: rate limiting, access control, usage metering, SDK.

Deliverables:

- Rate limiting
- Granular access control
- Usage metering and billing hooks
- SDK or integration library
- Production API docs

Phase 3: Scale & Fine-Grained Expansion

`phase3/fine-grained-verticals`

Municipal codes, local zoning ordinances, sub-regulatory guidance. Maximum granularity.

Deliverables:

- Fine-grained corpus acquisition and ingestion
- Scale validation at high document volumes

`phase3/change-monitoring`

Proactive regulatory change monitoring: detect updates, assess impact, trigger alerts.

Deliverables:

- Source monitoring and change detection
- Impact assessment engine
- Alert and notification system

`phase3/performance-at-scale`

Optimization for large corpus volumes, concurrent queries, multi-tenant workloads.

Deliverables:

- Performance benchmarks
- Query optimization
- Caching strategies
- Multi-tenant isolation

`phase3/vertical-marketplace`

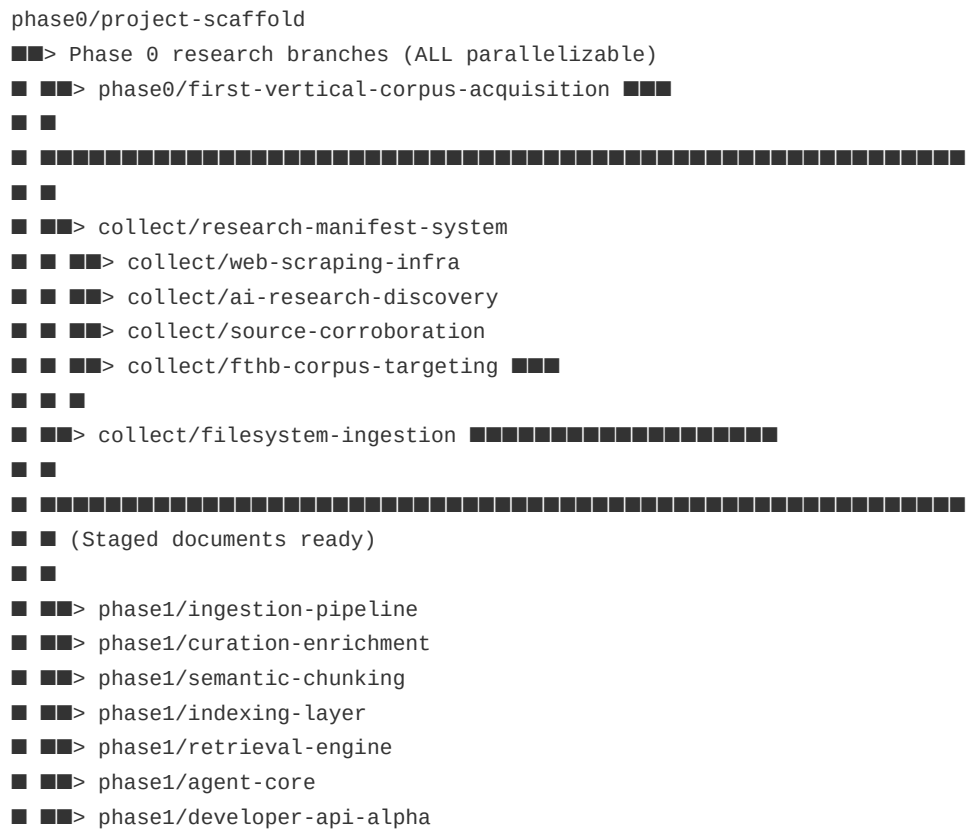
Plug-in model for vertical-specific configurations. Third-party contributions.

Deliverables:

- Plugin specification format
- Marketplace infrastructure
- Contribution and certification workflow

Sequencing & Dependencies

Critical Path



Parallel Work Opportunities

Phase 0: All research branches can be worked simultaneously. evaluation-framework and first-vertical-corpus-acquisition can also run in parallel.

Collection swim lane: Once first-vertical-corpus-acquisition produces the seed catalog, research-manifest-system should land first. Then web-scraping-infra, filesystem-ingestion, ai-research-discovery, and source-corroboration can develop in parallel. fthb-corpus-targeting is the integration branch that exercises them all.

Phase 1: Once indexing-layer lands, citation-provenance and version-control-regulatory can develop alongside retrieval-engine and agent-core.

Phase 2: All vertical onboarding and cross-corpus capability branches can run in parallel.

Branch Lifecycle

Step	Action	Description

1	Create	Branch from main using phase{N}/{name} or collect/{name}
2	Develop	Work the component per deliverables in this plan
3	Test	Validate against evaluation framework (Phase 1+)
4	PR	Submit pull request to main with documentation
5	Review	Peer review and approval
6	Merge	Squash merge to main

Appendix A: FTHB Regulatory Seed Manifest

Initial primary source targets for the First-Time Homebuyer vertical, derived from K4X Regulatory & Compliance Reference (K4X-REG-001 v1.0).

Federal Statutes and Regulations

Regulatory Domain	Citation	Primary Source
HUD Housing Counseling	24 C.F.R. Part 214	ecfr.gov, HUD.gov
Fair Housing Act	42 U.S.C. §§ 3601–3619	uscode.house.gov, HUD.gov
RESPA	12 U.S.C. §§ 2601–2617	uscode.house.gov, CFPB
RESPA Regulation X	12 C.F.R. Part 1024	ecfr.gov, consumerfinance.gov
ECOA	15 U.S.C. §§ 1691–1691f	uscode.house.gov, CFPB
ECOA Regulation B	12 C.F.R. Part 1002	ecfr.gov, consumerfinance.gov
GLBA	15 U.S.C. §§ 6801–6809	uscode.house.gov
FTC Safeguards Rule	16 C.F.R. Part 314	ecfr.gov, ftc.gov
FCRA	15 U.S.C. §§ 1681–1681x	uscode.house.gov, consumerfinance.gov

Agency Guidance and Enforcement

Source	URL Pattern	Content Type
CFPB Guidance & Rules	consumerfinance.gov/policy-compliance/	Guidance, bulletins, enforcement
CFPB Mortgage Resources	consumerfinance.gov/owning-a-home/	Consumer guidance, educational
CFPB AI/Chatbot Guidance	consumerfinance.gov	Enforcement signals, opinions
HUD Housing Counseling	hud.gov/program_offices/housing/	Program rules, HCS docs
HUD-9902 Reporting	hud.gov	Reporting formats, XML schemas
FTC Privacy Guidance	ftc.gov/legal-library/	Safeguards Rule, enforcement
DOJ Fair Housing	justice.gov/crt/fair-housing	Enforcement, settlements

GSE Seller/Servicer Guides

Source	URL Pattern	Content Type
Fannie Mae Selling Guide	singlefamily.fanniemae.com/originating	Full guide, updates
Fannie Mae Servicing Guide	singlefamily.fanniemae.com/servicing	Full guide, updates
Freddie Mac Guide	guide.freddiemac.com	Full guide, bulletins

Fannie Mae HomeReady	fanniemae.com (HomeReady)	Program requirements
Freddie Mac Home Possible	freddiemac.com (Home Possible)	Program requirements

State and Local (Initial Targets)

Source Category	Discovery Method	Priority
State fair housing statutes	AI research discovery + manual review	High
State housing finance agencies	Known URLs per state (50 targets)	High
State DPA program guides	AI research discovery per state	Medium
State data privacy laws	Known citations + AI expansion	Medium
Municipal DPA programs	AI research discovery per metro	Lower (Phase 2+)

Corroboration Sources

Source	Purpose
Federal Register (federalregister.gov)	Verify currency, find amendments and proposed rules
Congress.gov	Track legislative changes to underlying statutes
Legal commentary (law reviews, bar guides)	Corroborate interpretation of regulatory requirements
Agency FAQs and interpretive letters	Validate practical application of rules

---

v1.1 — Updated February 13, 2026 — Frank & Candi
Changelog: Added Collection & Research swim lane and FTHB Regulatory Seed Manifest (Appendix A)