

The Research Data Director

Building the Connective Tissue of Scientific Discovery

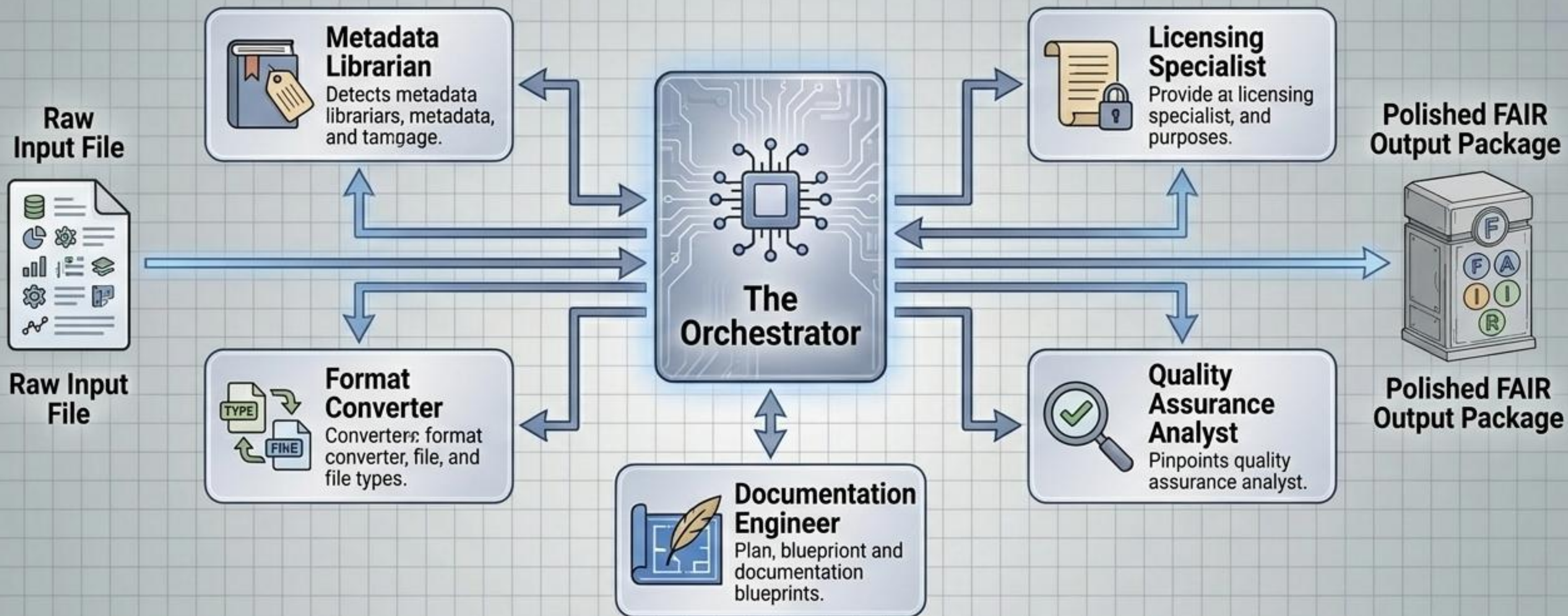
A Persona-Based Agentic
AI Architectural Blueprint
for Research Data,
Metadata, and Schema
Creation

Based on the RDA–Microsoft Research Data
Director Working Group System Design
(Version 0.8.1)



Ray Uzwysyn, Ph.D. MBA MLIS
Acting AUL and Director of
Research and Technology Services
University of California, Riverside
ruzwyshyn@gmail.com
[linkedin.com/in/rayuzwyshyn/](https://www.linkedin.com/in/rayuzwyshyn/)

The System: Six Agents, One Goal

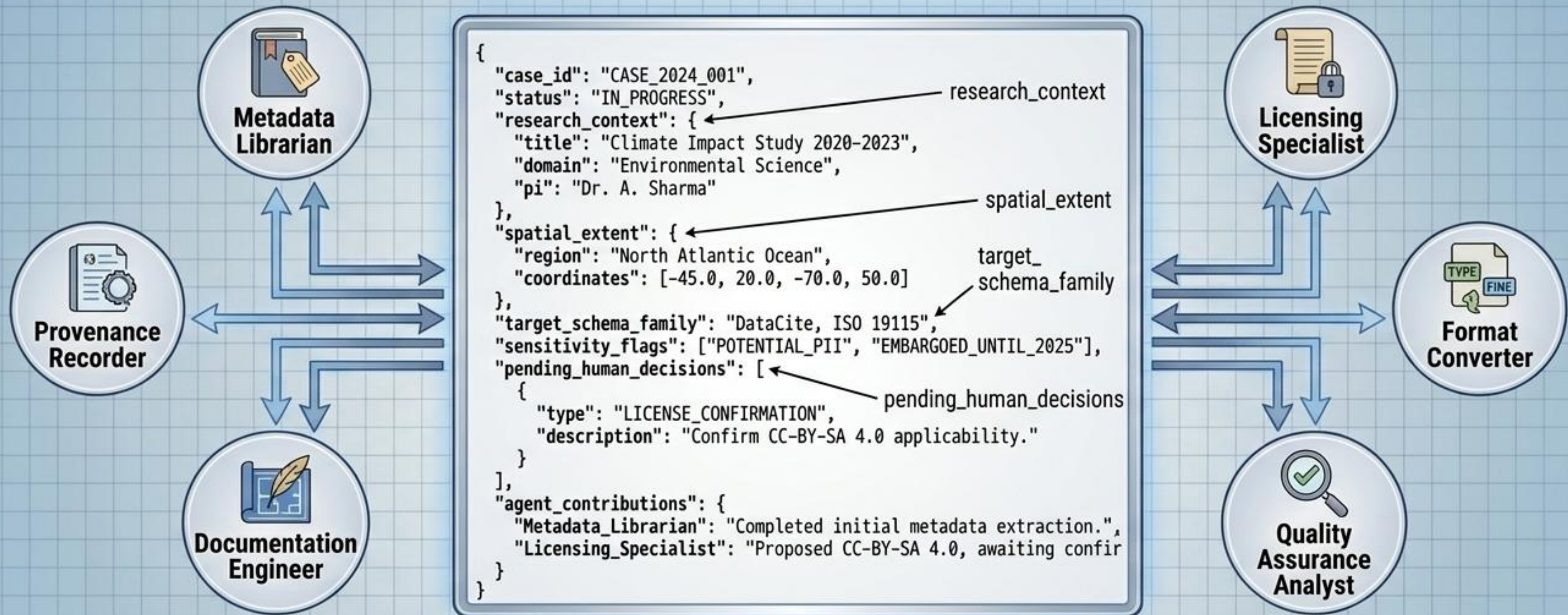


The **Research Data Director** is a team of six AI agents, each modeled on a real professional role in an R1 university research library. They take a researcher's raw dataset and produce a fully automated, FAIR-compliant, cross-disciplinary publication package. Language models reason from the situation rather than walk a predetermined script.

The Six Core Agents

AI Agent Name	Professional Analogue	Core Output	Trigger for Human Escalation	Under-the-Hood LLM Pattern
Research Data Director	Head of Data Services	Dispatch plan & gates	Irreversible acts (DOI minting)	Orchestrator
Research Data Librarian	First-Contact Intake Specialist	Working Brief (JSON)	Sensitivity/Privacy flags	Sequential Intake
Metadata Librarian	Senior Cataloger	DataCite/EML/DDI Record	Low-confidence schema mappings	Generator
Fact Checker	Research Integrity Specialist	Verification API Report	Unresolvable external APIs	Evaluator (Parallel)
FAIR Expert	FAIR Assessor	F-UJI Metric Score	Sub-threshold FAIR scores	Evaluator (Parallel)
Provenance Recorder	Digital Archivist	PROV-O Audit Log	(None - ambient listener)	Continuous Audit

Shared Memory: The Working Brief



Every agent reads from and writes to a single **shared object**. It is a structured JSON document that accumulates throughout the case. It is what a senior librarian's case notes have always been—now entirely legible to the entire team at once, and auditable by human reviewers.

The Intake: A Structured Sensemaking Conversation

Situation

"What is the research context?"



Gap

"What documentation is currently missing?"



Bridge

"What kind of help is most needed right now?"



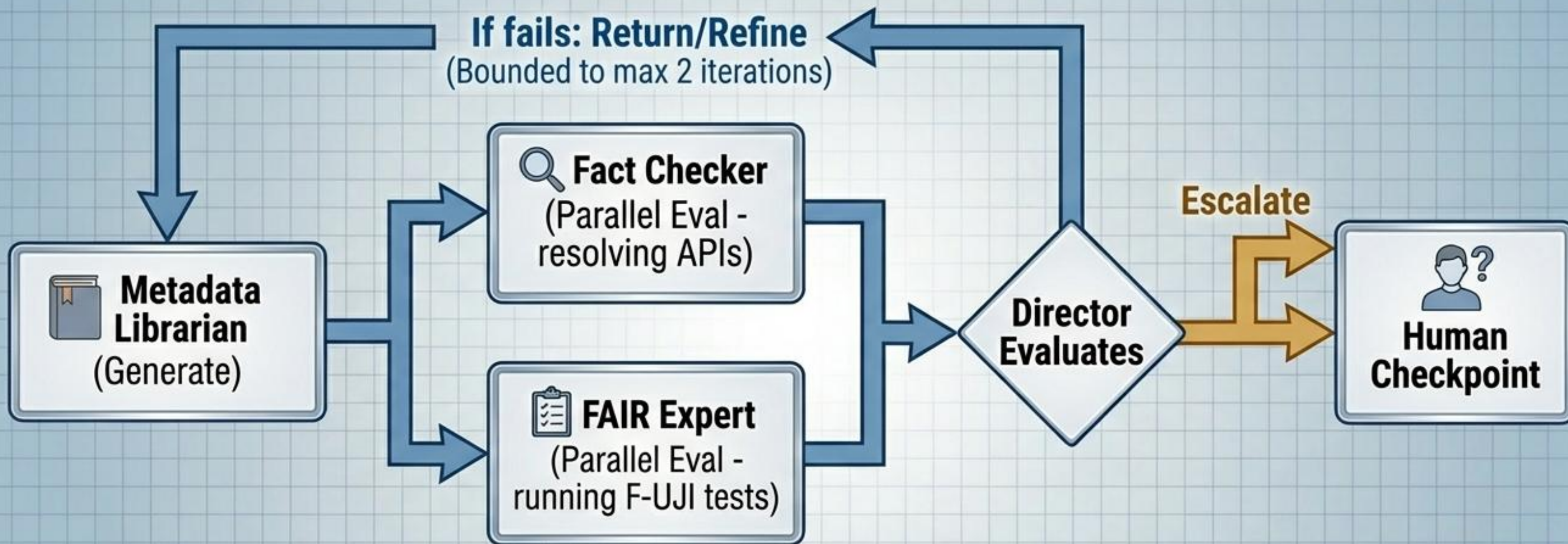
Use

"Who else might use this data in the future?"



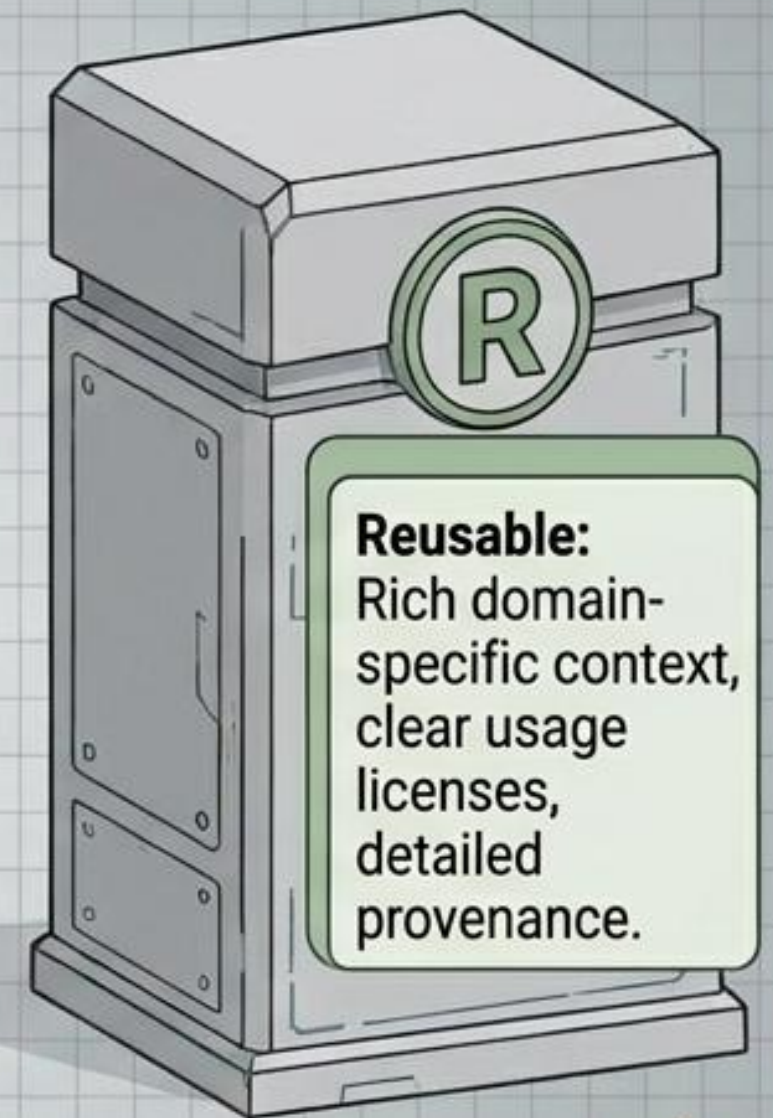
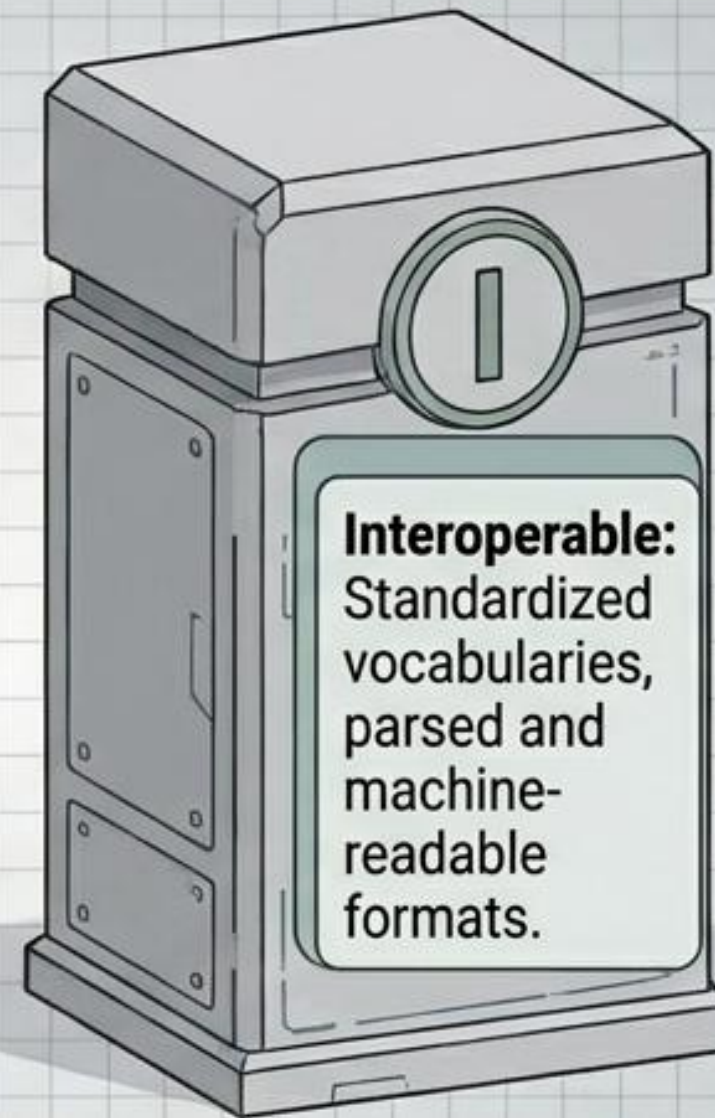
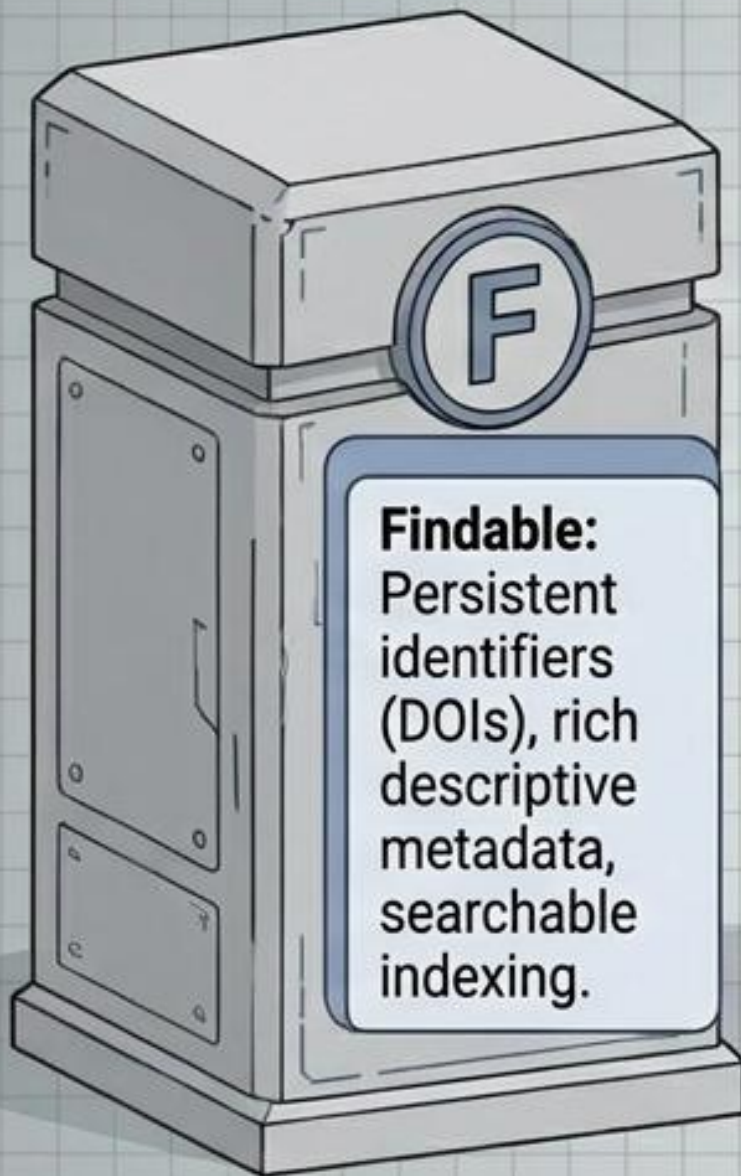
The Librarian reads manuscript drafts alongside the dataset to infer **instrument types** and **variables** before the interview even begins. She actively flags **sensitivity signals** (e.g., endangered species locations) to halt the process and wait for human instruction.

The Inner Loop: Generation and Parallel Critique



This automated inner loop iterates on metadata, combining parallel critique from different expert agents. It ensures robustness before escalation to a human.

FAIR is the Quality Standard for Metadata



A 2025 survey in **Nature Scientific Data** assessed 71 public drone datasets and found systematic failures in provenance and licensing. The global standards infrastructure is mature, but the human capacity to apply it is completely bottlenecked.

Sensitivity and Sovereign Data



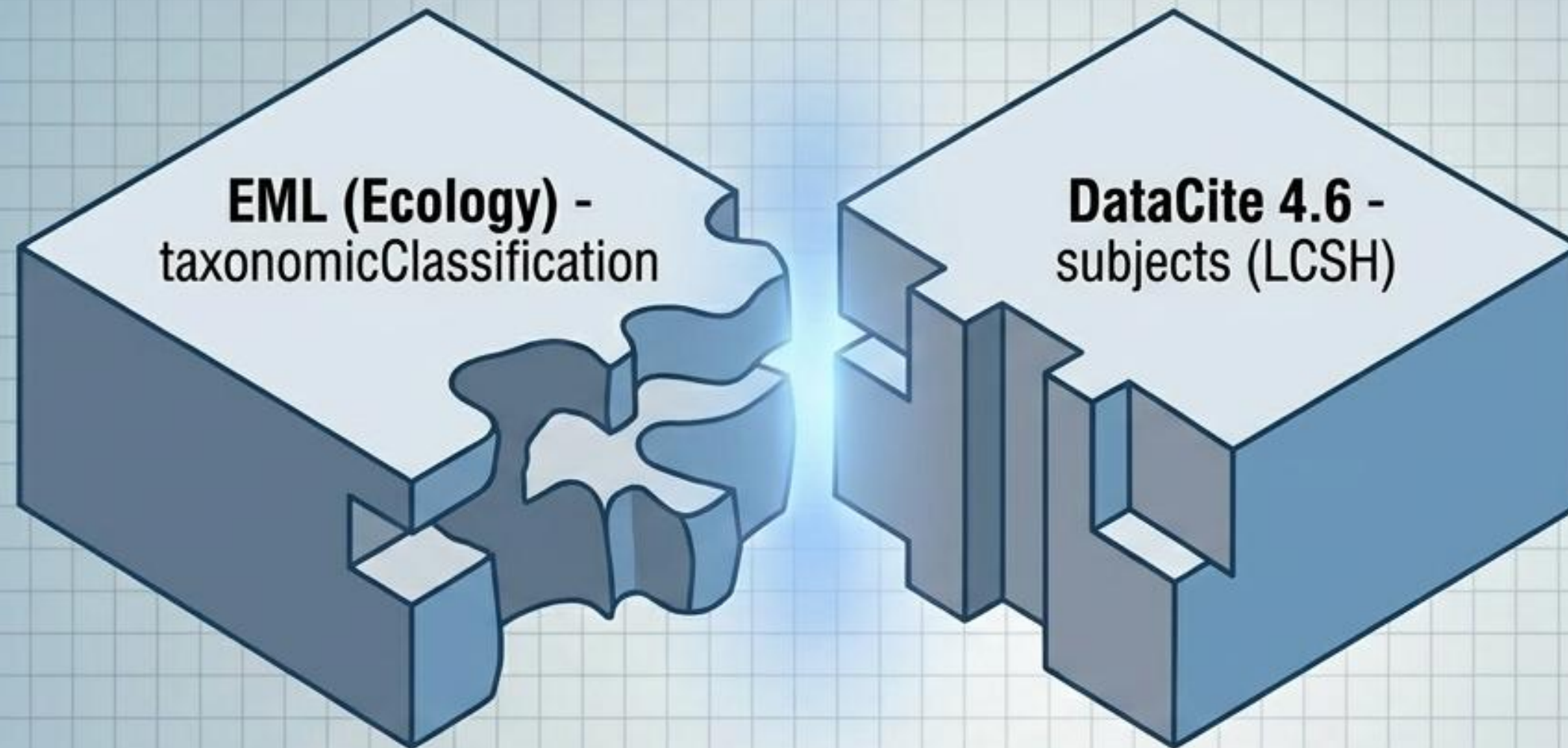
When the working brief detects signals of Indigenous provenance or sensitive locations, the system fires a CARE flag. The system has no authority to determine how CARE applies; it stops the case completely and routes it to the relevant human authority.

Eliminating Hallucination via Authoritative APIs



The defining failure mode of current AI is confident fabrication. The Fact Checker treats every claim as potentially hallucinated until an external tool call confirms it. Unverified claims are never silently deleted; they are flagged with proposed corrections.

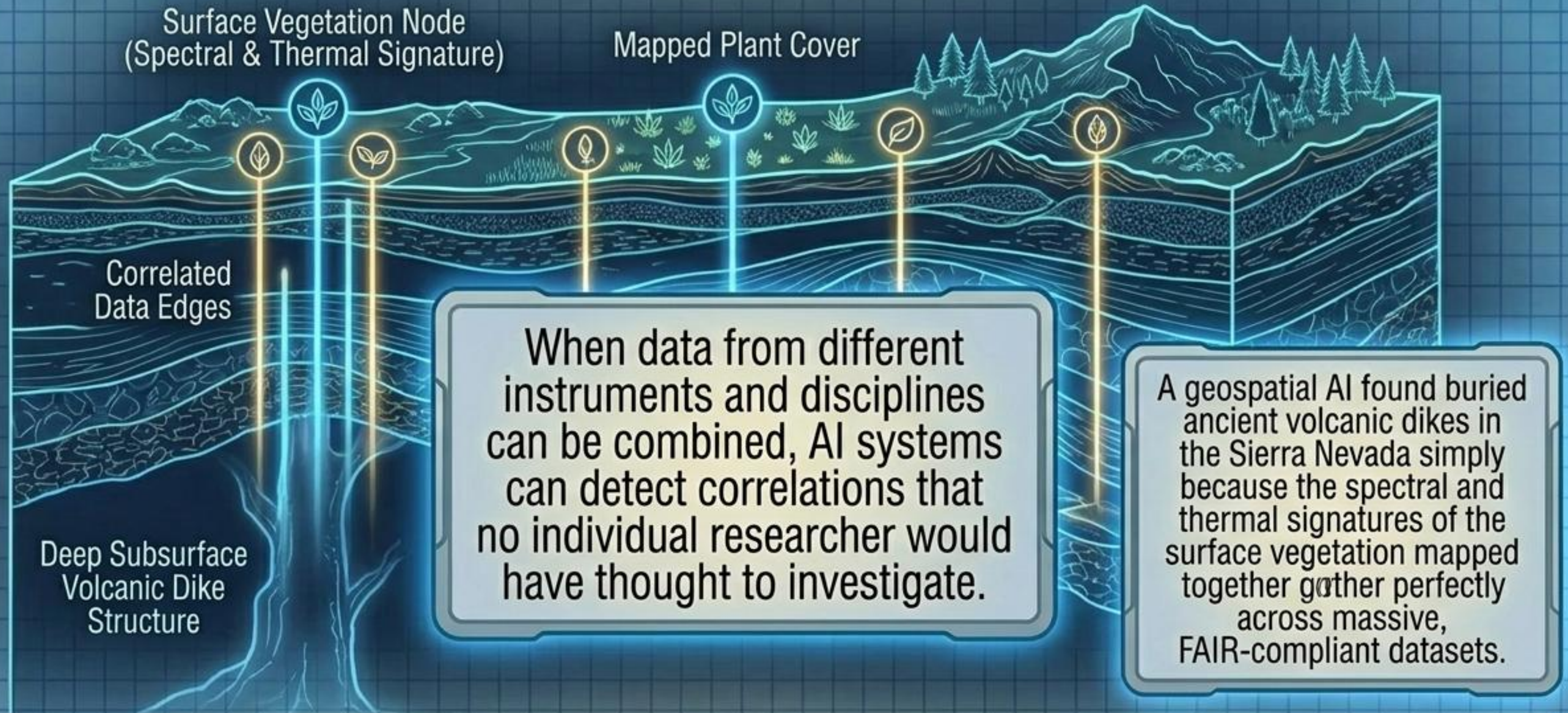
The Frontier of Discovery: Schema Crosswalks



Recent research (Andronikidis et al., 2025) shows **LLMs achieve 73.5% accuracy** on zero-shot ontology alignment. This is high enough to be incredibly useful, but low enough to mandate strict human oversight.

A metadata crosswalk is a field-by-field translation table between two distinct schemas. Routine crosswalks are pre-established by communities. The true frontier is the crosswalk that has never been built—connecting two disciplines that have never needed to talk before.

The Substrate for Unexpected Discovery



Cross-disciplinary metadata is not administrative overhead.
It is the precondition from which unexpected questions can emerge.

Where Humans Decide

AI Autonomy: Agent Proposals

Drafting Records

Mapping Schemas

API Verification

Scoring FAIR metrics

AI acts within predefined bounds.

— The Irreversible Checkpoint —

Human Authority

Minting DOIs

Assigning Licenses

Publishing to Repositories

Managing CARE/Sensitive Indigenous Data

Final, irreversible decisions are human-made.

Human oversight is a property each agent carries individually. The boundary principle: agents read freely and propose anything. The irreversible act belongs solely to a human.

Transparent Deliberation: The Thinking Window

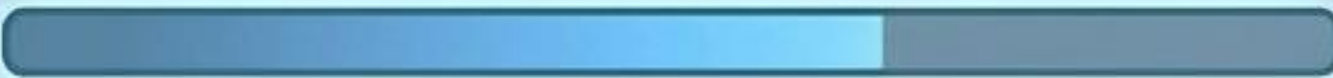
Structured Metadata Record

Record Title: Freshwater Ecology Data

Status: Compiling - 65%

Data Type: EML

Geographic Bounds: [Pending...]



The Thinking Window

07:23:15:17 PMT

Director: Working brief confirmed. Freshwater ecology. Target: EML.

07:23:26:07 PMT

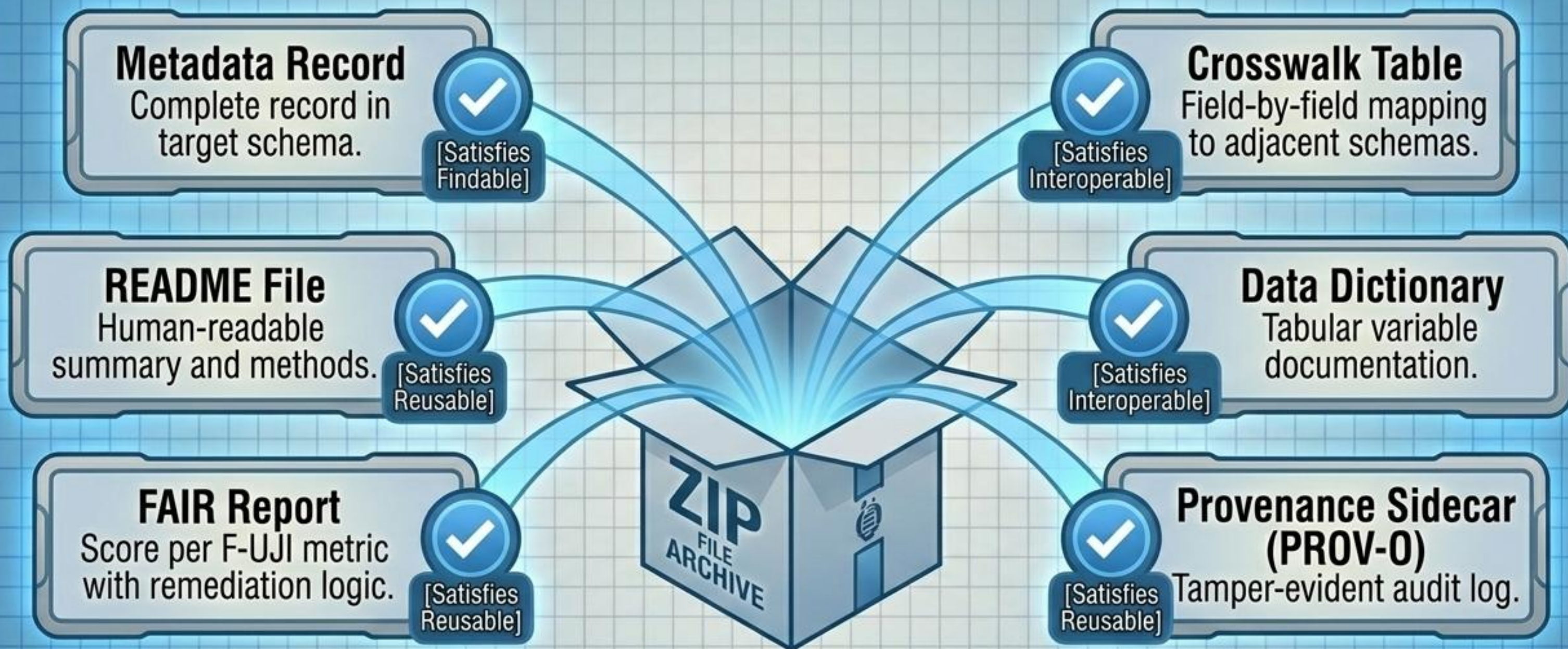
Ethics Manager: Nesting coordinates found for listed species. Recommending 1km obfuscation. Hold STAC geometry field.

07:23:33:19 PMT

Director: Noted. Metadata Librarian, use 1km bounding box placeholder pending human authorization.

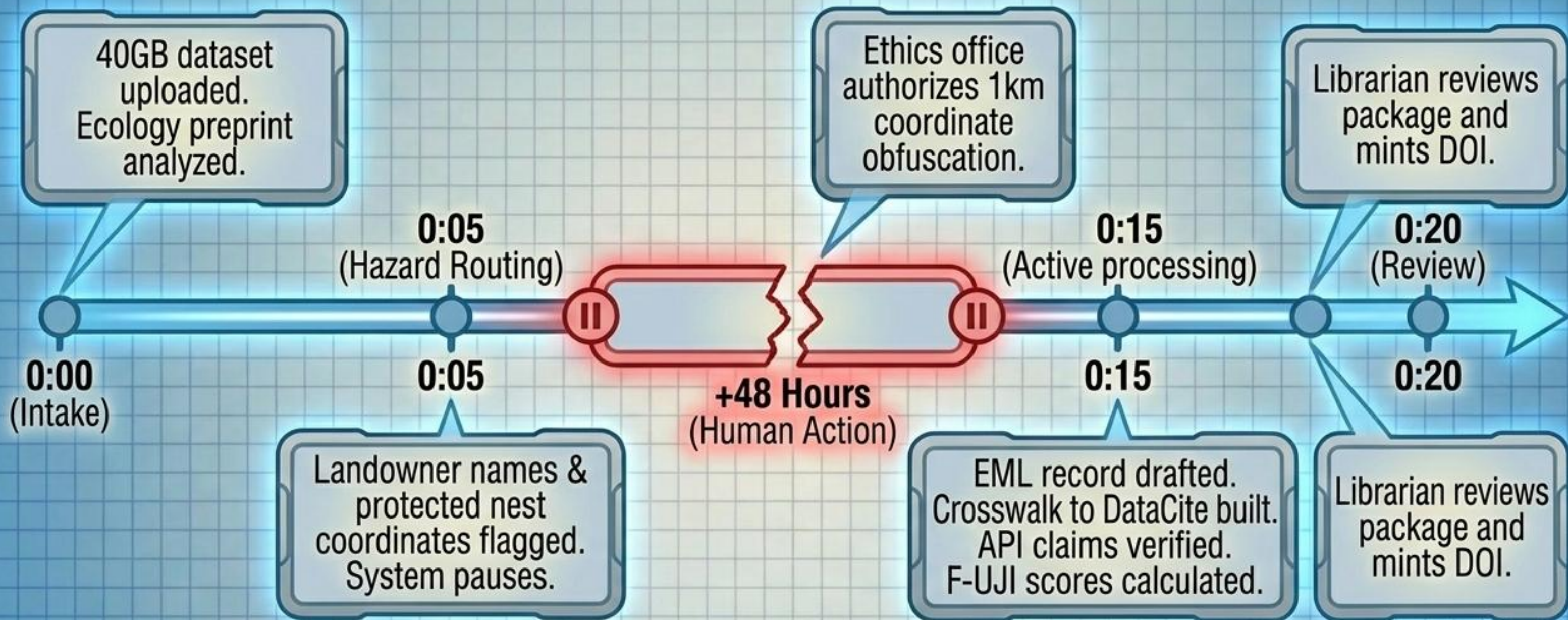
The Thinking Window surfaces the agents' internal deliberations as they navigate uncertainty. It is opt-in, fully archivable, and allows the researcher to intervene directly into the agents' workflow.

The Final Output Package



A complete, submission-ready bundle. The repository receives everything it needs; the researcher submits exactly once.

Worked Example: An Afternoon in the Life



Total active researcher time: approximately two hours. Every other task—schema selection, crosswalk construction, verification, and formatting—was completed by the system.

Comments & Questions

Building the Connective Tissue of Scientific Discovery

A Persona-Based Agentic
AI Architectural Blueprint
for Research Data,
Metadata, and Schema
Creation

Based on the RDA-Microsoft Research Data
Director Working Group System Design
(Version 0.8.1)

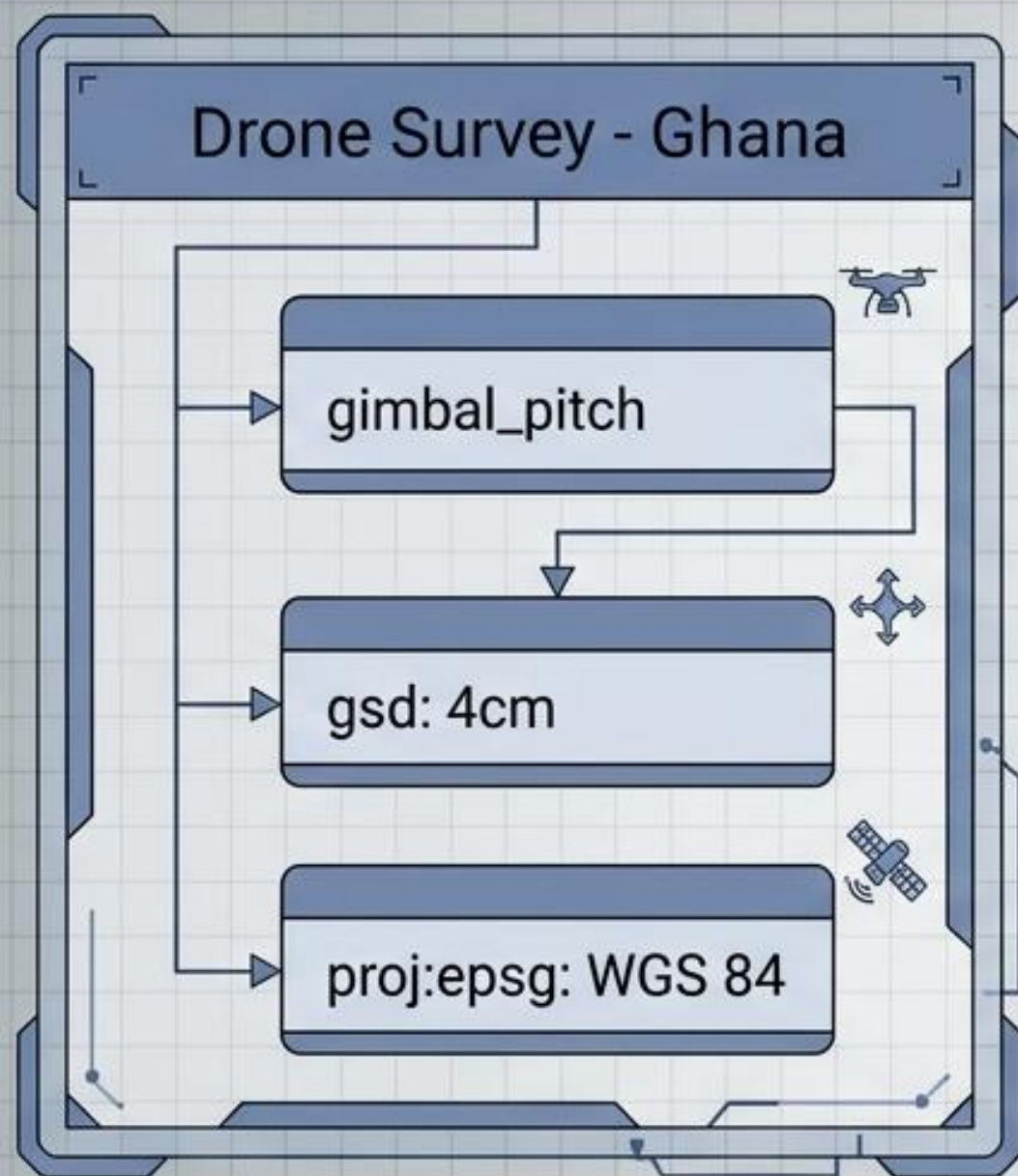


NotebookLM

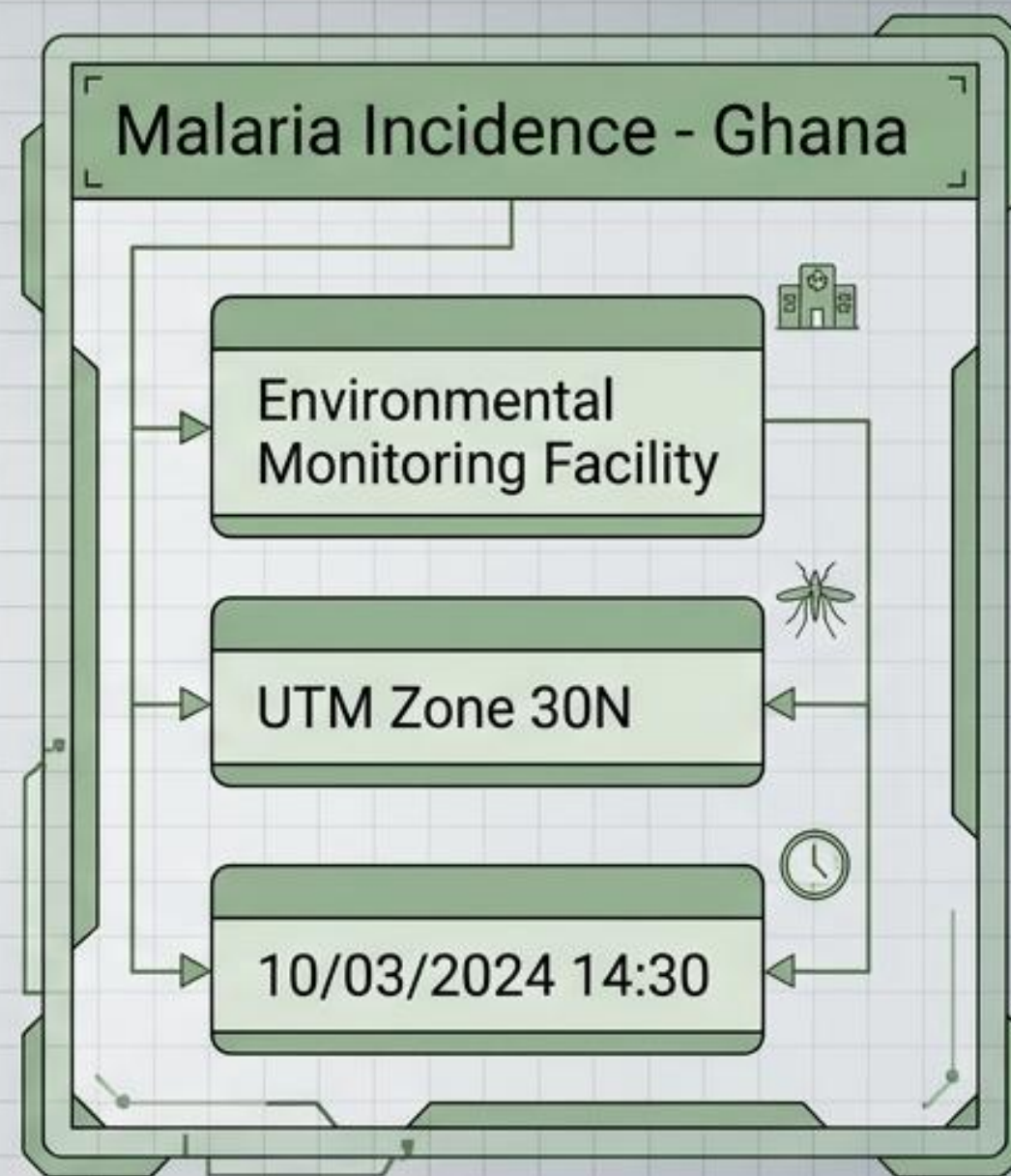
[Online Blueprint Draft and
Appendices: Personas,
Technical Details, Presentation Draft](#)

Ray Uzwysyn, Ph.D. MBA MLIS
Acting AUL and Director of
Research and Technology Services
University of California, Riverside
ruzwyshyn@gmail.com
[linkedin.com/in/rayuzwyshyn/](https://www.linkedin.com/in/rayuzwyshyn/)

When Data Cannot Find Data

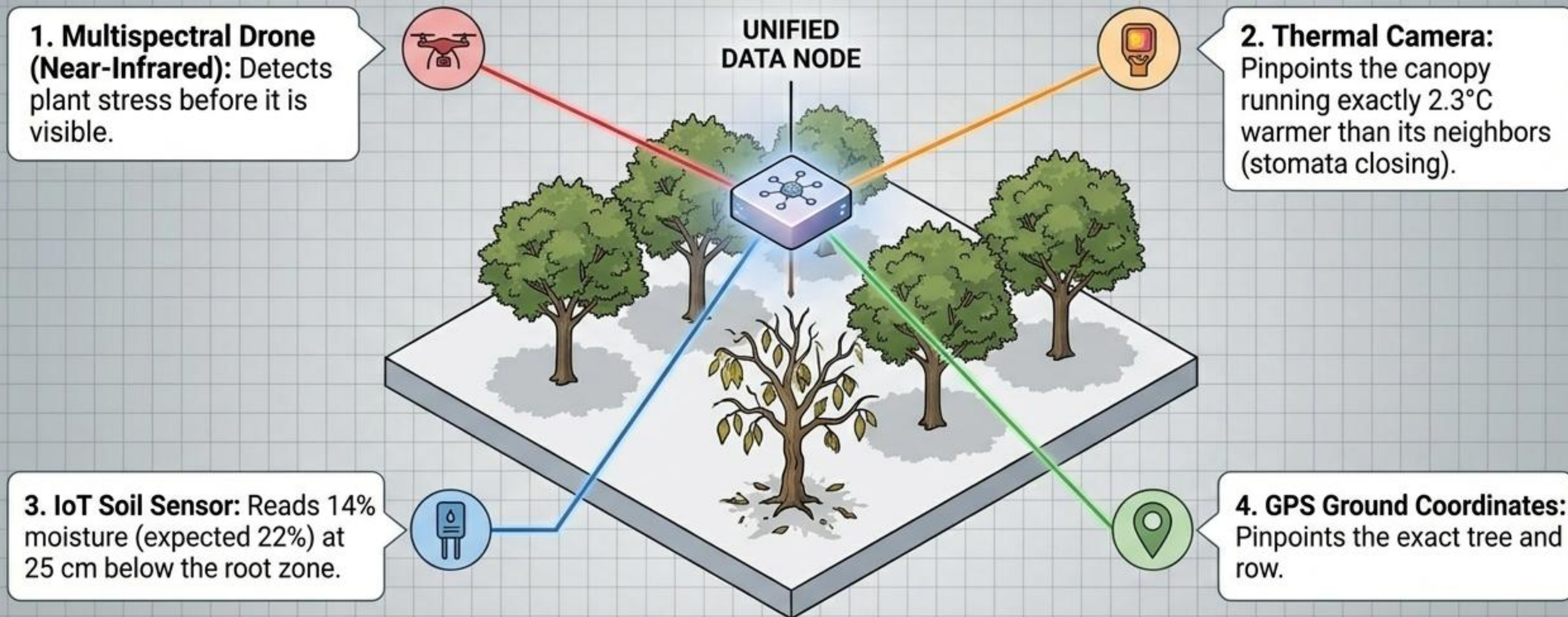


Synthesis Gap



Both datasets exist in public repositories, but cannot be combined to correlate standing water with disease outbreaks because their coordinates, timestamps, and schemas are fundamentally incompatible.

Discovery Requires Synthesis



No single instrument identifies the problem. **The knowledge lives only in the synthesis.** Without shared coordinate systems and calibrated timestamps in the metadata, diagnosing the failed irrigation emitter is impossible.

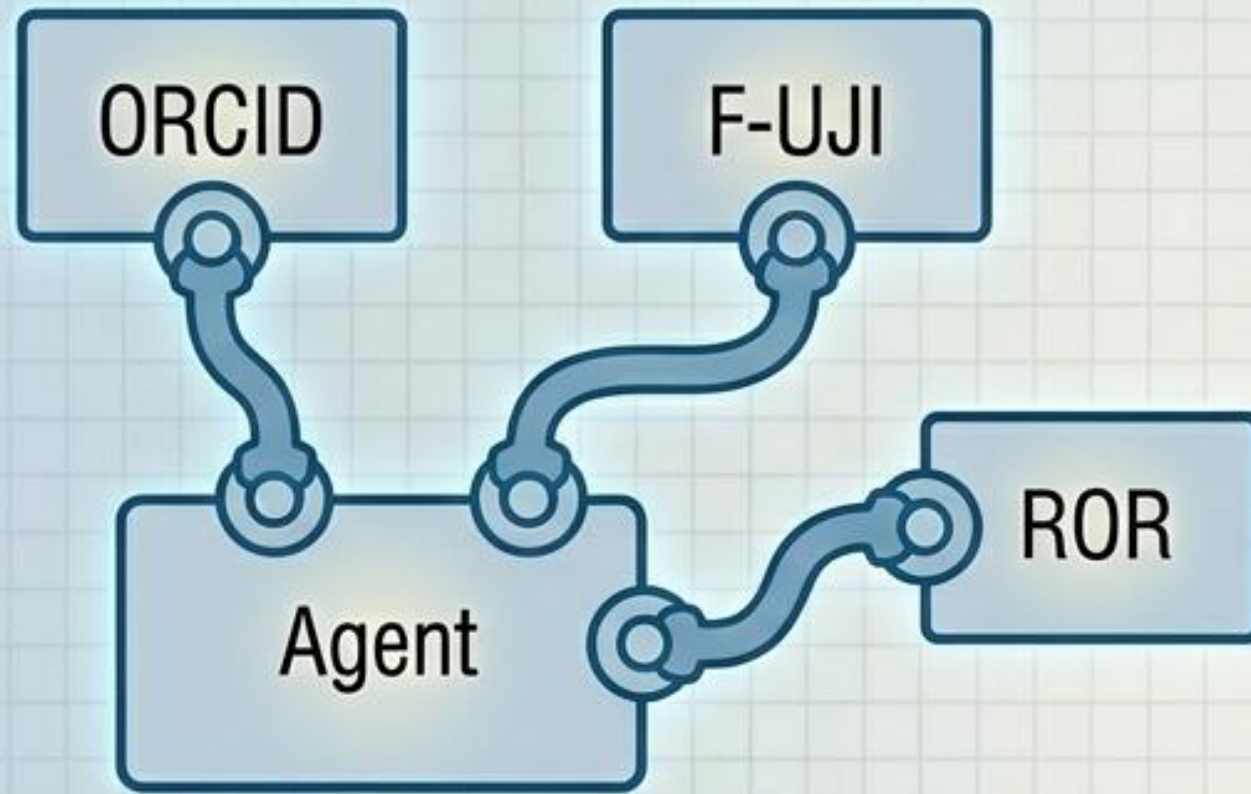
Reasoning the Bridge Between Disciplines



The system does not assert a bridge as fact;
it proposes a hypothesis for confirmation.

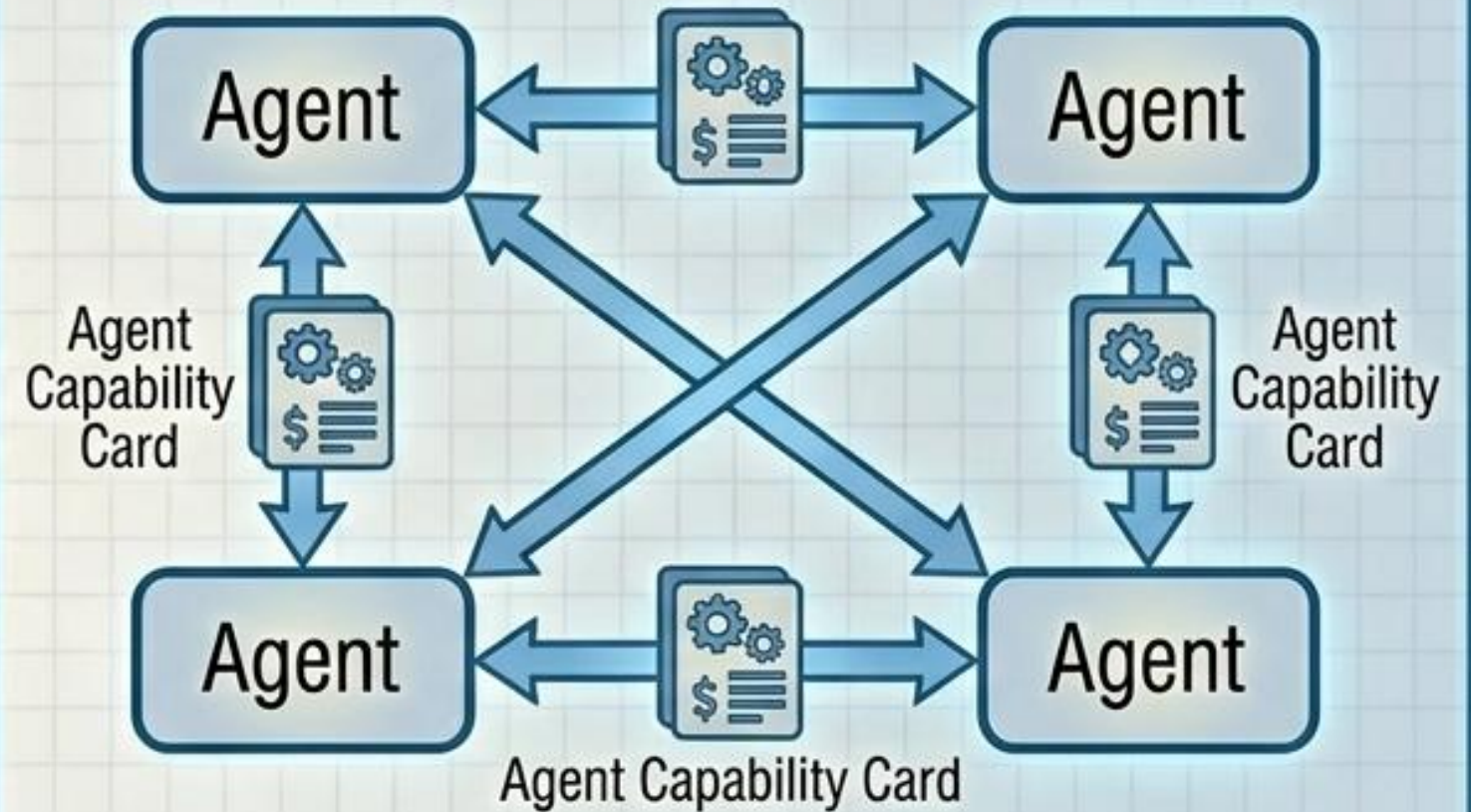
The Technical Substrate: Open Standards

MCP (Model Context Protocol)



Decouples agents from services;
open Linux Foundation standard.

A2A Protocol



Machine-readable declarations of capability; open
standard supported by AWS, Google, Microsoft.

The Research Data Director relies on open standards, not hand-wired integrations.
Tools are discoverable at run time, portable across deployments, and strictly auditable.