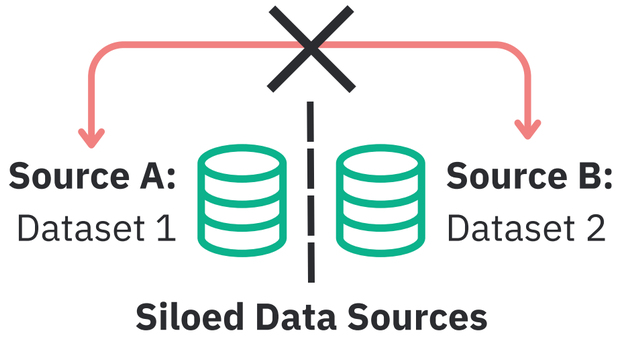
Data Integration and Platform Unification

IMPEDIMENTS AND CHALLENGES IN MARINE BIOINFORMATICS

Jaronchai Dilokkalayakul, *Information Biology Laboratory*,
Graduate School of Information Sciences (GSIS), Tohoku University

THE PROBLEM AND THE CONSEQUENCES

Main Pain Point

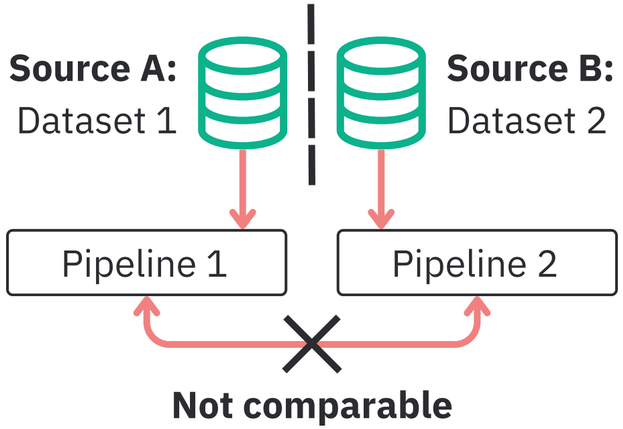


Data are distributed across **separate databases**

Cross-site comparisons and AI/LLM methods are **hard to apply reliably**

Consequences

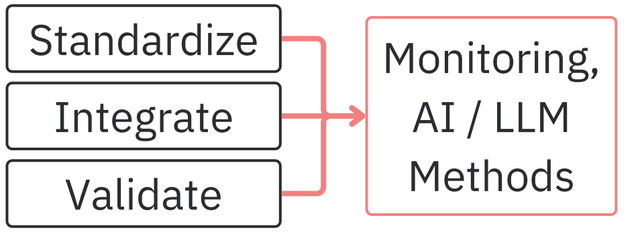
Repeated one-off pipelines and data management



Cross-site trends and anomalies are difficult to assess **consistently**

Addressable Gap

A shared framework across individual and central system



Integrated Platform

THE CHALLENGE WITHIN PLATFORM UNIFICATION

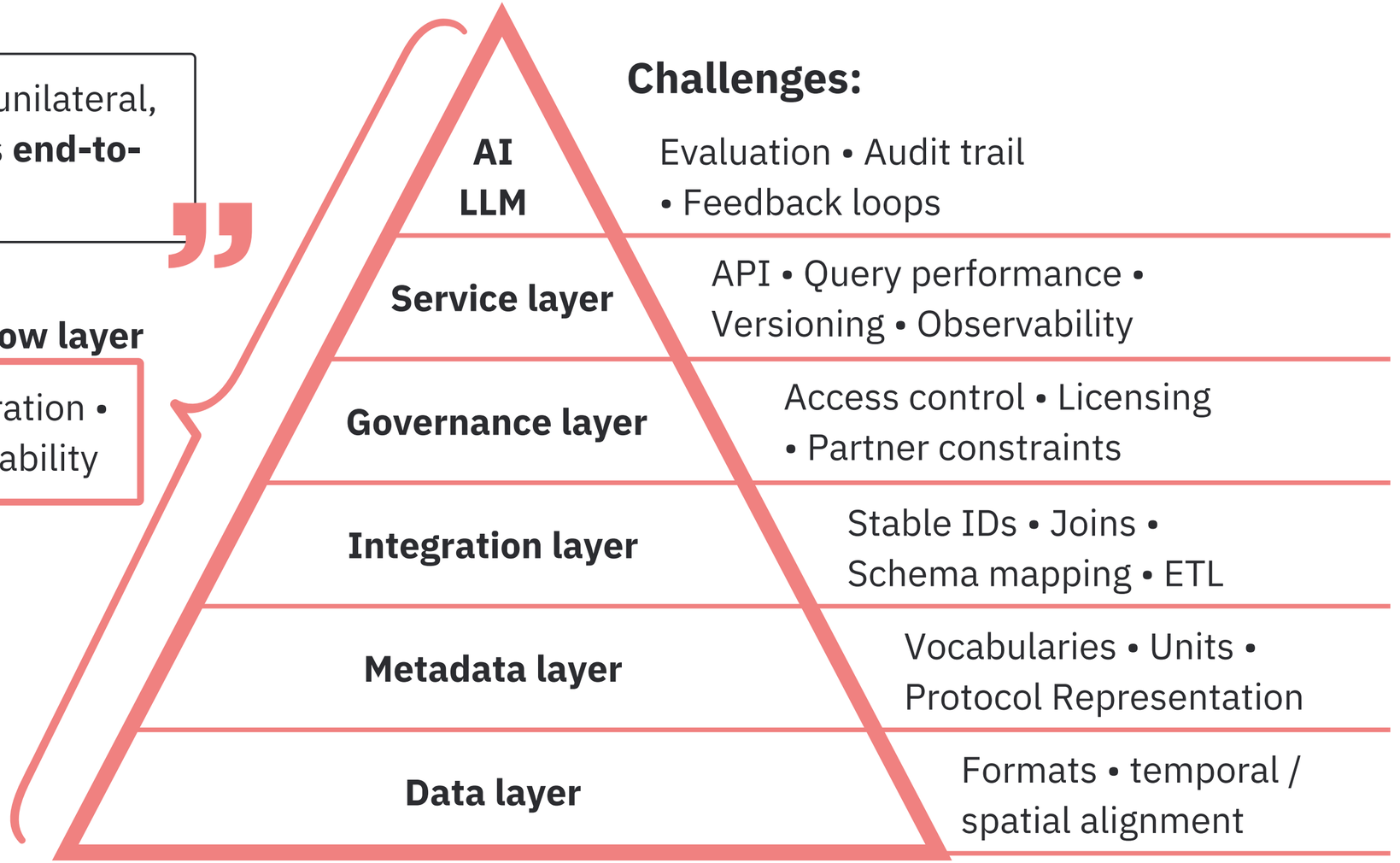
> Failures are rarely unilateral, unification requires **end-to-end** guardrails.

Human Workflow layer

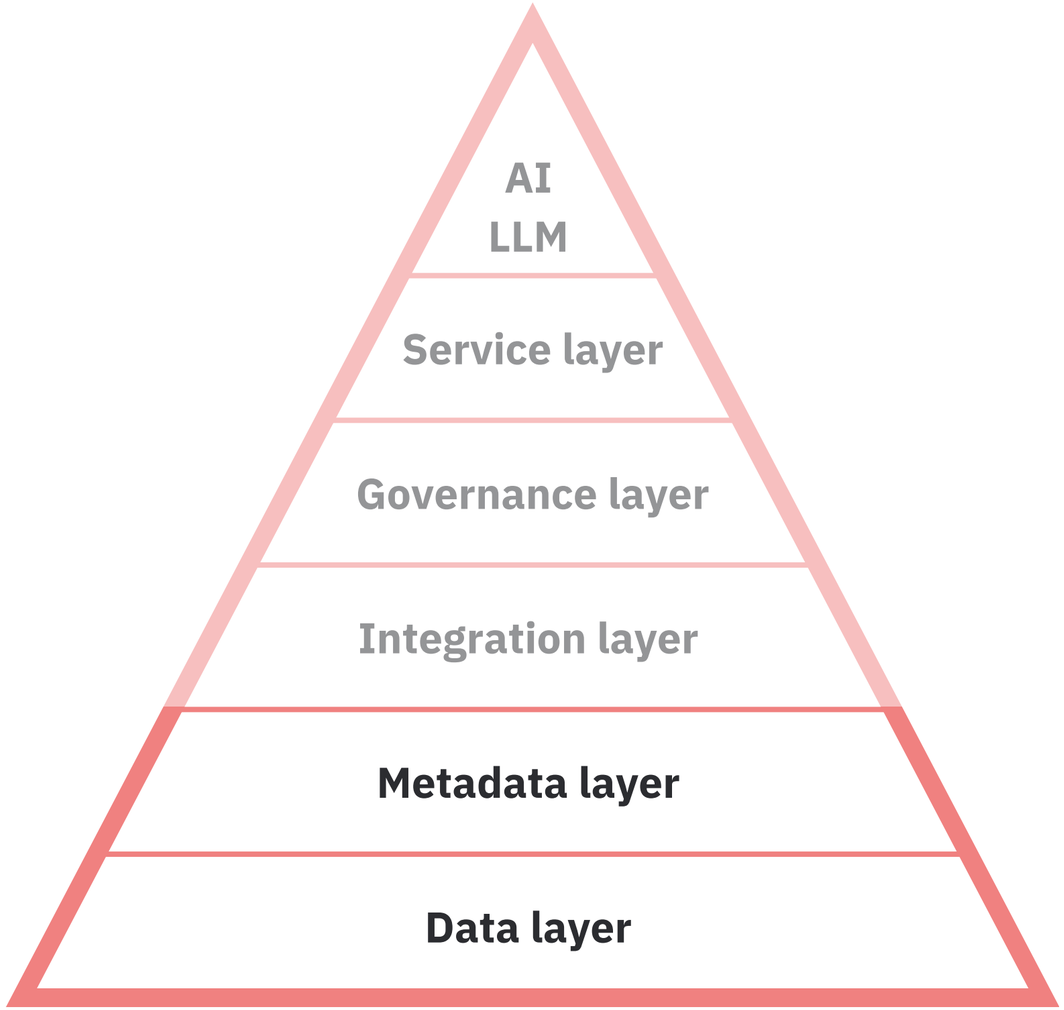
Interpretation • Curation • Handoffs • Accountability

Input Reality

Metagenomics
Oceanography
Satellite
Partner sites



DATA + METADATA = COMPARABILITY



> Without comparable metadata,
> ‘integrated’ data is not comparable

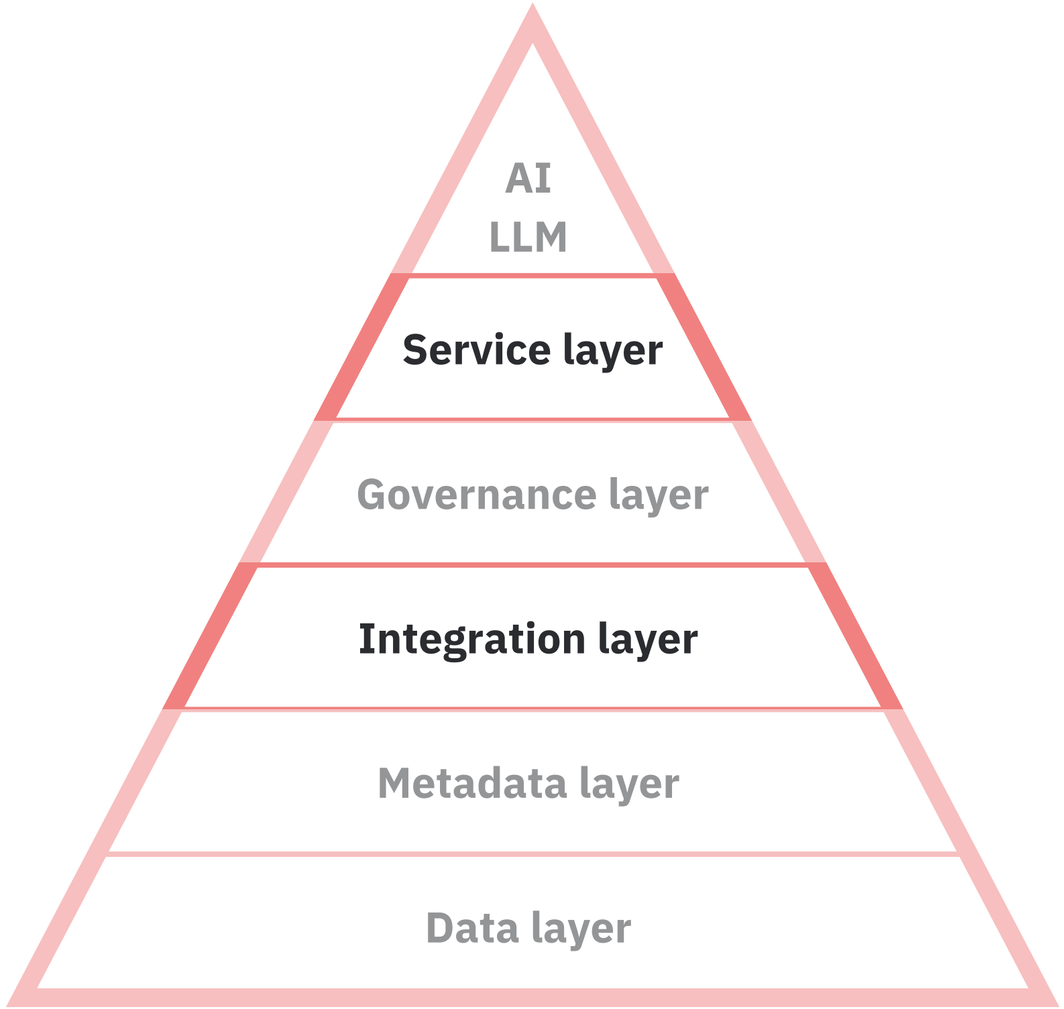
2 Independent Sites

Site A	Site B
Chlorophyll: 2.1 **µg/L**	Chlorophyll: 2.1 **mg/L**
Net mesh: **20** µm	Net mesh: **50** µm
Depth: **0–5** m	Depth: **0–10** m
Preservative: **X**	Preservative: **Y**

≠

Same name ≠ Same measurement

INTEGRATION + SERVICE = SUSTAINABILITY



> Without a framework, pipelines drift and teams revert to ad hoc fixes

Pipeline v1.0

Join key: *station_name*
Schema: *temp_C*
Output: *“Chl trend = ↑”*
API: */chlorophyll?site=A*

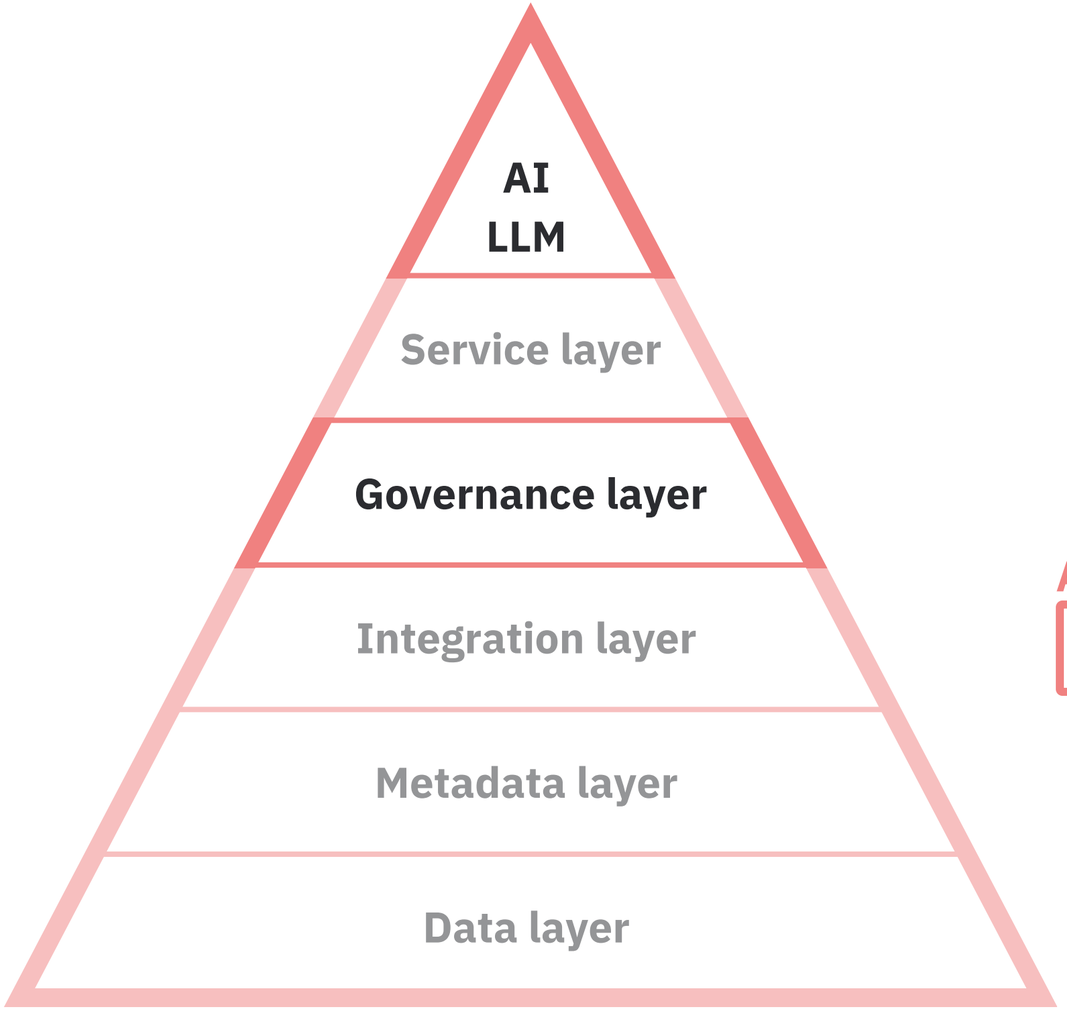
4-6 Months Later

Pipeline v2.0

Join key: *station_id*
Schema: *temperature_C*
Output: *“Chl trend = ↓”*
API: */chlorophyll?site=A*

Same query ≠ same result

GOVERNANCE + AI/LLM = TRUST & ANALYSIS



> AI/LLM is only useful when access, provenance, and evaluation are ready

User Question

Compare chlorophyll trends across the 2 sites

No guardrails

AI Answer:

"Both steadily increase"

Issues:

- No access control (Partner B restricted)
- No citations / audit
- Not reproducible

Guardrails enabled

AI Answer:

A increases; B's data are restricted, summary only."

Evidence: Dataset IDs, Methods, Links, Citations

Audit: Prompt + retrieved records logged

STEP-BY-STEP MITIGATION PLAYBOOK

> Mitigation is cross-layer: standards, interfaces, validation, and audit

STANDARDIZE	INTEGRATE	VALIDATE	SERVE	AUDIT & LEARN
Minimum metadata contract	Stable IDs	QC checks	Versioned APIs	Citations + audit trail
Controlled vocabularies/units	Explicit mappings	Comparability tests	Query contracts	Evaluation loop
Protocol context	Schema evolution rules	Provenance capture	Observability alerts	Feedback without drift

WHAT “SUCCESS CRITERIA” MEANS IN PRACTICE

Success criterion	What “done” means	**Evidence**	**Metric**
Comparable	Same variable has same meaning across sites	Units + protocol context + minimum metadata contract	% harmonized variables • comparability tests pass
Reproducible	Same query returns the same result later	Versioned datasets/APIs + stable IDs	Re-run reproducibility rate • drift incidents detected
Auditable	Answer is traceable and permission-aware	Access control + citations + retrieval/provenance logs	% outputs with citations • audit coverage

Thank you for your attention!

KEY TAKEAWAYS

- Failures are rarely unilateral, unification requires **end-to-end** guardrails.
- **Mitigation strategy:** Standardize → Integrate → Validate → Serve → Learn
- **Outcome:** interoperable platform, cross-site monitoring, trustworthy AI/LLM analysis

Jaronchai Dilokkalayakul, *Information Biology Laboratory*,
Graduate School of Information Sciences (GSIS), Tohoku University

jaronchai.com