Provenance-Preserving Framework for Integrating Heterogeneous Ecosystem Monitoring Data with Large Language Models

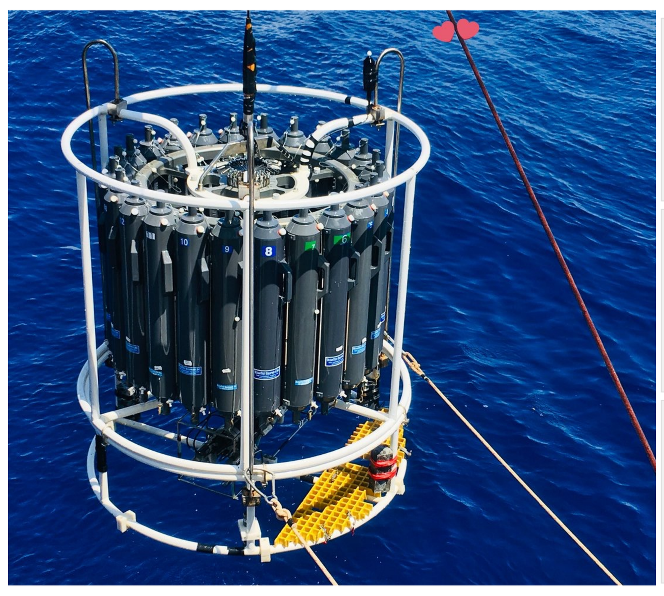
Jaronchai Dilokkalayakul [1], Takeshi Obayashi [1,2]

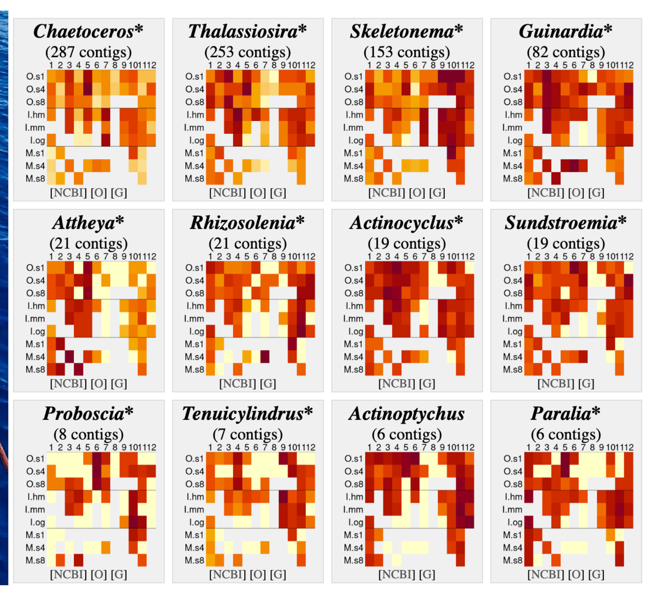
[1] Graduate School of Information Sciences (GSIS), Tohoku University
[2] Advanced Institute for Marine Ecosystem Change (WPI-AIMEC), Tohoku University

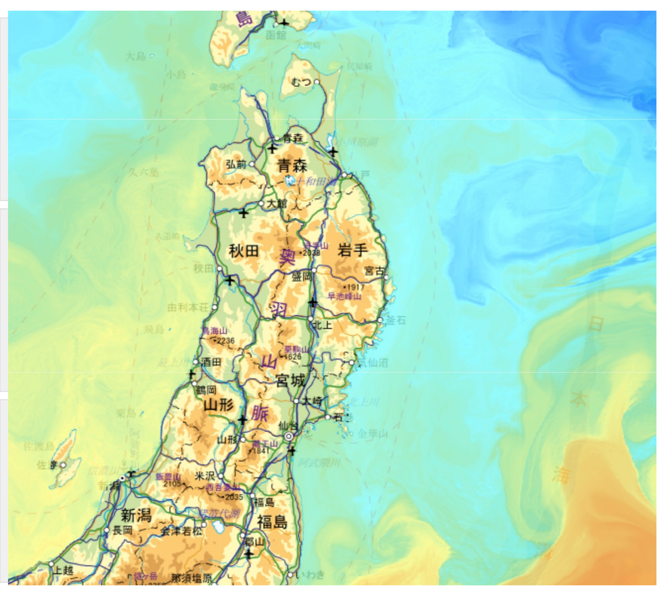
1. The Subtle Complexity of a Simple Research Questions

" What biological observations coincided with warmer surface conditions ? "

- What biological signals were observed? → Metagenome Taxa Tables
- Can each claim be traced? → Provenance Records
- Was the surface warmer? → Satellite SST
- What were the water-column conditions? → CTD profiles



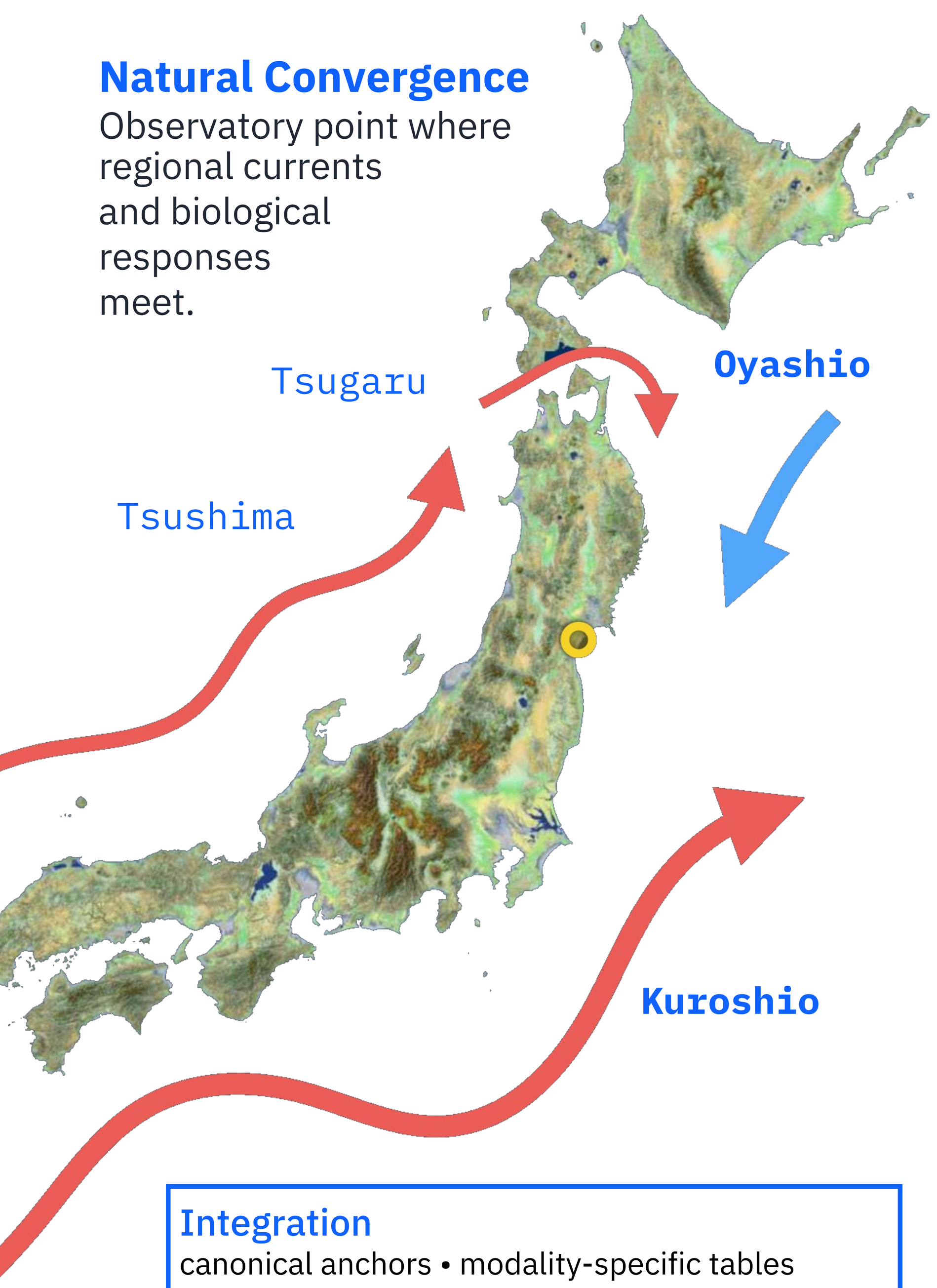




Source	Key Attributes	Integration	Challenges
CTD profiles	Structured water-column observations Temperature, salinity, depth, oxygen, related variables	Sampling date/time Station or bay Depth layer	Depth-dependent measurements; aligning profile time with biological samples.
Metagenome eDNA tables	Taxonomic abundance and sample metadata Genus / species-level biological signal	Sample ID Sampling date Site or bay	High-dimensional sparse taxa; multiple taxonomic levels; relative vs absolute abundance.
Satellite SST Himawari	Gridded sea-surface temperature Spatiotemporal raster / NetCDF subset	Timestamp Latitude/longitude Bay region	Surface-only context; spatial aggregation; matching grid cells to sampling sites.

2. Onagawa Bay as a Pilot Site

Natural Convergence
Observatory point where regional currents and biological responses meet.



- **Integration**
  canonical anchors • modality-specific tables
- **Provenance**
  source lineage • hashes • metadata
- **Retrieval**
  hybrid ranking • structured filters
- **LLM governance**
  citations • caveats • answer constraints

3. Core Framework Idea

A framework that converts heterogeneous ecosystem monitoring outputs data into source-specific, retrievable, citable, and auditable evidence units

4. Each Modality of the Heterogeneous Data Sources Required Specific Preprocessing Methods

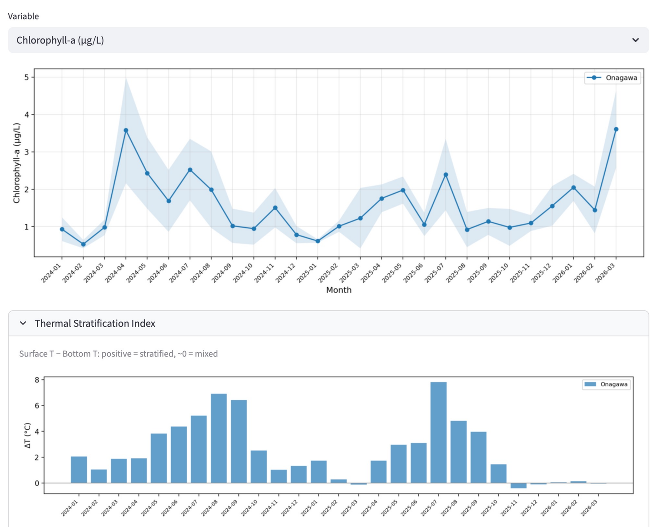
	1. Read source →	2. Standardize IDs →	3. Transform table →	4. Derive context →	5. Write tables →	6. Used by
CTD Water-column profiles	Raw TSV: CTD_Onagawa.tsv Depth-resolved Rows retained	label → sample_id date → ctd_date depth → depth_m	Canonical column names CTD_ALIAS_MAP Numeric coercion	Cast-level statistics: Surface / bottom mean, min, max depth range + point count	ctd_profile ctd_profile_standardized ctd_summary	Anchor events CTD retrieval documents Environmental context
Metagenome Taxonomic evidence	Raw sequencing: Run ID + Summary Kraken / MetaEuk Genus matrices	Map run_id → sample Parse bay / station Parse year_month	Reshape: wide genus × sample matrices → long records Add method labels	QC aggregation Upper taxonomic groups Consistency labels Top-N taxa context	run_qc sample_qc abundance + enriched serving context	Sample registry Metagenome retrieval Taxa analysis
SST Surface thermal context	NetCDF files: recursive *.nc Search in SST Subset directory	Parse timestamp time_utc time_jst	Extract nearest grid cell to Onagawa latitude / longitude Store SST value	Daily bay-scale bounding box: 38–39°N, 141–142°E mean, min, max, stdfile count	sst_point_timeseries sst_daily_summary typed time series	SST retrieval documents Thermal context Reliability checks

5. Pre-Analysis Memory

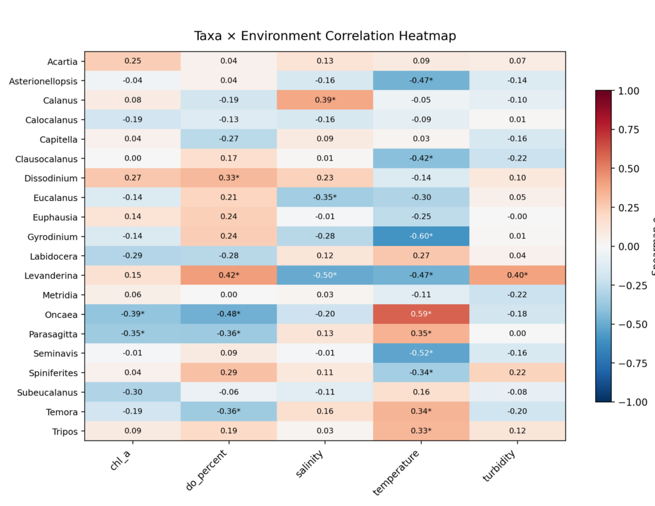
Keys preserved: sample_id · date/time · bay/station · lat/lon · provenance JSONL

Pipeline Hands-off: typed Parquet + serving context → anchor events / links → retrieval documents → PostgreSQL + pgvector

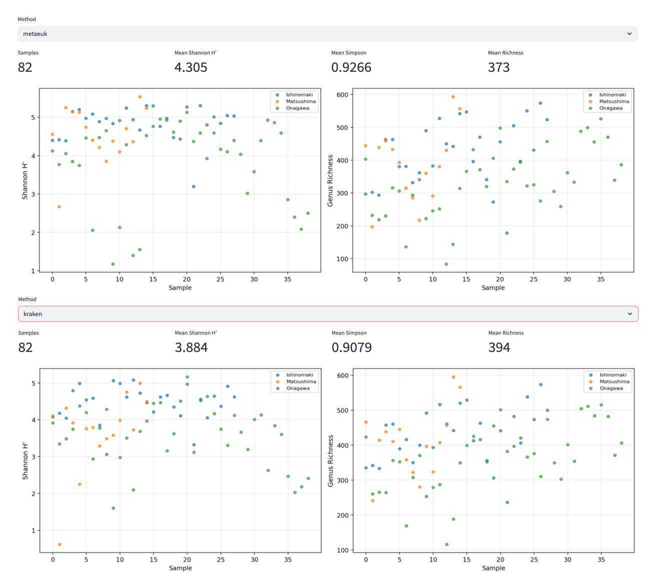
Structured ecological computations are precomputed before the user asks a question, analysis such as:



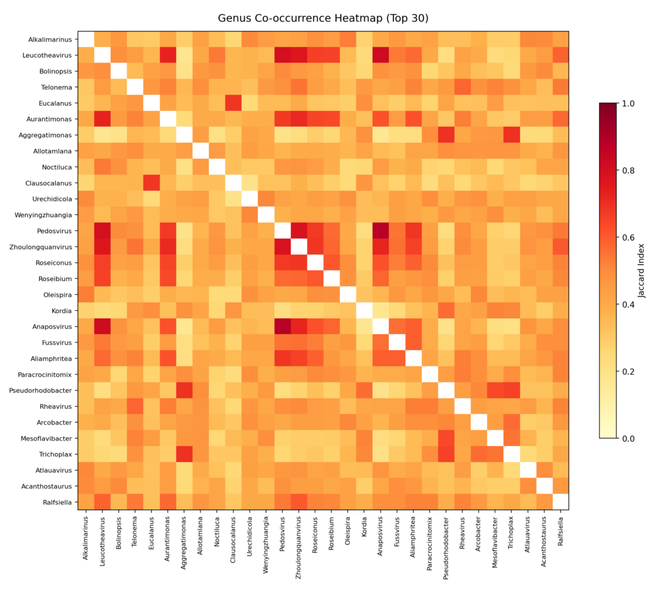
CTD Monthly Trends



Taxa-Environment Corr.



Diversity Indices



Taxa Co-occurrence

6. Data Integration Strategy

- Canonical observation records
- RetrievalDocument schema
- Hybrid retrieval context

Structured metadata

doc_id	citation target
source_type	modality boundary
event_id	anchor link
sample_id	sample link
time/site	spatiotemporal filters
provenance	audit path

Narrative Text

- Observed variables / taxa
- Time, bay, station, depth
- Relevant summary values
- Linked context documents
- Explicit caveats / gaps

The LLM sees readable evidence → the system keeps structure to filter, rank, join, and audit it.

Contact Information

Jaronchai "Potter" Dilokkalayakul
Personal Site: www.jaronchai.com
Email: dilokkalayakul.jaronchai.p8 [at] dc.tohoku.ac.jp

7. Hybrid Evidence Retrieval and Large Language Models

1. **Query controls**
   question + top-K source_type
2. **Backend dispatch**
   unified.retrieve() checks PostgreSQL
3. **Hybrid search**
   Semantic + FTS + SQL filters + RRF fusion
4. **Evidence assembly**
   ranked documents with metadata
5. **Prompt builder**
   evidence + pre-analysis + reliability + citation
6. **LLM Answer**
   returns answer + sources + n_sources

Input : Parsed Query + Filter

Vector Search | Full-Text Search

$$score_{RRF(d)} = \frac{w_v}{k + r_{v(d)}} + \frac{w_f}{k + r_{f(d)}}$$

Top-K Sources

Hybrid retrieval improves evidence selection by combining semantic similarity from pgvector with lexical precision from PostgreSQL full-text search, then fusing both ranked lists with RRF to produce citation-ready sources for the LLM

Expected answer structure

Answer claim →	cite [doc_id]
Modality evidence →	CTD / metagenome / SST separated
Numeric values →	units + precision
Interpretation →	supported, not causal by default
Caveats →	gaps, conflicts, resolution limits

The system enforces source-grounded answering, but the system still needs evaluations and whether injected context improves answers without over-conditioning them

8. User-Facing Interfaces Prototype