

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

JARONCHAI DILOKKALAYAKUL¹, TAKESHI OBAYASHI^{1,2}

¹ Graduate School of Information Sciences, Tohoku University

² Advance Institute for Marine Ecosystem Change (WPI-AIMEC), Tohoku University

Marine ecosystem assessment increasingly depends on the joint interpretation of heterogeneous observations, including in situ water-column environmental measurements obtained by conductivity–temperature–depth (CTD) measurements, metagenomic data, and satellite-derived remote sensing products. However, these modalities differ in structure, spatial and temporal resolution, and semantic meaning, which limits their combined use in exploratory analysis. This work proposes a provenance-preserving framework that integrates heterogeneous ecosystem monitoring data with large language models (LLMs) to support source-grounded scientific question answering. The framework standardizes each modality into a common retrieval-oriented representation while preserving source identity, acquisition context, spatiotemporal metadata, and measurement meaning. CTD-derived measurements provide structured physical profiles of the water column, metagenomic data capture biological and taxonomic signals, and remote sensing products contribute synoptic environmental context. These sources are linked through shared anchors such as time and location, enabling cross-modal retrieval and interpretation. An LLM-based reasoning layer synthesizes responses from the retrieved evidence while maintaining explicit traceability to the original records. The proposed system is intended to improve accessibility to complex marine datasets without collapsing modality-specific distinctions. By preserving provenance throughout ingestion, retrieval, and response generation, the framework supports transparency, interpretability, and reproducibility in LLM-assisted ecosystem data analysis.