

Poster | 33.Poster

Mon. Jun 29, 2026 1:00 PM - 1:50 PM JST | Mon. Jun 29, 2026 4:00 AM - 4:50 AM UTC | Poster 1(4FFoyer+401+402)

[P002] Poster

1:00 PM - 1:50 PM JST | 4:00 AM - 4:50 AM UTC

[P01-040] Provenance-Preserving Large Language Model Question-Answering System for Heterogeneous Biodiversity Monitoring Data○Jaronchai Dilokkalayakul¹ (1. Tohoku University (Japan))

Keywords : Large Language Model、 biodiversity informatics、 heterogeneous monitoring data、 question answering、 retrieval-augmented generation、 provenance、 data interoperability、 ecological information systems

Coastal fisheries and biodiversity monitoring programs generate valuable observations, yet these data are often difficult to use in practice because they are distributed across heterogeneous sources, including databases, spreadsheets, reports, and field notes. Some information is represented in fixed fields, such as sample identifiers, dates, locations, and abundance values. These structured records support efficient filtering, comparison, and computational analysis. Other information is preserved in free-text notes and reports, which capture sampling conditions, ecological context, and observational nuance that are difficult to encode in fixed schemas. These unstructured records are valuable because they preserve meaning and detail, but they are harder to search systematically. When structured and unstructured materials remain fragmented and inconsistently described, even simple ecological questions may require substantial domain knowledge and database expertise. This creates a need for a system that can connect both forms of information while preserving their respective advantages.

We present a provenance-preserving question-answering system for heterogeneous monitoring data. The system combines retrieval-augmented generation (RAG) with a large language model backend to enable source-grounded exploration of records and associated metadata. For each query, relevant metadata and text passages are retrieved and provided to the model, which generates answers with inspectable citations to the original enables users to search by keyword, location, and date while examining supporting evidence directly. By linking structured records with narrative documentation, this approach improves the usability and accessibility of biodiversity monitoring data. This work aligns with current biodiversity informatics efforts to make heterogeneous monitoring data more interoperable, reusable, and accessible to non-specialist users.