MaRESS: Mapping Research in Earth System Sciences

A Modular Web Application for Literature Analysis and Geographic Data Mapping in Earth System Scienes

Benjamin Schmidt[1], Marco Otto[1]

[1]Chair of Climatology, Institute of Ecology, Technische Universität Berlin

Rothenburgstraße 12, 12165 Berlin, Germany

benjamin.schmidt@tu-berlin.de, marco.otto@tu-berlin.de

Project Overview

In the process of conducting systematic literature reviews in Earth System Sciences, researchers often face challenges in efficiently geolocating study sites. While study sites are frequently mentioned in the text, they are rarely included in machine-readable formats or metadata. This gap hinders the ability to perform comprehensive spatial analyses and identify research trends.

MaRESS (Mapping Research in Earth System Sciences) is a modular web application designed to address these critical challenges in Earth System Sciences research data management and analysis.

Core Objectives:

- Automated extraction and mapping of research data from scientific literature
- Identification of geographic knowledge voids, thematic research gaps and citation network
- Extension of bibliographic metadata with spatial analysis capabilities
- AI-assisted categorization and semantic mapping of scientific publications

Target Application: Initial deployment will focuses on the expansive test dataset on High Andean Wetlands research, supporting comprehensive literature review and ecosystem service analysis

Modular System Design

MaRESS is structured in frontend/backend architecture, allowing usage as a standalone web application or integration into existing research data infrastructures.

Each part is separated technically and functionally into four main modules (Figure 2):

Module 1: Geographic Mapping

- Interactive clustered visualization using OpenLayers for Point data
- Coordinate extraction from research papers
- Spatial clustering analysis of study locations

Module 2: Semantic Mapping

- Zotero API integration for bibliographic management
- Citation/Author network analysis using graph algorithms
- Keyword extraction and clustering

Module 3: Research Data Mapping

- Integration with PANGAEA, Zenodo APIs
- DataCite DOI metadata extraction
- Automated dataset-publication linkage

Module 4: AI-Assistant

- Third-party API usage for LLM document categorization

Implementation Workflow

Data Processing Pipeline:

1. **Authentication:**
   Current: Email registration with Bearer Token
   Planned: OAuth2 integration with ORCiD user accounts
2. **Metadata Retrieval:**
   Automated download of bibliographic records and PDFs via *Zotero Web API*
3. **NLP Analysis:** spaCy pipeline processes full-text content for:
   - Geographic entity recognition (GPE, LOC tags)
   - Coordinate pattern extraction using regex
   - Sentence-level geolocation context analysis
4. **Spatial Analysis:** Algorithms implemented for:
   - Coordinate validation and projection
   - Clustering algorithms (DBSCAN) for study site grouping
5. **Visualization:** Real-time rendering of:
   - Interactive maps with study location markers
   - Citation network graphs with node clustering
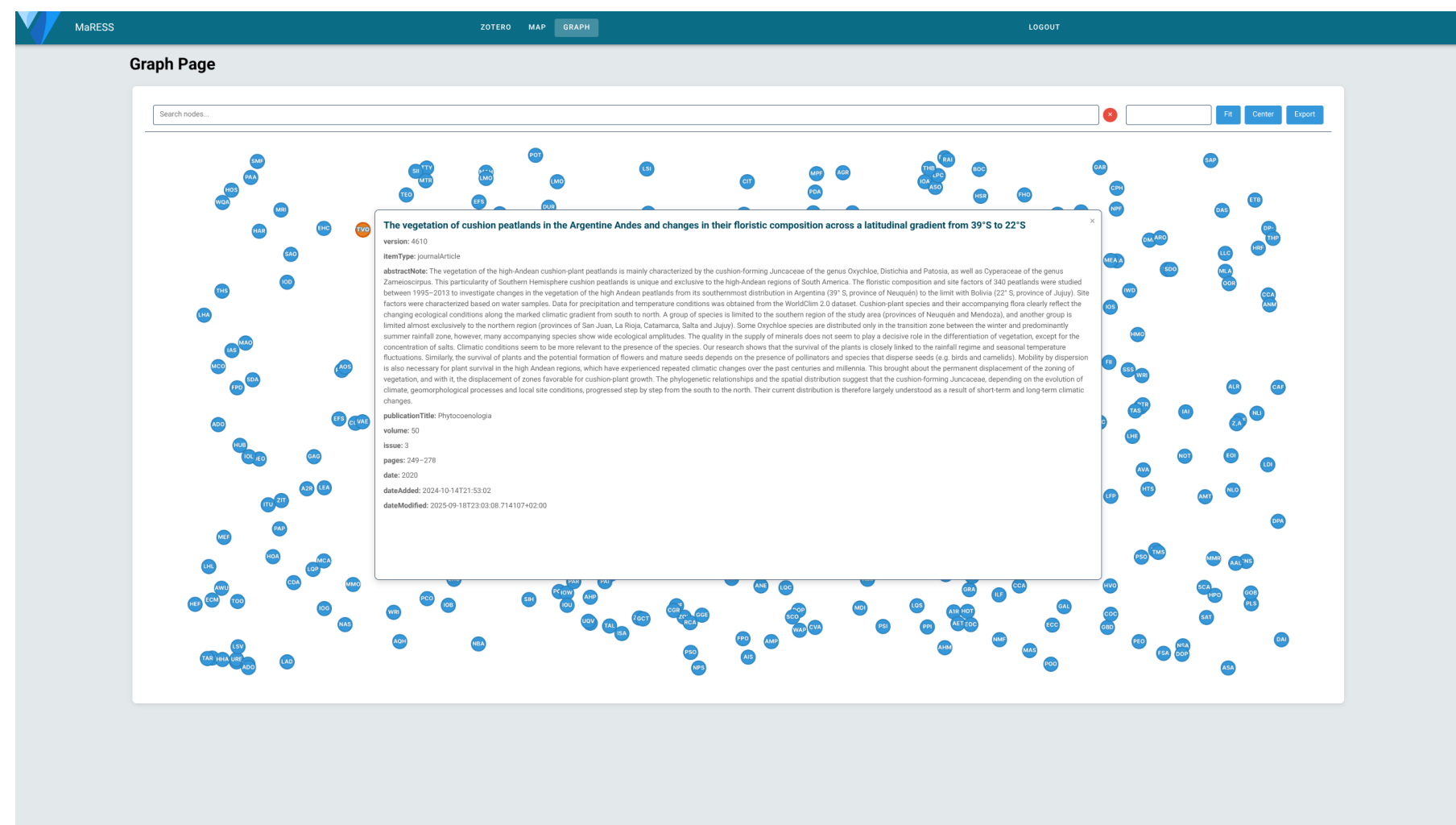   - Metadata dashboards with filtering options

Figure 1: *Interactive citation network and keyword visualization using Cytoscape.js with an overlay of metadata information of selected paper*
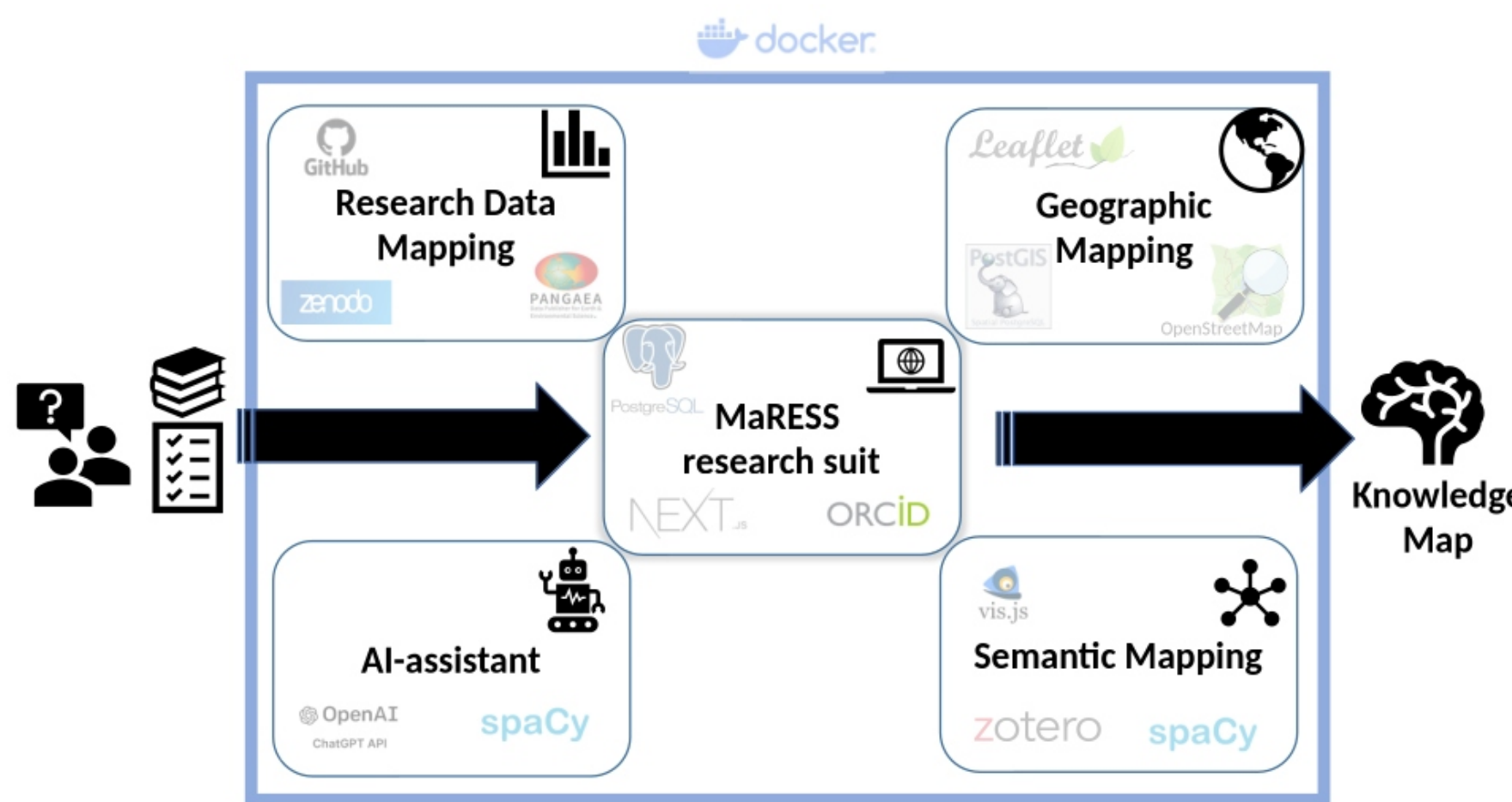

Figure 2: *MaRESS modular architecture integrating four core modules*
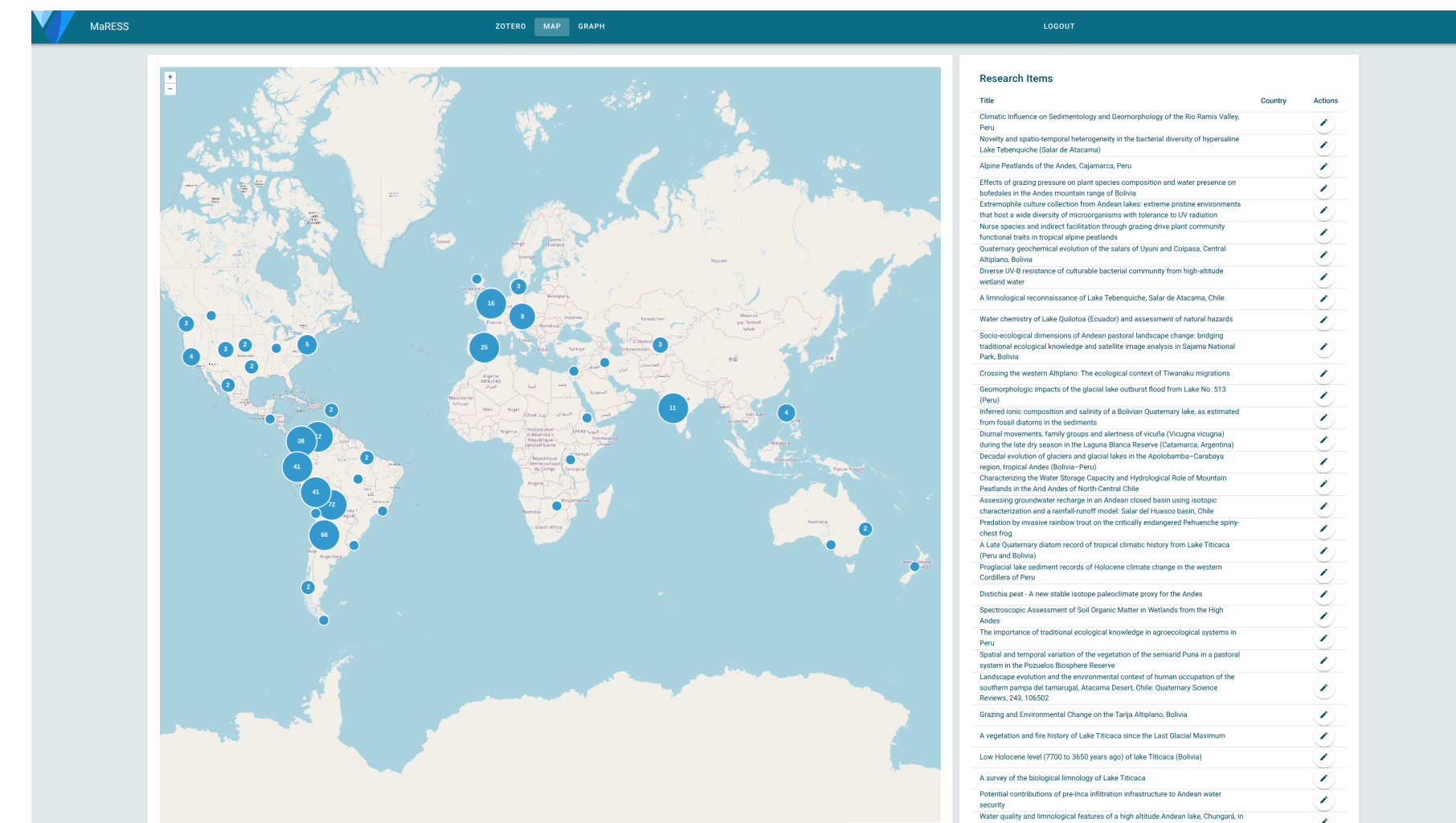

Figure 4: *Interactive map view displaying clustered research study locations extracted from Zotero library via Zotero Web API integration*

Technical Architecture

Backend Infrastructure:

- Python-based with asynchronous request handling
- Relational database with geoinformational extension for spatial data
- NLP model pipeline for Named and Custom Entity recognition
- Zotero Web API, DataCite DOI resolution, Zenodo integration

Frontend Components:

- Modern, reactive JS/TS-based Browser application
- Interactive geographic visualization with OpenStreetMap and clustering capabilities
- Dynamic data tables for metadata filtering and analysis
- Cytoscape.js for keyword and author network visualization

Compliance & Standards:

- Improves FAIR data principles (Findability, Accessibility)
- Open-source licensing (MIT/Apache 2.0)
- RESTful API design with OpenAPI 3.1 specification

Deployment:

Docker containerization ensuring reproducible environments and cross-platform compatibility

Self-hosted or server deployment options

Figure 3: *Metadata table with sorting, filtering, external links and download options*

Technical Specifications

Component	Technology Stack
Backend API	FastAPI 0.116+ with Pydantic v2
Database	PostgreSQL 17+ with PostGIS 3.3+
ORM	SQLAlchemy 2.0+ with asyncio support
Authentication	OAuth2 + JWT tokens
NLP Processing	spaCy 3.8+ with en_core_web_lg model
Frontend Framework	Vue.js 3.5+ with TypeScript
UI Components	Vuetify 3.10+ material design
Mapping Library	OpenLayers 10.6+ with vector layers
Graph Visualization	Cytoscape.js 3.33+

Current Status & Future Development

Development Timeline:

- **July-September:**
  MaRESS research suite development with User handling
  Zotero API integration for metadata retrieval
  Module 1 (Geographic Mapping) development
  Module 2 (Semantic Mapping) development
- **October:**
  Algorithm accuracy evaluation and improvements
  Missing feature implementation for Modules 1-2
  Implementation of Module 3 (Research Data Mapping)
- **November-December:**
  Final testing and bug fixing
  Module 4 (AI-Assistant) implementation

Key Features Implemented:

- User authentication and management system
- PostgreSQL spatial database with coordinate storage
- Automated PDF data processing pipeline (Implementation Workflow)
- Cytoscape graph visualization (Figure 1)
- Interactive clustered map visualization (Figure 4)
- Metadata table with filtering and export options (Figure 3)