Mapping **R**esearch in **E**arth **S**ystem **S**ciences (**MaRESS**): Developing an Application for Data Integration and AI-assisted Information Management

Dr. Marco Otto
Institute of Ecology (Chair of Climatology)
Technische Universität Berlin (TU Berlin)
Rothenburgstaße 12, 12165 Berlin
https://www.tu.berlin/klima

10.11.2024

Abstract

The project aims to develop a web application, **MaRESS** (**Ma**pping **R**esearch in **E**arth **S**ystem **S**ciences), designed to map research data from peer-reviewed literature to help researchers identify thematic research gaps and geographic knowledge voids. MaRESS will support researchers in formulating targeted questions and objectives within Earth System Sciences (ESS), advancing scientific understanding across ESS by providing a structured framework to build specific knowledge bases. Using a modular design, MaRESS will integrate geographic data ("geographic mapping"), an open-access reference management tool for knowledge organization ("semantic mapping"), support for data integration ("data mapping"), and AI-assisted categorization. These components will enhance data accessibility and information management for all research areas within ESS. The software will be portable and deployable using containerization (e.g., Docker or LXC) and will include comprehensive documentation, supporting FAIR data principles and facilitating open access. Initially, MaRESS will be applied to an existing knowledge base on High Mountain Wetlands, with potential for global expansion as additional regions and datasets are incorporated.

I. Introduction

The expected output is a web application, **MaRESS** (**Ma**pping **R**esearch in **E**arth **S**ystem **S**ciences), with capabilities including interactive mapping, open-access reference management, research data integration, and AI-based categorization features to streamline data entry (Fig. 1). Research in Earth System Sciences (ESS) often involves data from diverse disciplines and geographically dispersed ecosystems (e.g., High Mountain Wetlands). Organizing and integrating this scattered information poses significant challenges, particularly when working within specific ecosystems. MaRESS addresses this need by providing a web-based, AI-assisted mapping tool that allows researchers to organize and analyze data within the ESS domain into specific knowledge bases. Aligned with NFDI4Earth's mission, MaRESS will enhance data accessibility and promote Open Science and FAIR principles, offering a community-driven, innovative approach to research data management.
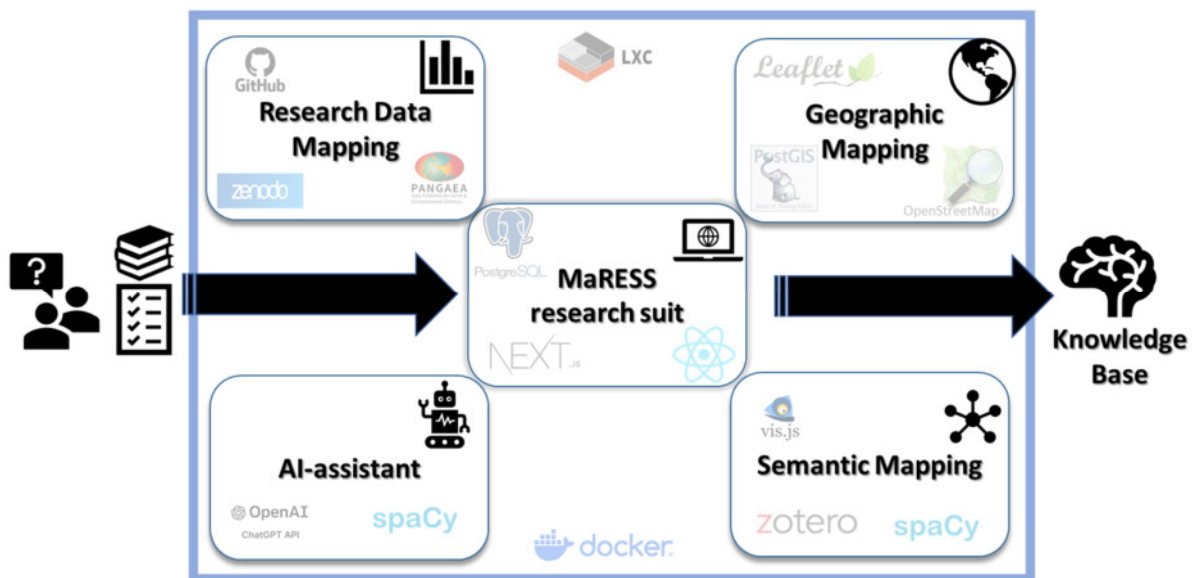
Figure 1: Modular design of the MaRESS research suite, integrating four key modules—Research Data Mapping, Geographic Mapping, Semantic Mapping, and AI-Assistant. Each module addresses distinct aspects of data management and analysis in Earth System Sciences (ESS) to build a comprehensive knowledge base from initial user input. Semitransparent annotations indicate selected open-source tools, libraries, and open-access platforms integrated into or connected with MaRESS.

II. Incubator Project description

The software development of MaRESS is mainly carried out by **Dr. Benjamin Schmidt** who has experience in software development. **Dr. Marco Otto** (TU Berlin) will oversee the project and bring his expertise in climatology and landscape ecology to ensure scientific rigor and relevance. The Chair of Climatology has extensive experience in implementing projects in the field of web development (e.g. https://uco.berlin/en) as well as the necessary infrastructure. A productive instance of MaRESS will serve as a functional prototype for research on ecosystem services in High Andean wetlands, forming the 'Mountain Wetland Repository for the Andes.' This prototype will support a comprehensive literature review compiled and published by a research team led by Dr. Marco Otto at TU Berlin.

Proposed Solution along the following work packages (WP in Table 1)
WP 1: Project Setup
- Decide on portable (containerization) software solution, tools and frameworks
- Set up software repository and project management
- Final team-setup and scheduling

WP 2: Core module development of MaRESS
- Set up a development server and database
- Develop core the module as a basis for all other modules

WP 3: Semantic mapping
- Integrate Zotero's API for bibliographic data management based on existing Zotero-library on High Andean Wetlands (> 600 peer-reviewed publications).
- Integration of citation network analysis and natural language processing

WP 4: Geographic mapping
- Develop an interactive map of the Andes with points referencing research locations.
- Use GIS data to allow users to navigate the map, select specific study locations, and access associated publications.

WP 5: AI-Assistant
- Categorization System through implementation of an AI feature (potentially using OpenAI's ChatGPT API or another cost-effective solution) to analyze document text and suggest geographic tags for new publications, aiding in location-specific categorization.

WP 6: Research data mapping
- Integration through the use of DataCite DOIs, API's (zenodo) or metadata files (PANGEA)

WP 7: Testing and Documentation
- Test the core module
- Test module integration to ensure usability, data categorization accuracy, and overall performance of the MaRESS Research Suite as a comprehensive application.
- Create comprehensive user documentation, including tutorials, and publish on GitHub/GitLab for open-source access
- Release final candidate as a containerized software package (Docker and LXC) for open access and public download.

Table 1: *Project timeline detailing the seven work packages (WP) across project months 1 to 6*

WP	month 1	month 2	month 3	month 4	month 5	month 6
1	■					
2	■	■	■			
3			■	■	■	
4			■	■	■	
5			■	■	■	
6				■	■	
7			■		■	■

Quality and Transferability Measures

- **Quality Standards**: The project will follow best practices in web development and data management, ensuring the app is user-friendly, efficient, and secure.
- **Transferability**: MaRESS features a modular design integrating open-source solutions for flexibility and adaptability. It utilizes existing open tools, e.g., Zotero for semantic mapping, Zenodo's open APIs for research data, and PostgreSQL with PostGIS for geographic mapping. Containerized deployment (e.g., Docker, LXC) ensures portability across server environments, making MaRESS easily adaptable to diverse ecosystems and research domains, aligning with NFDI4Earth's goals for interoperable, community-driven tools.

III. *Relevance for the NFDI4Earth*

MaRESS enhances data accessibility and aids ESS researchers, data curators, and educators in identifying knowledge gaps. Aligned with NFDI4Earth's mission, it introduces tools for data mapping, metadata integration, and user consultation to support thematic and geographic research. MaRESS will contribute to NFDI4Earth's living document on research trends and gaps, guiding future research and development. Its interactive mapping and AI-driven semantic capabilities fulfill NFDI4Earth's goals for metadata extraction, annotation, and semantic mapping, supporting user-driven exploration in research data management.

IV. *Deliverables*

A source code repository containing setup instructions and containers will be made publicly available including documentation, user guides, installation procedures, and example data for easy setup. Prototype: https://andes.mountain-wetlands-repository.info

V. *Finance plan*

The primary costs are allocated for **Dr. Benjamin Schmidt**, who will share the personnel expenses equally for software development equivalent to a full-time position (FTE) with a total cost of **24,300 € over 6 months** (4 PM 100%). Server hosting, development environment, and user documentation support for MaRESS will be provided by the Chair of Climatology at Technische Universität Berlin (Dr. Marco Otto).

VI. *VI. Änderungen nach Absprache mit NFDI4Earth:*

Herr Schmidt hat bei uns das Datenportal https://uco.berlin/en maßgeblich mitprogrammiert. Er ist derzeit auch in einer Phase zwischen abgeschlossener Promotion und neuem Arbeitsverhältnis (wahrscheinlich auch bei uns ab 1.10.). Kurz: Er hat die Zeit, die Qualifikation und das Interesse zusammen mit mir dann diesen Inkubator wie folgt umzusetzen:

- Herr Schmidt kann vom 1.06. – 30.09. hauptamtlich an dem Projekt über die TU Berlin arbeiten.
- Die Projektzeit läuft mit der Fertigstellung eines Prototypen zum Ende diese Jahres 2025 aus, sodass ab Oktober noch Zeit ist es kostenneutral abzuschließen.
- Die meisten Mittel sind Personalausgaben (Vertrag Herr Schmidt entspricht nach jetziger Schätzung etwas bei 95 % TVL 13 Stufe 3-4 Arbeitgeberbrutto bei 24 k€ über 4 Monate)
- Da das ursprüngliche Budget deutlich geringer ist, müssen folgende WPs im Umfang gekürzt werden:
  - WP 5: (KI) Reduktion auf eine KI ohne umfangreiche Testung anderer AI-Modelle (Entscheidung über Recherche),
  - WP 5: Reduktion des „Geo-Taggings" nur aus Text oder Tabellendaten (keine Kartendarstellungen oder Bilder) mittels AI-gestützte Extraktion der Geoinformation aus den wissenschaftlichen Artikeln.
  - WP 6 Reduktion auf einen Datensatz als beispielhafte Umsetzung (z.B. Zeitreihendaten mit CF-Standard, DWD-Daten oder UCO.Berlin-Daten)
  - Auslieferung des Inkubator-Prototypen nur als Docker-Compose (inkl. Dokumentation) mit Anwendungsbeispiel wie beschrieben im Antrag (Artikelsammlung: https://andes.mountain-wetlands-repository.info)

Im Großen und Ganzen werden wir versuchen, alle beantragten Funktionalitäten des Inkubators mit gegebenen finanziellen Rahmen von 24 000 € Personalkosten zu entwickeln, jedoch eher mit exemplarischen Character und weniger über die Masse unterschiedlicher Datenformate, Produkte etc. Als Sachmittel bräuchten wir allerdings noch **300 €** für die Nutzung einer API basierten KI (Beispielrechung basiert auf OpenAI API: 2x800 wissenschaftliche Artikel (GPT-4 Turbo) + Testing). Entwicklung, Testen und Onlinestellen des Inkubators nach abgeschossener Entwicklung erfolgt auf unserer hier am Fachgebiet Klimatologie verfügbaren Infrastruktur und frei zugänglichen Repositorien.

Die Gesamtkosten liegen somit bei **24.300 €** aufgeschlüsselt nach:
Personalmittel: **24 000 €**
Sachmittel: **300 €**
Projektdauer: **01.06.-31.12.2025** (Personalausgaben: 1.6-30.09.2025)