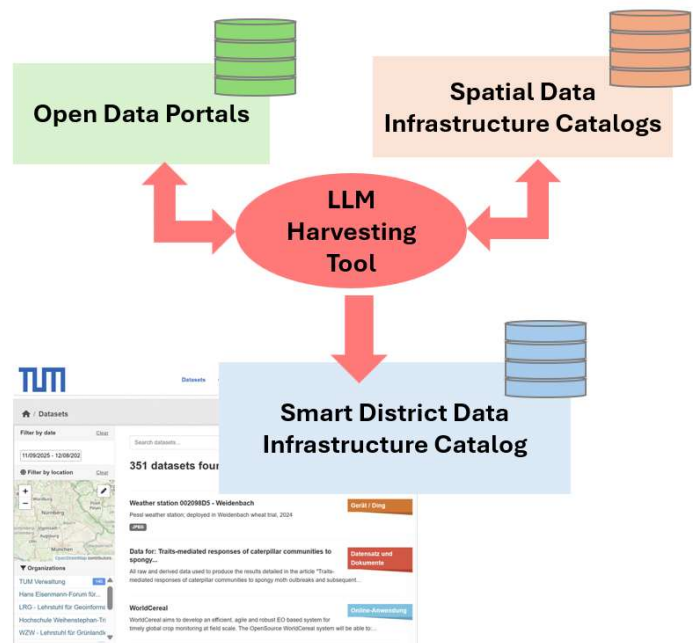


Proposed topic for Master's thesis

LLM-Based Metadata Extraction and Categorization for Automated Data Catalog Harvesting

Efficient metadata harvesting is a fundamental component of metadata infrastructures, particularly within distributed environments where datasets must be discoverable, accessible, and semantically well-organized across diverse data providers. Open Data Portals (ODPs) and Spatial Data Infrastructure (SDI) catalog systems provide access to large amounts of metadata, but these entries often vary significantly in structure, completeness, and categorization quality. Such heterogeneity makes it difficult to integrate datasets, reduces interoperability, and is a challenge for users who are looking for datasets on specific topics. Recent progress in Large Language Models (LLMs) such as *ChatGPT*, *Llama*, *Mistral* or *DeepSeek* enables new possibilities for semantic metadata extraction, topic classification, metadata quality assessment, and catalog enrichment. Unlike traditional approaches, LLMs can for example identify thematic domains and propose standardized categories.



This master's thesis will investigate how LLM-based methods can support the automated harvesting, extraction, categorization, and enrichment of metadata from ODP and SDI catalogs, aligned with standards such as ISO 19115 and DCAT, into the Smart District Data Infrastructure (SDDI) Metadata Catalog. The research will focus on analyzing heterogeneous external metadata sources, designing a scalable harvesting pipeline, and developing AI-driven methods to interpret and transform harvested metadata into the Smart District Data Infrastructure (SDDI)¹ framework. The thesis will explore how interactive user input within the harvesting process can be combined with LLM reasoning to ensure domain-specific alignment with SDDI requirements. This approach aims to combine automation and expert control to create high-quality, semantically consistent metadata that are suitable for integration into the SDDI ecosystem. The outcome will be a structured and extensible workflow for integrating diverse external metadata sources into SDDI using AI-supported methods.

Supervisors Marija Knezevic, Dr. Andreas Donaubaueer,
Office 0107, 0122,
Phone +49 89 289 22974, +49 89 289 22532
E-mail marija.knezevic@tum.de, andreas.donaubaueer@tum.de

¹ <https://www.asg.ed.tum.de/en/gis/projects/smart-district-data-infrastructure/>