

LLM-Assisted Variable Annotation using the I-ADOPT Framework

Arvin Rastegar¹, Barbara Magagna² and Christof Lorenz¹

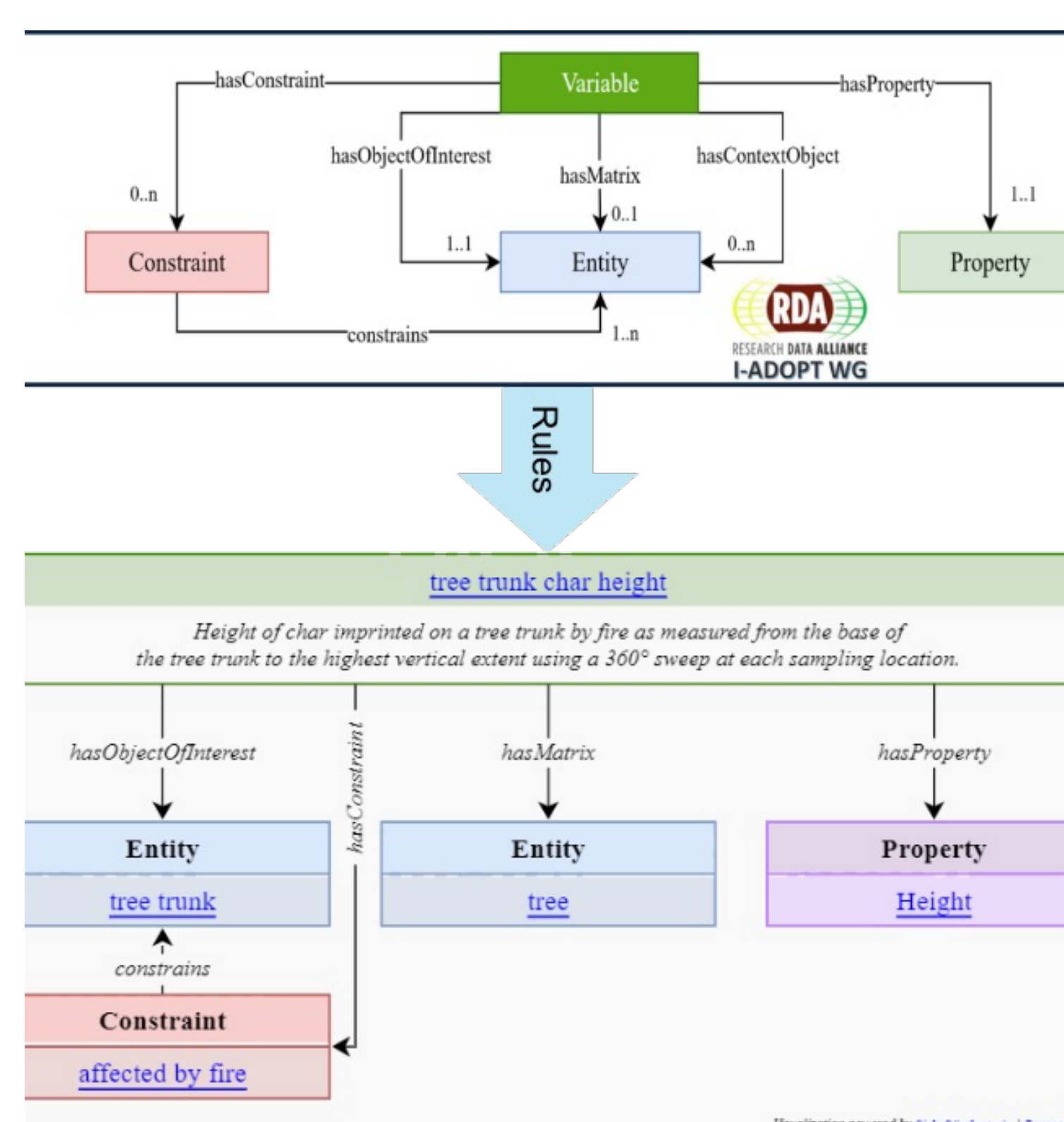
1: Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research – Atmospheric Environmental Research (IMK-IFU), Garmisch-Partenkirchen, Germany
2: GO FAIR Foundation, Leiden, The Netherlands

Automating Semantic Variable Annotation for Environmental Data Interoperability

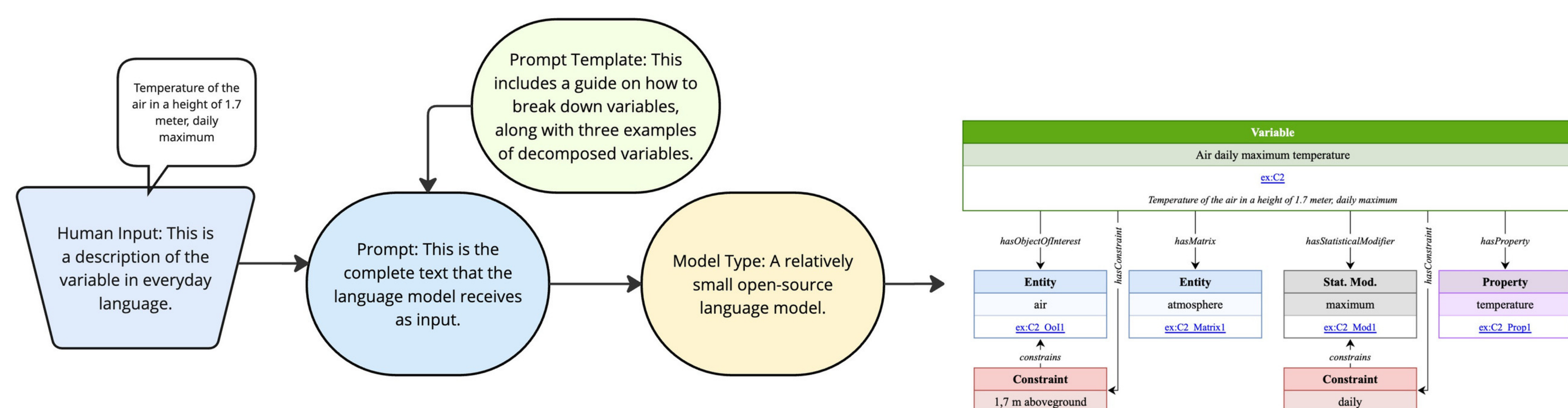
Within our NFDI4Earth-Pilot, we are jointly developing an LLM-assisted variable annotation service that leverages recent advances in Large Language Models (LLMs) to automate the semantic decomposition of variable descriptions. Our approach employs the community-driven I-ADOPT framework to break down natural-language variable definitions into essential atomic elements, ensuring naming consistency and interoperability across domains. The system incorporates retrieval-augmented generation (RAG) to access relevant literature and controlled vocabularies, enabling more precise annotations and reducing manual effort for data producers. By aligning AI-driven methods with established semantic standards, our work addresses several focus areas in Earth System Sciences — including Foundation Models & LLMs, metadata management, and data workflows — while also supporting the broader objectives of NFDI and higher-level initiatives like the EOSC. This approach, hence, enhances reproducibility, interoperability, and cross-disciplinary collaboration in day-to-day research data stewardship.

I-ADOPT and why an LLM based service plays a crucial role for its use

Consolidation of Semantics in Metadata **LLM-enabled I-ADOPT: Extraction of FAIR Variable Descriptions**

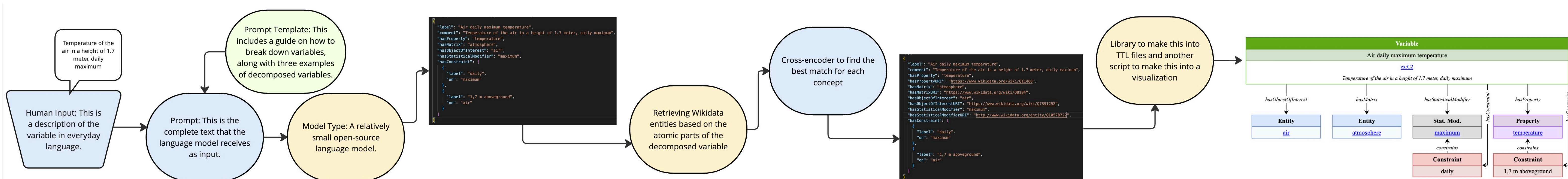


The I-ADOPT framework provides the semantic foundation for mapping and consolidating controlled vocabularies by decomposing observable properties into essential atomic parts, ensuring naming consistency and interoperability.



The LLM-enabled workflow decomposes human-readable variable descriptions (e.g., “Temperature of the air at 1.7 m; daily maximum”) into atomic I-ADOPT elements. The model maps each variable into structured I-ADOPT components, using a simple prompt template with reference examples and no further context. Using this workflow, we have evaluated the performance of 30 different LLMs for identifying the most suitable model for our application. When comparing the decomposed variables with our chosen ground truth, this simple workflow already showed a performance of $F1 = 0.55$.

LLM-Powered Mapping of Variables to Wikidata Concepts



In the next phase, we enhanced the previous workflow with API calls for querying Wikidata. With this extension, we want to retrieve candidate concepts for reference definitions (Incl. URLs, PIDs, etc.) of the decomposed elements. A cross-encoder then ranks the candidates, and the best matches are linked into the structured, machine-readable variable representation. This approach achieves strong performance ($F1 = 0.815$, Precision = 0.709). This enhancement is a crucial step for ensuring consistency of our decomposed variables with the I-ADOPT-framework.

All of these experiments are executed in Jupyter notebooks using standard libraries for easy reuse and collaboration.