LLM-enabled I-ADOPT Variable Extraction using Semantics

Authors: Barbara Magagna (GO FAIR Foundation, Leiden, Netherlands), Arvin Rastegar, Christof Lorenz, Christian Chwala (Karlsruhe Institute of Technology – Institute for Meteorology and Climate Research, Garmisch-Partenkirchen, Germany)

Date of submission: 31.10.2024

Required funding: 6 months

Abstract

Researchers annotate data with keywords for describing the physical properties that are observed or modeled. For ensuring findability and interoperability of this metadata, the keywords should be machine-readable and adhere to standardized vocabularies or ontologies. The I-ADOPT framework provides guidelines for expressing such keywords in alignment with the FAIR principles; however, transforming commonly used terms into atomic I-ADOPT components remains a highly manual task requiring both semantic and domain expertise. In response, we propose an LLM-based workflow to generate FAIR-compliant descriptions of variables that align with the I-ADOPT Framework.

I. *Introduction*

The fulfillment of the FAIR principles[1] is a central element of modern research. But particularly data findability and reusability highly depend on the quality and interoperability of their metadata. While it is widely recognized that metadata should extend beyond loosely defined keywords to specify domain-relevant concepts and adhere to community standards, achieving consistent and uniquely referenceable naming of geoscientific variables still remains a significant challenge. And according to the FAIR principles, metadata should be machine-interpretable using semantic annotations. But in practice, the terminologies used to describe datasets and observed variables vary a lot in their granularity, quality, governance and interconnectivity which, in turn, limits their interoperability. The Research Data Alliance (RDA)-endorsed I-ADOPT Framework[2] addresses this issue by breaking down descriptions of observed variables into five well-defined atomic components *ObjectOfInterest*, *Property*, *Matrix*, *Constraint* and *Context* (see Fig. 1), anticipating their annotation with generic terms from FAIR semantic artefacts. The potential of this framework has also been recognized by the Open Geospatial Consortium (OGC), that currently evaluates the integration of I-ADOPT as official extension to the Observation, Measurement, and Sample Standard[3] and the SensorThings API[4]. But, as of today, the I-ADOPT decomposition is still a highly manual process that requires semantic and domain knowledge.

[1] https://doi.org/10.1038/sdata.2016.18

[2] Magagna, B. et al. (2022): InteroperAble Descriptions of Observable Property Terminologies (I-ADOPT) WG Outputs and Recommendations (1.1.0), Zenodo, https://doi.org/10.15497/RDA00071

[3] https://www.ogc.org/publications/standard/om/

[4] https://www.ogc.org/de/publications/standard/sensorthings/

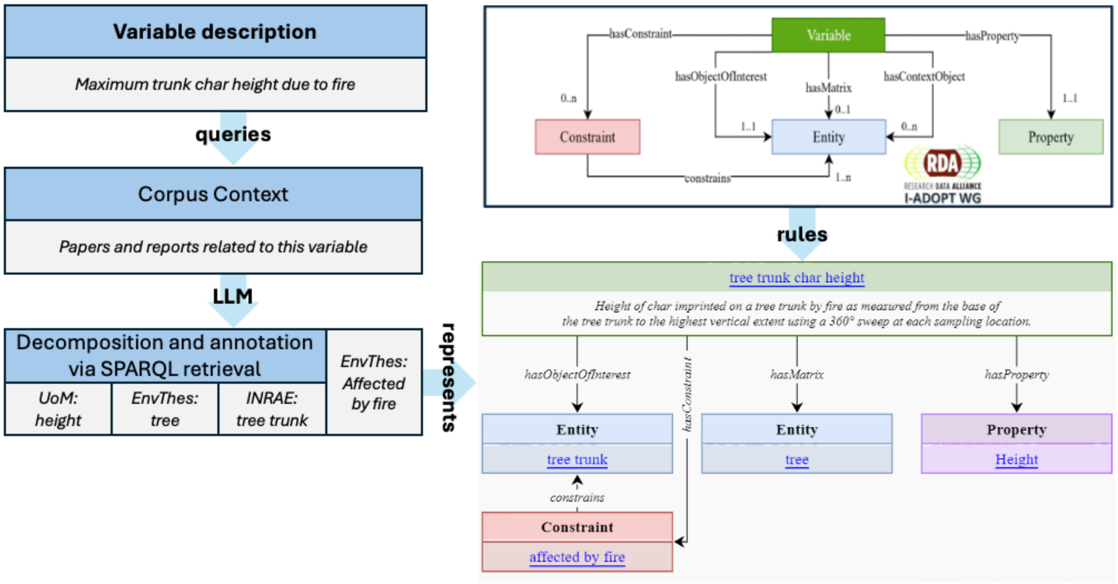

Figure 1: Schematic overview of the LLM-adaptation and the transformation of keywords and variable descriptions into atomic I-ADOPT-components.

We, hence, propose to augment a Large Language Model (LLM) for facilitating the transformation of scientific terms into I-ADOPT-aligned descriptions. This approach will ultimately allow domain experts to generate machine-interpretable representations directly from natural language descriptions of observational research. For developing this model and a first demonstrator, we will build on our previous experience in developing the I-ADOPT Framework, in transfer learning and fine-tuning neural networks, FAIR data stewardship, research data infrastructures and research software engineering. Our project will be further linked to several other ongoing activities and initiatives both on a national and also European level, which allows us to directly evaluate the performance of our LLM by potential end-users and communities.

II. *Incubator Project description*

We contextualize an existing LLM to automate the creation of I-ADOPT-aligned variable descriptions. This will enable researchers to generate machine-readable and human-understandable descriptions, adhering to the FAIR principles and reducing the manual workload involved in minting new variables. Developing a demonstrator for such a tool requires four subsequent tasks:

1. **Analysis of LLMs Capabilities (1 month)**: Evaluate different LLMs current ability to process scientific terms, focusing on its performance to break down variables into atomic components according to the I-ADOPT framework.
2. **Data Collection and Preprocessing (1.5 months)**: Leverage Graph RAG's (Graph Retrieval-Augmented Generation) to retrieve relevant literature and resources aligned with each variable - this step will enable the model to consider the user's intent within the query, providing essential context for the LLMM; retrieve terms from interoperable vocabularies to annotate atomic components in variable descriptions.
3. **Model Adaptation, Training and Validation (3 month)**: Fine-tune the LLM using the collected datasets enriched with controlled vocabularies for enhancing the model's ability to

decompose scientific terms into meaningful, standardized and I-ADOPT compliant components; integrate feedback from semantic and domain for assessing the model's quality, accuracy and relevance in generating variable descriptions

4. **Documentation and Deployment (0.5 month)**: Provide documentation and deploy a first demonstrator of a stand-alone open-source tool, allowing seamless integration into various research workflows and platforms for scientists working with earth science observations.

Due to the integration of our consortium in other initiatives and communities like the RDA, the OGC, the NFDI4Earth or the DataHub of the Helmholtz Research Field Earth and Environment, we will further evaluate the potential impact and usability within a broader research data infrastructure. Transparency and transferability are further ensured by the publication of our model via open repositories, making it easy to adopt in other projects and also to scale our solutions for future needs.

III. *Relevance for the NFDI4Earth*

The main purpose of our demonstrator is to support the transformation of natural language descriptions of observational properties into machine-interpretable and interoperable I-ADOPT-compliant representations. With our solution, data providers and curators will be able to automatically receive a selection of FAIR variable descriptions. We, hence, provide a crucial tool required for the large-scale transformation of existing vocabularies into I-ADOPT enriched knowledge graphs that can be used in metadata and data annotations. As our system also supports the entry of descriptive sentences in natural language, we help data producers to provide I-ADOPT-compliant terms without the need to search through extensive vocabularies and terminology services. This approach allows infrastructure providers and system integrators to rely on the well-defined components instead of complex and highly domain-specific terms. Thus, by using reusable, machine-interpretable components, we streamline the integration of (meta)data from different disciplines into interconnected, interoperable systems. Regarding the NFDI4Earth repositories and infrastructures, our demonstrator has the potential to serve as a connecting building block, integrating (meta)data from various sources into, e.g., one central metadata portal, significantly enhancing the findability and interoperability of the underlying data.

IV. *Deliverables*

D1: Curated list of I-ADOPT-compliant vocabularies, the corresponding API-queries and well-structured database of variable terms and their respective I-ADOPT representation, that serve as reference for training and evaluating our augmented LLM.
D2: Demonstrator of the trained LLM via openly accessible Jupyter Notebooks
D3: Publication of all results and material (software via Helmholtz Codebase, data and results via open repositories)

V. *Finance plan*

For developing our demonstrator, we require three person-months FTE for a semantic- and metadata expert as well as three person-months FTE for a data scientist and software developer at the Karlsruhe Institute of Technology, i.e., in total six person months FTE corresponding to 35.800€.