SE 25 SOFTWARE ENGINEERING

deRSE CONFERENCE
KARLSRUHE 2025

Contribution ID: **153** Type: **Poster**

From Idea to Prototype: Using BITS and LLMs to automate the annotation process for SGN Collection Data

Wednesday 26 February 2025 19:40 (20 minutes)

Being cross-disciplinary at its core, research in Earth System Science comprises divergent domains such as Climate, Marine, Atmospheric Sciences and Geology. Within the various disciplines, distinct methods and terms for indexing, cataloguing, describing and finding scientific data have been developed, resulting in a large amount of controlled Vocabularies, Taxonomies and Thesauri. However, given the semantic heterogeneity across scientific domains (even within the Earth System Sciences), effective utilisation and (re)use of data is impeded while the importance of enhanced and improved interoperability across research areas will increase even further. The BITS Project (BluePrints for the Integration of Terminology Services in Earth System Sciences) aims to address the inadequate implementation of encoding semantics by establishing a Terminology Service that may serve the whole ESS Community on national, european and international level. It will be developed based on the existing TS of the TIB, supplemented by an ESS Collection that already contains relevant terminologies for Earth and Environmental Sciences and to which further relevant terminologies will be added. The implementation of this TS within two data repositories (WDCC at the German Climate Computing Center and a Data Collection at Senckenberg) will showcase the benefits for such different data regarding e.g. enhanced and improved discoverability of research products or automated metadata annotation.

We will present a workflow at SGN that combines the usage of BITS outcome (i.e. ESS collection of the TIB TS) with GPT4all in order to identify gaps in terminologies on the one hand, and provide assistance to scientists, who are working on new collections on the other hand. Based on two major data management challenges facing SGN, Legacy Data Digitisation (historical grown data require systematic transformation into machine-readable formats) and Data Proliferation Management (continuous input of data generated by ongoing collection efforts and research activities), our prototyping process can be divided into several areas:
- Identifying nominal phrases (NPs) in the collection data and annotating them using BITS TS. Our primary goal was to achieve reliable detection, with a focus on minimising false negatives, while accepting some false positives during annotation.
- During the prototyping phase, several obstacles were encountered referring to poor NP detection quality in scientific texts and a lack of reliability in conjunction splitting and singularization using common tools. It is also not always possible to determine the correct language of the text, especially with mixed-language content.
- Revising our requirements had let us choose GPT4all as our preferred solution, specifically the Meta-Llama-3-8B-Instruct.Q4_0.gguf model.
- This allows us to perform high quality NP detection and transformation, but with very high computational and time requirements. To optimise resource utilisation, GPT4all is employed only for high-level operations. Other operations can be performed by tools with less hardware requirements.
- Using statistical logging allows us to identify various significant information about the NP detection and usage. This data we can reuse in later development steps.

By leveraging the strengths of BITS and GPT4all, SGN is paving the way for more accurate processing of complex scientific data to improve research outcomes.

I want to participate in the youngRSE prize

Primary authors: WOLODKIN, Alexander (Senckenberg –Leibniz Institution for Biodiversity and Earth System Research); MARTENS, Claudia (German Climate Computing Center)

Presenter: WOLODKIN, Alexander (Senckenberg –Leibniz Institution for Biodiversity and Earth System Research)

Session Classification: Poster and Demo Session together with Reception

Track Classification: Research Software: AI and ML in a research context