AI4DiTraRe: Towards LLM-Based Information Extraction for Standardising Climate Research Repositories

Anna M. Jacyszyn[1], Shufan Jiang[1], Genet Asefa Gesese[1,2], Sven Hertling[3,1], Tobias Kerzenmacher[4], Peer Nowack[5,4], Sabine Barthlott[4], Etienne Posthumus[1], Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

[2] Karlsruhe Institute of Technology, Institute of Applied Informatics and Formal Description Methods, Kaiserstr. 89, 76133 Karlsruhe, Germany

[3] University of Mannheim, Data and Web Science Group, School of Business Informatics and Mathematics, B 6, 26, 68159 Mannheim, Germany

[4] Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research-Atmospheric Trace Gases and Remote Sensing, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

[5] Karlsruhe Institute of Technology, Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Germany, Am Fasanengarten 5, 76131 Karlsruhe, Germany

Anna.Jacyszyn@fiz-Karlsruhe.de, Shufan.Jiang@fiz-Karlsruhe.de, Genet-Asefa.Gesese@fiz-Karlsruhe.de, sven.hertling@uni-mannheim.de, kerzenmacher@kit.edu, peer.nowack@kit.edu, sabine.barthlott@kit.edu, Etienne.Posthumus@partners.fiz-Karlsruhe.de, Harald.Sack@fiz-Karlsruhe.de

Abstract

In the petabyte-era of climate research, harmonising diverse environmental and geoscientific datasets is critical to improve data interoperability and support effectiveness of interdisciplinary studies. This paper presents an idea of designing an LLM-based tool to extract and standardize metadata from climate research repositories. The solution leverages the adaptability of LLMs that are able to understand contextual nuances. By addressing common inconsistencies such as varying parameters (observation types), units, and definitions, the proposed tool will significantly improve effective data integration. It will be the first step to facilitate the creation of a unified metadata schema adhering to the FAIR principles.

Introduction

In the era of Earth data repositories growing to petabytes sizes (e.g. more than 24 PB of NOAA data; Willett et al. (2023)), a key challenge in climate research is managing and publishing large datasets. A good example are publications relating to the Infrared Atmospheric Sounding Interferometer[1] (IASI; Schneider et al. (2022)), which is deployed on the Metop satellites of European Organization for the Exploitation of Meteorological Satellites (EUMETSAT). The challenge of publishing data consistently is complicated by the fact that climatologists harvest their data from multiple and different sources, including ground-based observatories and stations, balloons, and aeroplanes. On top of the task of extracting knowledge from these huge datasets, this necessitates a proper synchronisation of datasets, integrating different data formats, and mapping standards (Ramachandran 2023).

[1]MUSICA IASI full retrieval product, https://www.imk-asf.kit.edu/english/musica-iasi-v321.php

Numerous existing Earth data repositories contain datasets originating from various research groups which use different standards. The lack of successful standardisation of metadata leads to metadata inconsistencies, such as datasets including identical measurements but tagged inconsistently (e.g., *Latitude/Longitude* vs. *Geocoordinates*). These inconsistencies hamper interoperability and data reuse, particularly in climate research requiring cross-repository analyses.

Leveraging Large Language Models (LLMs) for metadata standardisation is a promising route to address these challenges. LLMs are effective at recognising patterns, resolving ambiguities in natural language descriptions, and generating standardised metadata entries (Zhao et al. 2023). This paper proposes an idea of a novel LLM-based tool for extracting and harmonising metadata in climate research repositories.

The task of standardising metadata is a complex one and can be divided into two main parts: standardising the existing datasets, and creating an environment to ensure a proper standardisation of newly added datasets. In this paper we will concentrate on one of the steps connected to the latter part, namely creating an LLM-based information extraction tool. The tool is supposed to be included in the repository and automatically extract available information from the text accompanying the dataset to be uploaded.

This research is carried out within the scope of the Leibniz Science Campus *Digital Transformation of Research*[2] (DiTraRe) (Razum et al. 2023; Jacyszyn et al. 2024).

Background and Analysis

Earth Data Portal

In this study we exploit the Earth Data Portal (Heß et al. 2023). It serves as a collaborative platform for discovering, visualizing, and downloading environmental sciences

[2]DiTraRe web page, https://www.ditrare.de/en

data. Its primary goal is to provide scientists with access to reusable datasets to enhance research efficiency. Although the portal currently displays metadata information and dataset parameters (observation types), the abstracts associated with these datasets may also contain valuable information that could be extracted to further improve their usability.

Preliminary Analysis of the Earth Data Portal

We use the search API[3] to extract metadata from the Earth Data Portal. We find that (1) only dataset records imported from the PANGAEA repository[4] (Felden et al. 2023) are indexed with *parameters*, while dataset records imported from other repositories such as RADAR[5], Natural Environment Research Council[6] and World Data Center for Climate[7] describe the dataset in unstructured abstracts or titles.

(2) Existing parameters are not standardised. Figure 2 shows duplicate terms (e.g., *Age* vs. *AGE*) and nested concepts (e.g., *Speed* contains *Wind speed*).

In this study, we study the metadata extraction with LLMs exploiting their adaptability. Unlike traditional keyword-matching algorithms, LLMs understand contextual nuances in dataset descriptions. For instance, they can infer that *sea surface temp* [8]. and *SST* refer to the same parameter.

Proposed Approach

LLMs have been applied to improve heterogeneous data integration in several subtasks, such as schema extraction (Bai et al. 2023; Chen and Koudas 2024) and entity linking (Oshima et al. 2024). The implicit knowledge represented in LLMs can improve text processing tasks (Chen and Koudas 2024; Zhao et al. 2023) if there are few labelled data available.

Based on the preliminary analysis of the Earth Data Portal, we suggest building the following components with LLMs:

1. **A Terminology for Dataset Parameters**
   - Describe different parameters within datasets by providing (a) canonical forms and their variants (e.g.,*TEMP* vs *temperature*), which help mapping field names used in the datasets to a standardised parameter name; (b) information about units and measurement (e.g., "km/h" and "m/s" for *Speed*).
   - Integrate existing terminologies such as PANGAEA's catalogue (Diepenbroek et al. 2017) to ensure interoperability.
2. **Automate Parameter Detection and Linking**
   - We have extracted 28k dataset records containing both abstract and structured parameters. These records can serve as training data to train a system to automatically detect mentions of parameters in unstructured textual fields such as titles and abstracts.
   - Leverage LLMs to map mentions of parameters to their corresponding entries in the reference terminology.
3. **A Chatbot for Dataset Import and Retrieval**
   - Assist users with dataset importation tasks:
     - Propose candidate parameters based on the users' input and the reference terminology.
     - Prompt users to add missing parameters when none are detected in their input.
   - Improve dataset discovery:
     - Recommend related data set records based on the metadata.
     - Guide the user to specify their query.

Summary

This study proposes development of an innovative LLM-based tool aimed at standardising metadata across diverse climate research repositories. It is a part of ongoing research within the Leibniz Science Campus *DiTraRe*. The goal of the project introduced in this article is to use LLMs to support partial automatisation of providing basic information about the dataset to the Earth Data Portal. In the first step, all relevant metadata information available in the accompanying paper or abstract should be extracted and fed to the portal. In the next step, an LLM-based chatbot is to be developed to support users during the upload process by providing suggestions to harmonise metadata by mapping into a chosen existing standard.

By resolving common inconsistencies in parameter definitions, naming conventions, and units, the tool will facilitate the creation of a unified schema aligned with FAIR data principles. It will support the integration of over 800,000 datasets from major repositories, such as PANGAEA and the World Data Center for Climate. This harmonisation will enhance the ability to perform comparative analyses and predictive modelling, with implications for addressing global challenges like climate change and biodiversity loss. The project represents a significant step towards automating metadata standardisation, with future plans for scaling the tool to include research papers and developing a user-facing chatbot for metadata mapping.

Analysis of the parameters in Earth Data Portal

Figure 1 shows that, as an example, *Wind Speed* can be described in the title, abstract, or parameters fields in the Earth Data portal. These texts may also contain other information which could be extracted and used.

In Figure 2, the 100 most frequent parameters in the Earth Data Portal are visualized, clustered using SBERT embeddings and DBSCAN, with bubble size representing parameter occurence frequency and plotted via multidimensional scaling.

[3] https://earth-data.de/rest/search
[4] https://www.pangaea.de/
[5] https://www.radar-service.eu
[6] https://eidc.ac.uk/
[7] https://www.wdc-climate.de/ui/
[8] https://en.wikipedia.org/wiki/Sea_surface_temperature

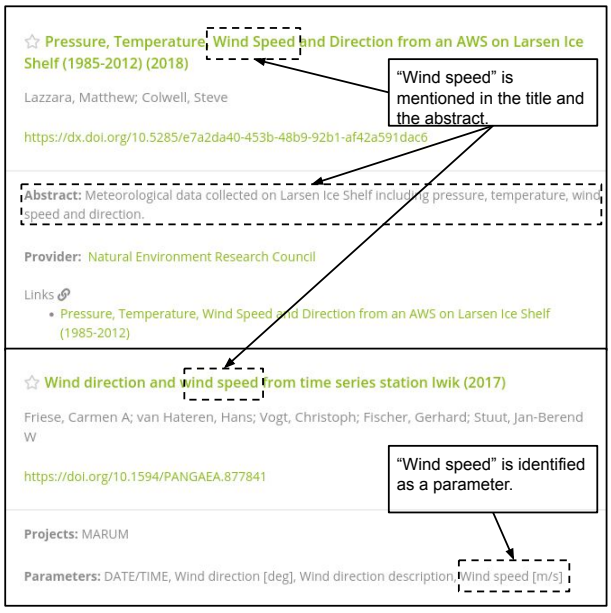

Figure 1: Mentions of *Wind speed* in the Earth Data Portal.

Acknowledgments

The Leibniz Science Campus *Digital Transformation of Research* (DiTraRe) is funded by the Leibniz Association.

References

Bai, F.; Kang, J.; Stanovsky, G.; Freitag, D.; Dredze, M.; and Ritter, A. 2023. Schema-driven information extraction from heterogeneous tables. *arXiv preprint arXiv:2305.14336*.

Chen, K.; and Koudas, N. 2024. Unstructured Data Fusion for Schema and Data Extraction. *Proceedings of the ACM on Management of Data*, 2(3): 1–26.

Diepenbroek, M.; Schindler, U.; Huber, R.; Pesant, S.; Stocker, M.; Felden, J.; Buss, M.; and Weinrebe, M. 2017. Terminology supported archiving and publication of environmental science data in PANGAEA. *Journal of biotechnology*, 261: 177–186.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Felden, J.; Möller, L.; Schindler, U.; Huber, R.; Schumacher, S.; Koppe, R.; Diepenbroek, M.; and Glöckner, F. O. 2023. PANGAEA-data publisher for earth & environmental science. *Scientific Data*, 10(1): 347.

Heß, R.; Albers, K.; Konopatzky, P.; Koppe, R.; and Walter, A. 2023. The Earth Data Portal for Finding and Exploring Research Content. In *EGU General Assembly Conference Abstracts*, EGU–9952.

Jacyszyn, A.; Sack, H.; Group, D.-S.; Bach, F.; and Razum, M. 2024. DiTraRe: AI on a Spider's Web. Interweaving Disciplines for Digitalisation. In *4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment*, volume Vol-3780.

Oshima, Y.; Shindo, H.; Teranishi, H.; Ouchi, H.; and Watanabe, T. 2024. Synthetic Context with LLM for Entity Linking from Scientific Tables. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, 202–214. Bangkok, Thailand: Association for Computational Linguistics.

Ramachandran, R. 2023. From Petabytes to Insights: Tackling Earth Science's Scaling problems in Data, Information and Processes. In *AGU Fall Meeting Abstracts*, volume 2023, IN41A–01.

Razum, M.; Bach, F.; Brünger-Weilandt, S.; Scherz, C.; Böhm, F.; Sack, H.; and Volkamer, M. 2023. Proposal for a Leibniz ScienceCampus – Digital Transformation of Research (DiTraRe). Project proposal.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Schneider, M.; Ertl, B.; Diekmann, C.; Khosrawi, F.; Weber, A.; and et al. 2022. Design and description of the MUSICA IASI full retrieval product. *Earth System Science Data*, 14(2): 709–742.

Willett, D. S.; Brannock, J.; Dissen, J.; Keown, P.; Szura, K.; Brown, O. B.; and Simonson, A. 2023. NOAA Open Data Dissemination: Petabyte-scale Earth system data in the cloud. *Science Advances*, 9(38): eadh0032.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

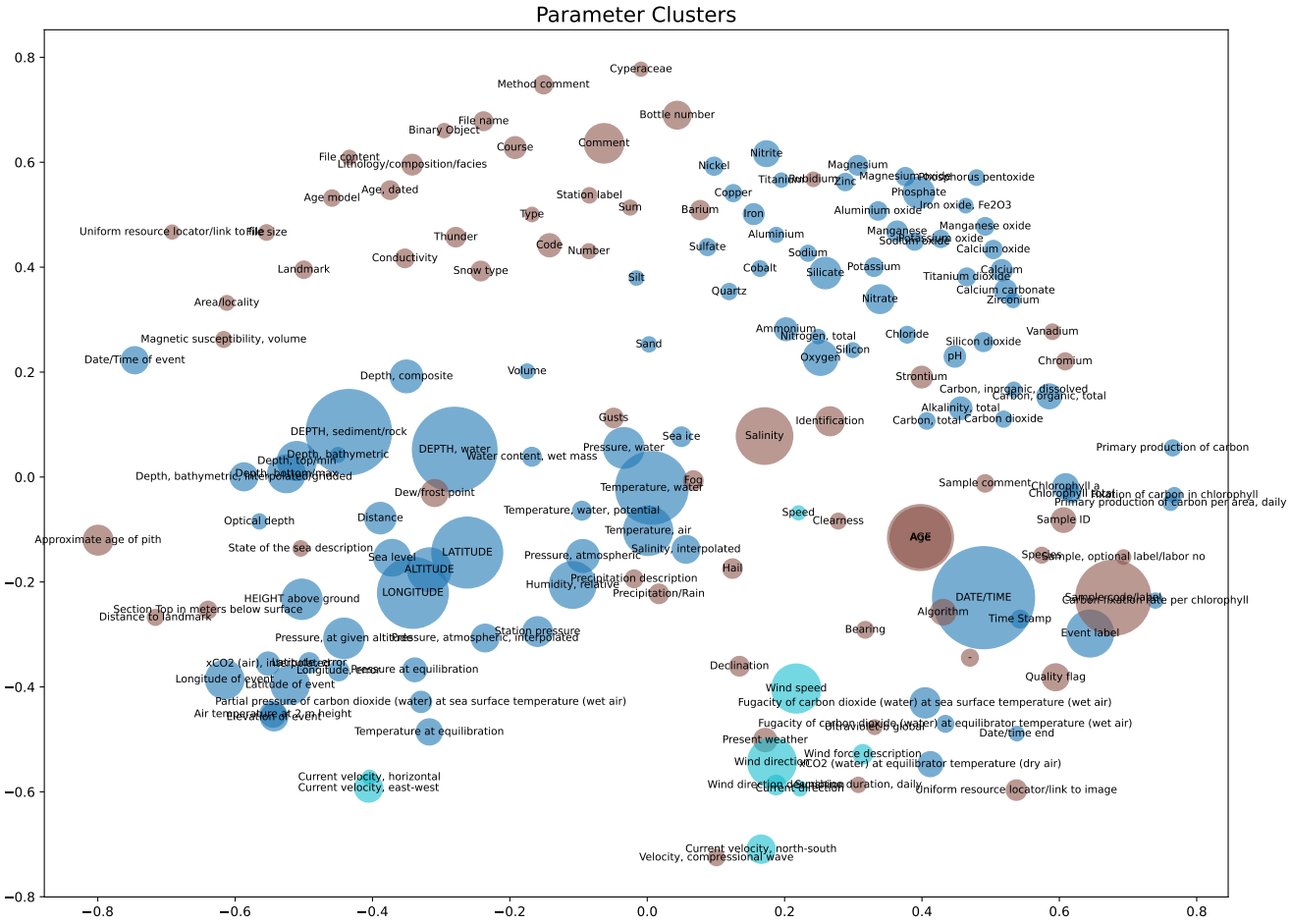

Figure 2: 100 most frequent parameters in Earth Data Portal. We use Sentence BERT (Reimers and Gurevych 2019) to get embeddings of the parameters and calculate the semantic similarity of parameters based on cosine similarity of the embeddings. We use DBSCAN (Ester et al. 1996) to group similar parameters into clusters (bubbles with the same colour). Then we use multidimensional scaling to plot the embeddings in two dimensions. Size of the bubble represents frequency of the occurence of the parameter.