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

Leibniz ScienceCampus
Digital Transformation of Research

Anna Jacyszyn

with S. Jiang, G.A. Gesese, S. Hertling, T. Kerzenmacher, P. Nowack, S. Barthlott, E. Posthumus, H. Sack

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

KIT
Karlsruhe Institute of Technology

AI 4 Scholarly Communication, Bridge @ AAAI, 25th-26th February 2025

DiTraRe = Digital Transformation of Research

Leibniz Science Campus

Leibniz ScienceCampus
Digital Transformation of Research

Regional growth core to **establish new research branch**.

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
+
Karlsruhe Institute of Technology (KIT)

- Planned as a **4+4 years** project (start: September 2023).
- Funded by the Leibniz Association + FIZ KA + KIT.
- Analyse the process of **digitalisation of research**.
- Multilevel **interdisciplinary** approach.
- Very broad, general scope.

DiTraRe Use Cases

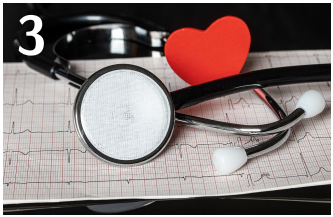
Leibniz ScienceCampus
Digital Transformation of Research

Sensitive Data in Sports Science

KIT Institute of **Sports and Sports Research**

Chemotion Electronic Lab Notebook

KIT Institute of **Biological and Chemical Systems**



AI in Biomedical Engineering

KIT Institute of **Biomedical Engineering**

Publication of Large Datasets

KIT Institute of **Meteorology and Climate Research**

Managing and publishing in the petabyte era

UC4: Publication of large datasets (climate research)

An example: Infrared Atmospheric Sounding Interferometer (IASI) data.

- Instrument on the MetOp satellite.
- Collects data on atmospheric temperature and humidity.
- Current size of IASI full retrieval product: 25 TB (15 GB per day).
- No exploration of data possible (i.e. map).
- RADAR(4KIT) currently doesn't enable download of sub-files.

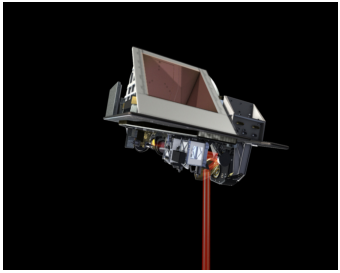


Figure: Infrared radiation directed into the IASI instrument. *Credit: ESA.*

Managing and publishing in the petabyte era

UC4: Publication of large datasets (climate research)

- Climatologists harvest their data from multiple sources:
  - ground-based observatories and stations
  - balloons
  - aeroplanes
  - satellites

Figure: *Credit: KIT-IMKASF*

The problem of standardisation

UC4: Publication of large datasets (climate research)

- Extracting knowledge from these huge datasets is challenging.
- Proper synchronisation of datasets, integrating different data formats, and mapping standards is needed.
- Different standards → metadata inconsistencies.
  - I.e. datasets including identical measurements but tagged inconsistently (e.g., *Latitude/Longitude* vs. *Geocoordinates*).
- Data reusability and interoperability is strongly limited.
  - Climate research requires cross-repository analyses.

→ *AI4DiTraRe goal in UC4: Support creation of a uniform data management platform.*

Idea: LLMs to the rescue

UC4: Publication of large datasets (climate research)

- Leveraging LLMs for metadata standardisation is a promising route.
- Effective at:
  - Recognising patterns.
  - Resolving ambiguities in natural language description.
  - Generating standardised metadata entries.

Idea: LLMs to the rescue

UC4: Publication of large datasets (climate research)

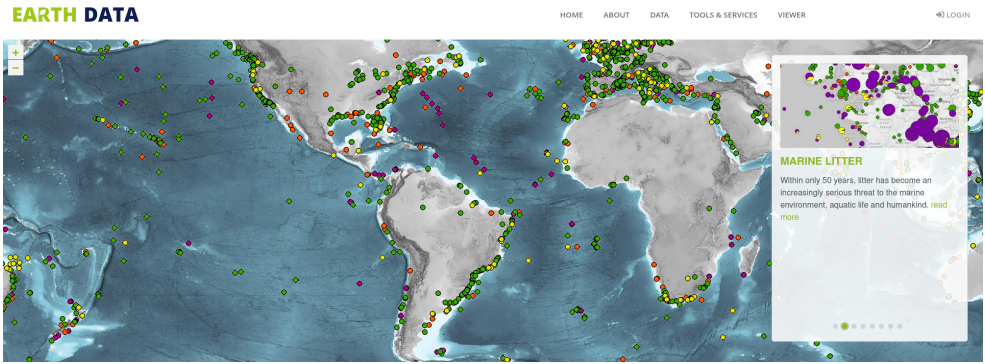
- Leveraging LLMs for metadata standardisation is a promising route.
- Effective at:
  - Recognising patterns.
  - Resolving ambiguities in natural language description.
  - Generating standardised metadata entries.
- An idea of a novel LLM-based tool for extracting and harmonising metadata in climate research repositories.

Earth Data Portal

Leibniz ScienceCampus
Digital Transformation of Research

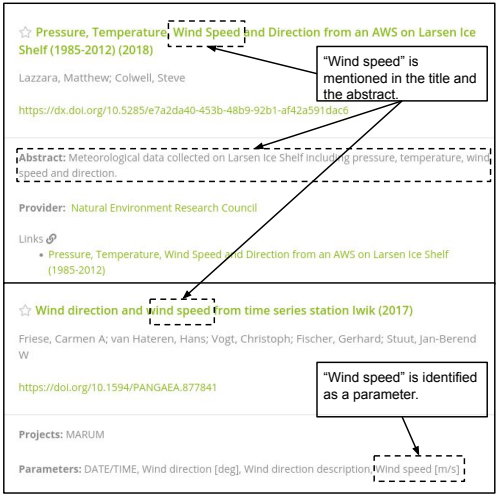
UC4: Publication of large datasets (climate research)

- Collaborative platform for discovering, visualising, and downloading environmental sciences data.
- Provide access to reusable datasets to enhance research efficiency.



Earth Data Portal

UC4: Publication of large datasets (climate research)



- Abstracts and titles also contain information.
- I.e. *wind speed*.

Preliminary analysis

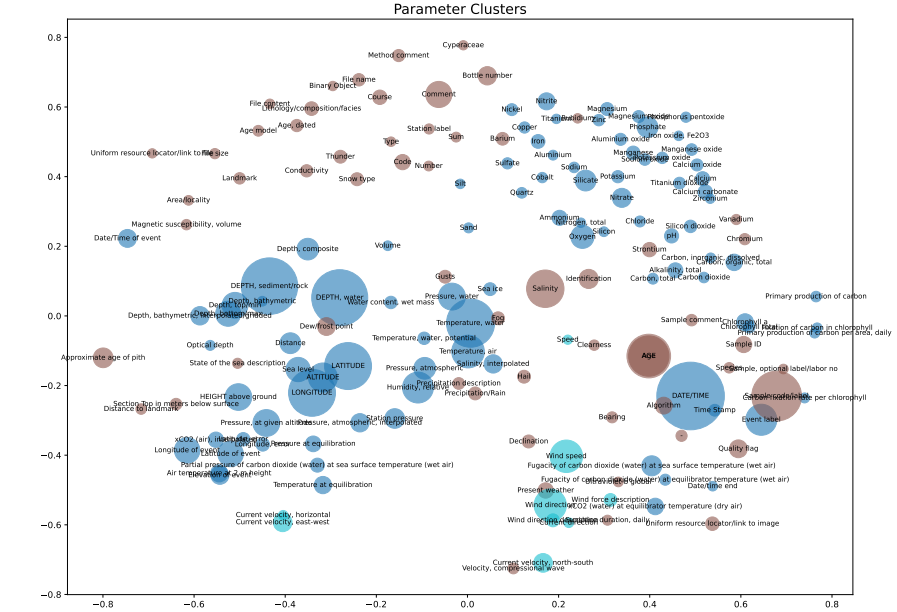
UC4: Publication of large datasets (climate research)

- API to extract metadata from the Earth Data Portal.
  1. Only datasets imported from the PANGAEA repository are indexed with parameters.
  2. Existing parameters are not standardised.

Preliminary analysis

UC4: Publication of large datasets (climate research)

- API to extract metadata from the Earth Data Portal.
  1. Only datasets imported from the PANGAEA repository are indexed with parameters.
  2. Existing parameters are not standardised.
- Sentence BERT → embeddings of the parameters, semantic similarity of parameters based on cosine similarity of the embeddings.
- DBSCAN → group similar parameters into clusters (bubbles with the same colour).
- Multidimensional scaling → plot the embeddings in two dimensions.
- Size of the bubble represents frequency of the occurence of the parameter.



- Duplicate terms (i.e. *Age/AGE*).
- Nested concepts (i.e. *Wind speed* in *Speed*).

Proposed approach

UC4: Publication of large datasets (climate research)

1. A terminology for dataset parameters.
   - Canonical forms and their variants (e.g. *TEMP* vs *temperature*).
   - Integrate existing terminologies such as PANGAEA.

Proposed approach

UC4: Publication of large datasets (climate research)

1. A terminology for dataset parameters.
   - Canonical forms and their variants (e.g. *TEMP* vs *temperature*).
   - Integrate existing terminologies such as PANGAEA.
2. Automate parameter detection and linking.
   - Dataset: abstracts and structured parameters → train to detect parameters in unstructured text (titles and abstracts).
   - LLMs to map parameters to the reference terminology.

Proposed approach

UC4: Publication of large datasets (climate research)

Leibniz ScienceCampus
Digital Transformation of Research

1. A terminology for dataset parameters.
   - Canonical forms and their variants (e.g. *TEMP* vs *temperature*).
   - Integrate existing terminologies such as PANGAEA.
2. Automate parameter detection and linking.
   - Dataset: abstracts and structured parameters → train to detect parameters in unstructured text (titles and abstracts).
   - LLMs to map parameters to the reference terminology.
3. A chatbot for dataset import and retrieval.
   - Assist users with dataset importation tasks (i.e. make suggestions).
   - Improve dataset discovery (i.e. recommend related datasets).

Outlook

UC4: Publication of large datasets (climate research)

- This paper is a vision paper presenting an idea which is currently being developed.
- How to clean up the data for a unified portal?
  - Earth Data Portal has flat categories – different granularities of definitions.
  - How to introduce controlled vocabularies?
  - Combination of historical data and new uploads.
  - PANGAEA repository as the most complete – use it as a training set.
- **Neurosymbolic AI**: ML + symbolic logic

Summary

- DiTraRe = Digital Transformation of Research.
- Ongoing interdisciplinary research.
- **Role of AI** within the process of digitalisation?

DiTraRe

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

KIT
Karlsruhe Institute of Technology

Summary

- DiTraRe = Digital Transformation of Research.
- Ongoing interdisciplinary research.
- **Role of AI** within the process of digitalisation?

→ DiTraRe Symposium: December, Karlsruhe, Germany

→ Anna.Jacyszyn@fiz-karlsruhe.de

→ ditrare@fiz-Karlsruhe.de

DiTraRe

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

KIT
Karlsruhe Institute of Technology

Summary

- DiTraRe = Digital Transformation of Research.
- Ongoing interdisciplinary research.
- **Role of AI** within the process of digitalisation?

→ DiTraRe Symposium: December, Karlsruhe, Germany

→ Anna.Jacyszyn@fiz-karlsruhe.de

→ ditrare@fiz-Karlsruhe.de

DiTraRe

Thank you for your attention! ☺

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

KIT
Karlsruhe Institute of Technology

Which LLMs are we going to use?

- On our server we have:
  - Mistral-7B
  - Falcon-7B
  - Llama-7B
- Open LLMs to ensure reproducibility.
- Firstly, we need to prepare a ground truth to evaluate existing LLMs for our task.
  - Any suggestions?