Enhancing Dataset Discovery and Usage Tracking in Earth Sciences: Integrating Knowledge Graphs and Large Language Models

AMS 2025
January 12-16, 2025
Irina.Gerasimov@nasa.gov

https://disc.gsfc.nasa.gov/ NASA/Goddard Earth Sciences Data and Information Services Center (GES DISC)

Irina Gerasimov[1,2], Armin Mehrabian[1,2], Jerome Alfred[1,2], Binita KC[1,2], Andrey Savtchenko[1,2], Kendall Gilbert[1,2], Jennifer Wei[1]
[1]Code 619, NASA Goddard Space Flight Center, Greenbelt, MD, USA [2]ADNET Systems Inc., Lanham, MD, USA

Abstract

NASA's Data Active Archive Centers (DAACs) have played a crucial role in supporting a wide range of applied research in Earth and Environmental sciences. To date, over 20,000 publications have been collected, citing more than 3,000 NASA Earth science datasets.

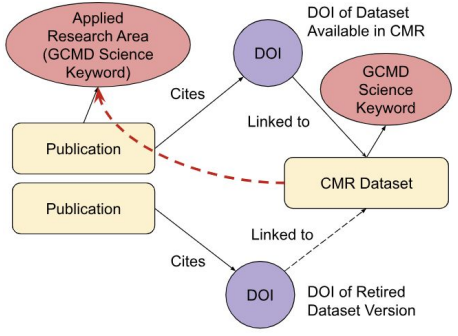
We present an innovative approach that links datasets and collected publications through a knowledge graph (KG). This KG enables the tracking of dataset citations throughout the dataset's lifecycle, revealing patterns of dataset usage across various applied research areas. We fine-tuned the pre-trained NASA IMPACT INDUS-Base Retriever Large Language Model (LLM) using a set of labeled publication abstracts. Our results indicate that 87% of the publications were classified into one of twenty applied research areas, while the remaining 13% were categorized into non-applied research areas.

The classified publications linked to datasets are used to discover datasets by users interested in specific applied research and by dataset providers to determine dataset usage for applications.

Building Knowledge Graph

Collect publications referencing the Earth Observing System Data and Information System (EOSDIS) dataset DOIs from Web of Science, Scopus, Crossref, Google Scholar, and DataCite; for details see Gerasimov, et al. (2024), 10.5334/dsj-2024-001.

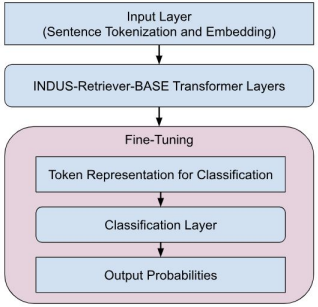
Determine linkage of EOSDIS dataset DOIs to Common Metadata Repository (CMR) data collections. Assign Applied Research Areas to publications using the Global Change Master Directory (GCMD) keywords. When possible, link DOIs of retired dataset versions to their current versions in CMR.



Connecting publications to CMR data collections in Knowledge Graph allows for linking publications' Applied Research Areas to data collections currently offered to CMR users.

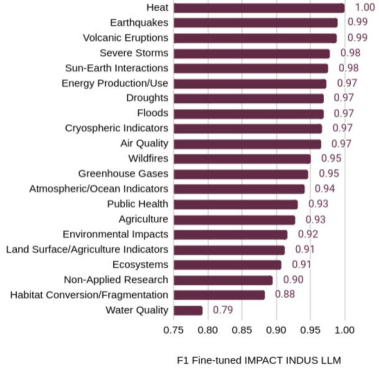
Fine-tuning LLM for Publications Classification into Applied Research Areas

To classify the diverse range of collected publications, the pre-trained NASA INDUS-Base Retriever LLM (10.48550/arXiv.2405.10725) was fine-tuned on a labeled set of 1,065 publications. As a result, ~87% of collected publications were successfully categorized into one of twenty applied research areas.



Publications ratio for training/validation is 80%/20%
Model performance monitored using accuracy, precision, recall, and F1-score
The best-performing model was evaluated using independent test set comprising 1,036 unseen publications

F1 Scores for Evaluation of 1,036 Abstracts



Dataset Discovery and Usage Tracking

The GES DISC Publications website https://disc.gsfc.nasa.gov/information/publications includes an *Applied Research Areas* facet for Publications browsing. The facet allows users to browse publications and discover the datasets used in an Applied Research Area of user interest.
Dataset providers can navigate GES DISC Publications interface to learn about research areas their datasets were used in.
The proposed solution allows for adding new Applied Research Areas in the future and re-training LLM to re-classify publications.
As dataset versions retire, GES DISC Publication system KG tracks DOIs of those retired versions so collected publications can be used to determine applied research areas for the new versions of datasets.

Citation Counts of EOSDIS DOIs in Publications Arranged by Applied Research Area