Discovering Research Areas in Dataset Applications through Knowledge Graphs and Large Language Models

Irina Gerasimov
ADNET Systems, Inc
NASA Goddard Space Flight Center
Greenbelt, MD, USA
0000-0003-0224-5004

Armin Mehrabian
ADNET Systems, Inc
NASA Goddard Space Flight Center
Greenbelt, MD, USA
0000-0003-0587-9387

Binita KC
ADNET Systems, Inc
NASA Goddard Space Flight Center
Greenbelt, MD, USA
0000-0001-6126-5369

Jerome Alfred
ADNET Systems, Inc
NASA Goddard Space Flight Center
Greenbelt, MD, USA
0000-0002-1203-7113

Michael P. McGuire
Department of Computer and Information Sciences
Towson University
Towson, MD, USA
0000-0001-7585-8018

***Abstract*—Scientific datasets are increasingly cited in peer-reviewed journal publications, facilitating easy access to research utilizing those datasets. Datasets undergo a life cycle where older versions of datasets are replaced by newer versions often due to improvements in data resolution, algorithms, and other factors. Unlike peer reviewed documents registered with a single Digital Unique Identifier (DOI), datasets can be updated over time and the newer version of the datasets are registered with a new DOI which is not necessarily linked to the previous version of the dataset. It is challenging when publications citing a dataset need to be traced over the entire life cycle of that dataset. We provide an innovative approach to link the dataset versions and publications using a knowledge graph (KG). KG can help to trace the dataset cited in publications over the entire dataset life cycle and shed light into dataset usage in various applied research areas. We fine-tuned the pretrained NASA IMPACT INDUS Large Language Model (LLM) on a set of labeled publications abstracts. Our results showed that 87% of the publications were classified into one of twenty applied research areas, while the remaining 13% were classified into non-applied research areas. By linking datasets to applied research areas through the KG and employing Global Change Master Directory (GCMD), a well-established controlled vocabulary of scientific keywords describing Earth science datasets, we contribute to a transparent and advanced search and discovery mechanism for datasets across the Earth data ecosystem. The integrated KG and LLM approach is now incorporated and operational in dataset publication management at one of NASA's Earth science data archival centers.**

***Index Terms*—LLM, Knowledge Graph, data citation**

I. Introduction

Earth science data are derived from diverse sources, such as satellites, suborbital field campaigns, computational models, terrestrial monitoring stations, and ocean buoys, each generating data to study Earth's environmental and geophysical characteristics. At NASA, these data are meticulously archived, distributed, and managed by the Distributed Active Archive Centers (DAACs), under the umbrella of Earth Observing System Data and Information System (EOSDIS) [1]. In accordance with NASA's commitment to open science, as detailed in the Scientific Information Policy for NASA's Science Mission Directorate [2], these Earth science data are available with free and unrestricted access. Despite being open access, the management of this vast data encounters significant challenges including ensuring data provenance, enhancing accessibility, maintaining transparency, enabling reproducibility, and upholding accountability.

A key component of ensuring the datasets' relevance, accessibility, versioning, and verifiability in the scientific ecosystem is the integration of persistent digital identifiers, such as DOIs, for datasets, coupled with the establishment of rigorous citation protocols requiring them to be cited with DOIs in scholarly articles [3]. While the practice of citing datasets in scholarly articles has existed for over three decades, it has been challenging to link datasets to scholarly articles because citation rates of datasets are still low [4]. Creating a comprehensive database of scientific articles that cite the datasets requires exhaustive searching of different bibliographic sources [5].

There are challenges to creating and maintaining publications databases due to multiple versions of data over data life cycle. For example, the scholarly article citing the dataset versions that are no longer publicly available presents difficulty of associating this article with the dataset versions that are currently offered to users. Another challenge is extracting knowledge from collected publications and reflecting dataset usage in applied research areas in a transparent way. To address these challenges, our research focuses on modern tools

This work was funded by NASA SESDA IV contract No. 80GSFC17C0003.

that enable users to search dataset-citing articles, extract information, and use metadata to enhance search and discovery. We achieved this through unification of KG and Large Language Models (LLM) technologies [6], in particular, augmenting KG with LLM output for graph extension.

In recent years, the rise of semantic technologies, particularly KGs, has greatly impacted various scientific fields, enhancing the organization of domain-specific knowledge and the application of graph analytics. The use of KGs in geosciences has become increasingly important, as they can be used throughout the entire data life cycle [7]. Google Dataset Search, based on Schema.org, which provides metadata schemas to markup datasets shared on the Web, exemplifies the growing popularity of these technologies [8]. The geosciences communities took several initiatives to develop domain specific ontologies. Two prominent NASA led projects are the GCMD keywords hierarchy [9] and Semantic Web for Earth and Environmental Terminology [10], or SWEET, a formal ontology built based on GCMD hierarchy. The GCMD science keywords are hierarchical controlled vocabularies that cover 14 different categories of keywords in Earth science. NASA has used GCMD science keywords for over three decades to manage dataset metadata for Earth Science datasets.

It has been demonstrated that LLMs such as GPT [11] and BERT [12] trained on vast amounts of data are capable of understanding and generating natural language. It was further shown that for science specific domains, LLMs trained on the domain specific corpora such as SCIBERT [13] produce better accuracy on the domain specific applications. In our research we used the recently developed INDUS LLM suite [14], developed for Earth sciences, biological and physical sciences, heliophysics, planetary sciences, and astrophysics. INDUS was trained using curated scientific corpora drawn from these diverse data sources.

The primary contributions of our work can be outlined as follows:

- A novel KG design augmented with an LLM for extracting and integrating metadata for dataset research outputs analytics, discovery, and search.
- An empirical demonstration of the significance of the dataset versions and research contribution for search and discovery of the dataset that are available to the public.
- Our operational publication management system utilizes our KG with LLM to navigate over 12,000 dataset-citing publications by their applied research areas and discover and narrow down datasets by application areas.

II. Methodology

A. Knowledge Graph design

Our KG consists of the nodes for dataset DOIs, the nodes for dataset DOI citing publications, the nodes for dataset metadata used for EOSDIS dataset search and order, and the nodes for GCMD keywords describing datasets and publications. For EOSDIS datasets, the dataset metadata is stored and managed by Common Metadata Repository (CMR) [15], which is the authoritative management system for all EOSDIS metadata and used for search and discovery datasets and their data files archived by DAACs. The dataset metadata nodes in our KG are referred to as 'CMR datasets'. The CMR dataset metadata describes attributes of Earth science datasets such as dataset ShortName, LongName, VersionID, description, instruments, missions, projects, dataset measured variables, e.g. 'Ozone', 'Precipitation', etc, and dataset possible applications, e.g. 'Air Quality', 'Droughts', etc. The instruments, missions, projects, measured variables, and applications are governed by GCMD Science Keywords hierarchy and populated by the dataset metadata curators. A Science Keyword for a measured variable reflects its content, and is selected accordingly, whereas a Science Keyword for a dataset application is selected based on its predicted or evidential use. With our KG design, we incorporate evidential dataset usage extracted from dataset-citing publications in the form of GCMD vocabulary. This dataset applied usage can be integrated seamlessly with the CMR dataset metadata through linking this usage to publications and publications to datasets.

Since CMR contains only current data collections available to Earth data users, and does not contain collections for retired dataset versions, connecting dataset citing publications with CMR data collections is challenging. The CMR system contains a DOI for the data collection, which allows linking CMR data collections with registered EOSDIS DOIs. Upon reviewing approximately 11,000 EOSDIS DOIs registered as datasets in the DataCite DOI registry [16], we found that only 70% of these DOIs have corresponding data collections in the CMR. Upon investigating the remaining 30% of registered EOSDIS DOIs, we determined that 16% of these DOIs correspond to retired and duplicate versions of CMR data collections. The remaining 14% are either collections not identified as retired dataset versions, collections retired without successor versions, or collections never deposited into CMR. Mapping dataset to CMR data collection allows for extracting persistent GCMD metadata for the KG, and thus the datasets that were not mapped to CMR data collections were not included into KG.

We faced several challenges in the process of mapping DOIs for older versions of the datasets to the newer versions. Our main finding was that, while the recent DataCite schema [17] allows for adding DOIs for preceding and succeeding datasets into a DOI record, this feature is not being used for EOSDIS datasets. As a result, it prevents automated retrieval of provenance of dataset versions. Thus, we had to use heuristic approaches to get version provenance. One of our heuristic approaches relied on the 'ShortName' CMR dataset attribute that is common between dataset versions and a required attribute in CMR and recorded for many datasets in DataCite as an optional attribute. The other heuristic approach was to rely on common parts of DOIs for different dataset versions. We found that although these approaches allowed us to determine version provenance in many cases, these approaches cannot be generalized. In cases when these heuristics were not applicable we examined the DOI landing web pages for old dataset versions. On many of these pages the succeeding version of the dataset was indicated. In addition, the DOI landing web pages of current dataset versions contained dataset version history that we used for the provenance. Since all these web pages were created in a non-uniform way between the datasets, we were not able to create a solution allowing us to automatically retrieve dataset version history.

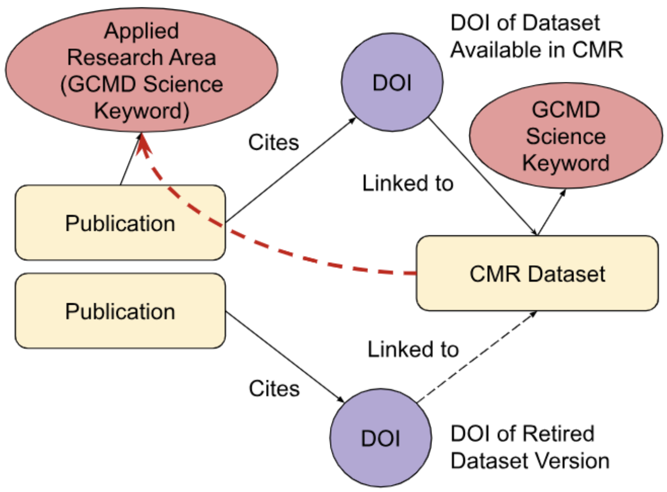

Fig. 1. KG high level design.

Fig. 1 displays the high level design of our KG. The dataset nodes are presented as CMR datasets with EOSDIS dataset DOI nodes linked to them with a solid line for DOIs of collections existing in CMR or with a dashed line for DOIs that are for the retired versions of the CMR data collections. The 'GCMD Science Keyword' nodes are all the Science Keywords that were assigned to CMR data collection. The 'Applied Research Area' nodes are the GCMD Science Keywords that were derived from publications content. Since the GCMD Science Keywords are organized into hierarchy, our KG reflects that with the subcategory edges between Science Keyword nodes. Tables I and II list node types, properties, and relationships between the nodes.

The content for the 'Publications' nodes in our KG is sourced from the EOSDIS Publications Library [18]. An exhaustive search for publication citations was conducted using bibliographic sources such as Web of Science, Scopus, Crossref, Google Scholar, and DataCite. Details of the collection process are outlined in [5]. The library includes citations for books, journal articles, conference papers, dissertations, and reports, while excluding webpages, preprints, discussions, and works without DOIs.

To update our knowledge graph, the following steps are necessary. When new publications citing EOSDIS dataset DOIs are identified, they can be added to the KG as new nodes with edges linking them to the relevant dataset DOIs. If a dataset DOI is not yet present in the KG, the corresponding CMR record must be located and linked to the dataset DOI. If no CMR record is found, the publication and dataset DOI will not be added to the KG.

The more challenging KG updates occur when new CMR entries are created for newer dataset versions and when existing dataset versions are retired from CMR. When a CMR entry is created for a newer dataset version, the CMR dataset node for an older version is identified in the KG. All dataset DOIs linked to the older CMR dataset node will have linkage created to the new CMR dataset node. When the older dataset version is retired from CMR the corresponding node and its linkages are removed from KG. KG DOIs for older dataset version nodes and linked publication nodes that have no more linkages to datasets linked to CMR must also be deleted if there is no successor version for the retired CMR entry.

The primary challenge in creating our KG lies in mapping EOSDIS dataset DOIs to CMR data collections. A significant issue arises in linking older dataset versions to newer ones due to the lack of consistent provenance for older versions. This problem can be resolved by implementing standardized procedures for storing provenance information. In summary, the novelty of our graph is highlighted by two key features. First, the graph preserves dataset version provenance, allowing for access to publications spanning the entire life cycle of a dataset. Second, the graph captures information about dataset application usage in a way that enables the extraction of usage data from the dataset's descriptive metadata, facilitating enhanced search and discovery of datasets.

TABLE I
KNOWLEDGE GRAPH NODE TYPES AND THEIR PROPERTIES

Node Type	Properties
CMR_Dataset	globalId, CMR Concept ID, ShortName, LongName, VersionID, Abstract
DOI	globalId, DOI, Title, Year
Publication	globalId, DOI, Title, Year, Abstract
ScienceKeyword	globalId, Name

TABLE II
RELATIONSHIP TYPES BETWEEN NODE TYPES

Relationship Type	From Node Type	To Node Type
Linked_to_CMR_Dataset	DOI	CMR_Dataset
Has_ScienceKeyword	CMR_Dataset, Publication	ScienceKeyword
Subcategory_of	ScienceKeyword	ScienceKeyword
Cites_DOI	Publication	DOI

A limitation impacting the generalizability of our KG is its dependence on CMR and GCMD Science Keyword metadata. To implement a similar KG in another organization, the dataset metadata would need to be carefully reviewed to ensure proper handling of dataset versioning and the use of relevant terms for search and discovery.

B. Content Classification

For KG integration the publications were classified into applied research areas drawn from the GCMD Science Keywords vocabulary. Upon careful examination of 1,065 publication abstracts we identified 20 applied research areas presented in Table III. While some topics cover more specific content, such as 'Droughts' and 'Earthquakes,' others encompass broader content, such as 'Atmospheric/Ocean Indicators.' The selection of broader topics is due to the difficulty narrowing down the publications topics for closely interrelated research areas.

TABLE III
APPLIED RESEARCH AREAS MAPPED TO GCMD SCIENCE KEYWORDS HIERARCHY

Applied Research Area	GCMD Science Keywords Hierarchy
Agriculture	AGRICULTURE
Air Quality	ATMOSPHERE >AIR QUALITY
Atmospheric/ Ocean Indicators	CLIMATE INDICATORS > ATMOSPHERIC/OCEAN INDICATORS
Cryospheric Indicators	CLIMATE INDICATORS > CRYOSPHERIC INDICATORS
Droughts	HUMAN DIMENSIONS > NATURAL HAZARDS >DROUGHTS
Earthquakes	HUMAN DIMENSIONS > NATURAL HAZARDS >EARTHQUAKES
Ecosystems	BIOSPHERE >ECOSYSTEMS
Energy Production/Use	HUMAN DIMENSIONS > ECONOMIC RESOURCES > ENERGY PRODUCTION/USE
Environmental Impacts	HUMAN DIMENSIONS > ENVIRONMENTAL IMPACTS
Floods	HUMAN DIMENSIONS > NATURAL HAZARDS >FLOODS
Greenhouse Gases	N/A
Heat	HUMAN DIMENSIONS > NATURAL HAZARDS >HEAT
Habitat Conversion/ Fragmentation	HUMAN DIMENSIONS > HABITAT CONVERSION/ FRAGMENTATION
Land Surface/ Agriculture Indicators	CLIMATE INDICATORS > LAND SURFACE/ AGRICULTURE INDICATORS
Public Health	HUMAN DIMENSIONS > PUBLIC HEALTH
Severe Storms	HUMAN DIMENSIONS > NATURAL HAZARDS > SEVERE STORMS
Sun-Earth Interactions	SUN-EARTH INTERACTIONS
Volcanic Eruptions	HUMAN DIMENSIONS > NATURAL HAZARDS > VOLCANIC ERUPTIONS
Water Quality	TERRESTRIAL HYDROSPHERE > WATER QUALITY/WATER CHEMISTRY
Wildfires	HUMAN DIMENSIONS > NATURAL HAZARDS >WILDFIRES

The publications with non-applied research, that were mainly about datasets validation, calibration, and algorithms were placed into a separate category, 'Non-Applied Research'. Each publication was labeled with the most relevant single topic. The fine-tuning process began by leveraging the NASA-IBM INDUS model [14], a large language model pre-trained for natural language understanding tasks, to classify scientific publications. The dataset used for this fine-tuning consisted of 1,065 publication abstracts and titles, each annotated with the corresponding research area. The dataset was stratified and split into two main parts: a training-validation set comprising 1,065 publications and a separate test set of 1,036 unseen publications. The training-validation set was further divided, with 80% (approximately 852 publications) used for training and 20% (approximately 213 publications) for validation. The INDUS model was then fine-tuned on the training set, with its architecture modified to include a classification head tailored to the 20 target research areas. This head consists of fully connected layers that map the rich embeddings generated by the pre-trained model to the appropriate research area categories.

During training, the model's performance was continuously evaluated on the validation set, with key metrics such as accuracy, precision, recall, and F1-score computed for both overall and per-class performance. The best-performing model, based on validation accuracy, was then tested on the independent test set comprising 1,036 unseen publications. The architecture of the fine-tuned model, with its additional classification layers, allowed for an effective adaptation of the INDUS model to the specific nuances of the scientific text in the NASA GES-DISC domain. This fine-tuned model demonstrated a strong ability to accurately classify publications into their respective research areas, making it a valuable tool for enhancing publication and dataset discovery and for integration into knowledge graph applications.

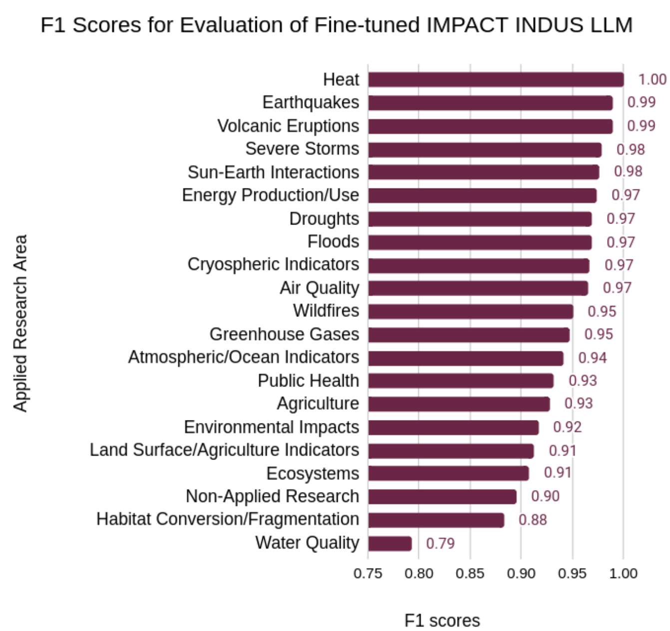

Fig. 2. F1 scores from fine-tuned IMPACT INDUS LLM arranged by applied research area in publications evaluation set.

The resulting fine-tuned model was evaluated by using an unseen labeled set of 1,036 publications. These publications were processed with the fine-tuned model and the output of the model produced a probability density function across all classes. The topic with the highest probability was evaluated against the most prominent labeled topic. F1 scores for 21 topics are shown in Fig. 2 and are ranging from 0.79 to 1.00, with an average F1 score of 0.94 over all topics. Out of all evaluated publications, 5.6% of publications were incorrectly classified. As it is seen in Fig. 2, the model shows improved performance for applied research areas that are more narrowly defined, such as 'Heat' and 'Earthquakes' with performance degrading for broader defined research areas such as 'Habitat Conversion/Fragmentation' and 'Water Quality'.

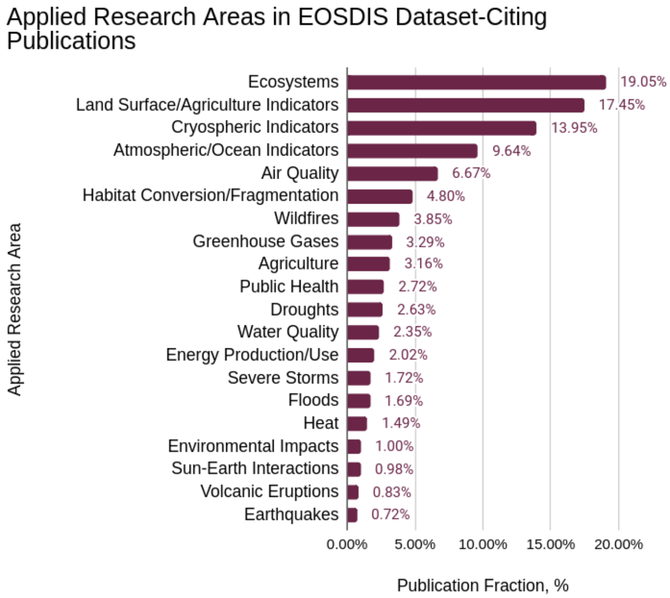

Fig. 3. Distribution of applied research areas in EOSDIS dataset-citing publications.

Using our fine-tuned model we classified 21,873 EOSDIS dataset-citing publications. Among those, 87% were assigned an applied research area while the other 13% of publications were classified as non-applied research. Fig. 3 shows the distribution of research areas over processed publications. Obtained Applied Research Areas were linked to Publication nodes in our KG described in the previous section.

III. Results

A. Obtaining Research Outputs over Dataset's Lifetime

In this section, we demonstrate how a KG that connects different dataset versions throughout its lifetime can provide a comprehensive history of the research outputs generated over that period. In the next section, we present a use case illustrating how this enriched collection of research outputs can enhance the search and discovery of datasets.

For constructing KG we took 21,806 publications from EOSDIS Publications Library [18], where each of the publications is citing one or more of EOSDIS DOIs. This collection of publications contained a total of 32,511 citations of EOSDIS DOIs registered at DataCite as datasets (while a majority of the EOSDIS DOIs are datasets, there is a small portion of DOIs which are documents, projects, or other bibliographic entities). Among these 3,556 DOIs were unique. Using techniques described in the 'Knowledge Graph Design' section each of these DOIs was traced to CMR data collection. Out of 3,556 unique DOIs 2,481 DOIs were found in CMR and 584 DOIs were not found in CMR but were determined as older versions or duplicate DOIs of the datasets that are currently in CMR. The rest of the DOIs, 491, were not matched to neither current CMR collections nor their older versions.

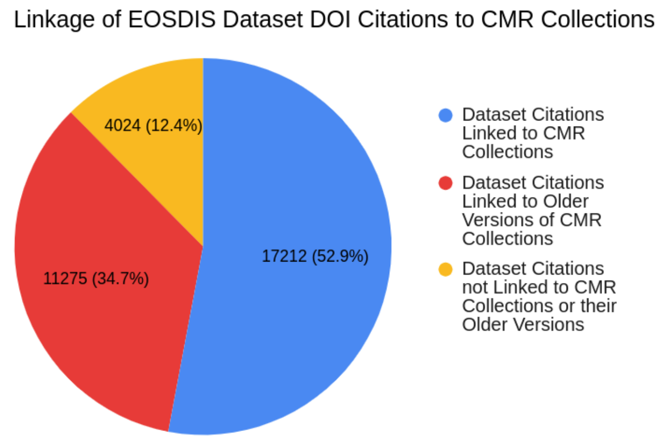

Fig. 4. Distribution of 32,511 EOSDIS DOI citations in 21,806 publications in regards to EOSDIS DOI linkage to CMR.

Based on these findings, the pie chart in Fig. 4 illustrates the counts and percentages of 32,511 citations: those whose DOIs can be linked to current CMR collections, those linked to their older versions, and those that cannot be linked to either. The data in this pie chart reveal that only 52.9% of dataset citations can be linked to the CMR data collections, thus underlining the importance of research outputs produced by the older dataset versions. In presenting these findings, we emphasize the need for streamlined solutions to link older collection versions to publicly available ones. This ensures that research conducted with older versions is preserved and can be used to trace the dataset's historical applications.

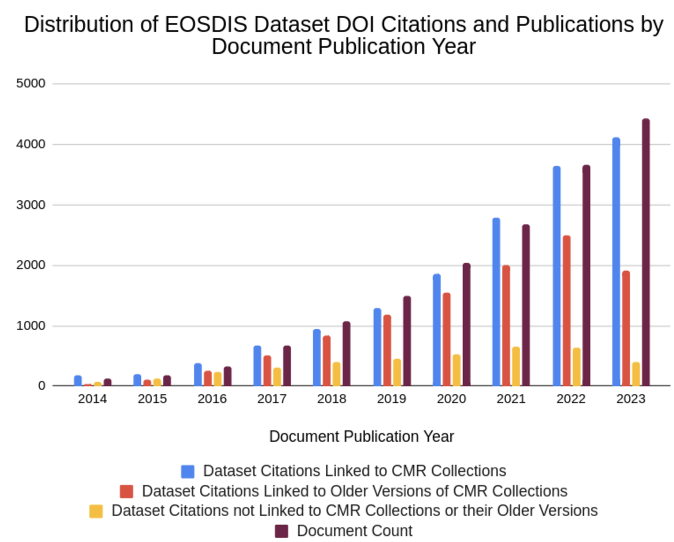

Fig. 5. Citations count for EOSDIS dataset DOIs arranged by the document publication year.

Fig. 5 presents a histogram of the yearly counts of publications by their publication dates and the counts of EOSDIS DOI citations related to dataset linkage to the CMR. The EOSDIS policy requiring DOI registration began in 2012, which is reflected in the progressive growth of dataset-citing documents starting in 2014, as shown in Figure 5. The histograms indicate continuous growth in the counts of citations for dataset DOIs, with data collections currently present in the CMR correlating with document counts.

The counts of DOI citations for older dataset versions were almost as high as those citations for CMR data collections until 2020 but then reduced significantly in 2023. This can be explained by an increase in publications citing CMR data collections. As those collections get retired over the years, we will observe a decrease in counts for current collections and an increase in counts for retired collections.

The counts of DOIs not matching CMR data collections are relatively high up to 2018 and significantly lower in recent years. This can be attributed to the fact that, over time, information about collections that existed five or more years ago becomes lost. Their data landing pages stop functioning, and some collections are never replaced with newer versions.

An additional finding is that out of 3,556 unique DOIs in the 21,806 publications considered, 477 current CMR collections had publications citing their older versions. Moreover, while older versions of 150 of these collections have been cited in publications, their newest versions present in CMR have not been cited yet at all. Overall for these 477 multi-version collections, 4,786 publications cited the newest collection version and 12,099 publications cited older versions.

This sends an important message to dataset creators and distributors: when releasing a new version of a dataset, it is crucial to maintain and provide access to the provenance of older versions. This includes making all dataset-citing works available to users and knowledge processing systems in an automated format.

B. Discovery of Dataset Applied Research Areas

In this section we demonstrate how dataset-citing publications can be used to extract the dataset application areas and how these areas are integrated in the KG to enhance dataset search and discovery. We also underline the significance the older versions of the datasets and publications that link them have in this discovery process.

In the 'Content Classification' section, we described the methodology used to classify publication content into applied research areas. Fig. 3 displays the distribution of 21,806 publications collected from the EOSDIS Publications Library [18] by applied research area. When a dataset DOI is cited in a publication and the publication is assigned an application area, it serves as an evidential use case of that dataset in the given application. The statistics on the linkage of cited EOSDIS dataset DOIs to the CMR highlight the importance of publications that use older versions of datasets in discovering the datasets' applications. The histogram in Fig. 6 shows the counts of dataset DOI citations distributed by the applied research area of the citing publications. These dataset citation counts are displayed as a stacked histogram, with the following color scheme: blue represents the count of dataset citations linked to current CMR collections, red indicates the count of citations linked to older versions of existing CMR collections, and yellow shows the count of citations not linked to CMR collections or their older versions. According to data in Fig. 6, a significant number of the dataset DOI citations come from publications citing old versions of datasets in most applied research areas.

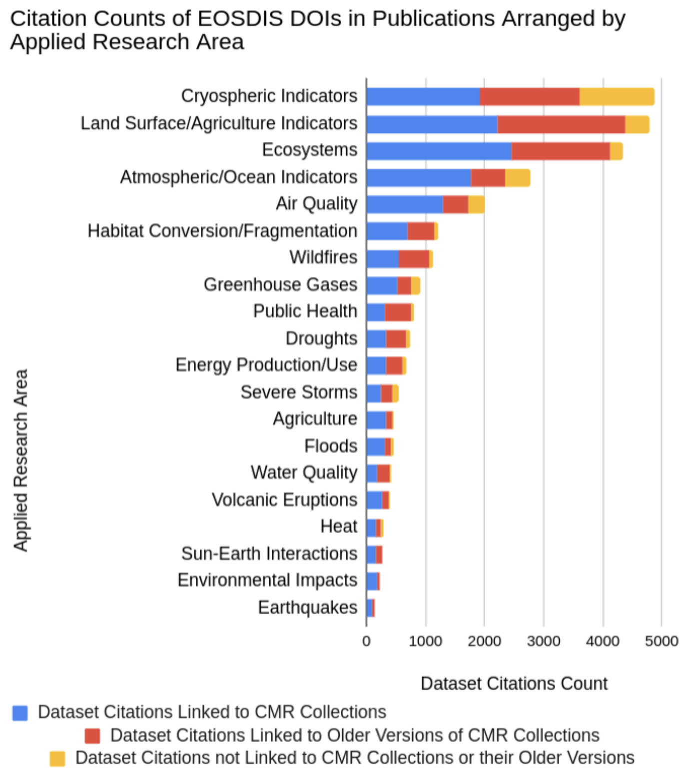

Fig. 6. Stacked histogram of EOSDIS dataset DOI citations in publications, categorized by applied research area and the linkage of the dataset citations to CMR datasets and their older versions.

To demonstrate the use of our KG for discovering applied research areas for a dataset throughout its lifetime, we present a use case of the "ATLAS/ICESat-2 L3A Land Ice Height, Version 6" dataset with the DOI 10.5067/ATLAS/ATL06.006 [19]. This is the latest version of the 'ATL06' dataset published in 2023 and available in CMR for public use. This dataset version is new and thus has no published works referencing it. This dataset's GCMD science keywords describing dataset science content include only "GLACIERS ICE SHEETS", providing a high level information that this dataset contains measurements of the glacier ice sheets. This dataset version supersedes five other versions, each cited in several publications. Table IV presents the data collected from KG for this dataset current and prior versions. This information include dataset's version DOI, dataset version DOI publication year, count of publications citing that specific dataset version DOI, and applied research areas extracted from publications. Some publications focus on dataset validation and algorithms, so not all publications cover applied research.

TABLE IV
EXAMPLE OF DATASET APPLICATIONS THROUGHOUT ITS LIFETIME

Dataset DOI	**Year**	**Pubs Cnt**	**Applied Research Areas**
10.5067/ATLAS/ATL06.001	2019	5	Cryospheric Indicators: 5
10.5067/ATLAS/ATL06.002	2019	8	Cryospheric Indicators: 1
10.5067/ATLAS/ATL06.003	2020	25	Cryospheric Indicators: 17 Public Health: 1 Atmospheric/ Ocean Indicators: 1
10.5067/ATLAS/ATL06.004	2021	8	Cryospheric Indicators: 6 Volcanic Eruptions: 1
10.5067/ATLAS/ATL06.005	2021	15	Cryospheric Indicators: 13 Atmospheric/ Ocean Indicators: 2
10.5067/ATLAS/ATL06.006	2023	0	**None**
All Dataset Collections		60	Cryospheric Indicators: 42 Atmospheric/ Ocean Indicators: 3 Public Health: 1 Volcanic Eruptions: 1

As shown in Table IV, the "ATL06" dataset is primarily used in studies focused on "Cryospheric Indicators." This research area involves the measurement and observation of changes in the Earth's cryosphere, including glaciers, ice caps, ice sheets, sea ice, snow cover, and permafrost. These indicators are essential for understanding climate change, offering insights into temperature trends, sea level rise, and the health of polar and mountainous regions. Beyond the "Cryospheric Indicators" applied research area, our KG has also revealed additional, unexpected applications of the "ATL06" dataset throughout its lifetime, including research in areas such as "Atmospheric/Ocean Indicators," "Public Health," and "Volcanic Eruptions." Careful examination of research conducted in these areas can offer valuable insights to dataset creators and users regarding how the dataset is being utilized in these diverse fields.

C. The Analysis of Research Outputs for Dataset Groups

Our KG enables access to research outputs not only for individual datasets but also for groups of datasets linked by common metadata attributes, such as instrument, project, or mission. These metadata attributes are part of the CMR dataset metadata and are organized according to the GCMD Science Keywords dictionary. In our KG, these attributes are represented as GCMD Science Keyword nodes. By querying these nodes, we can retrieve research outputs for projects, instruments, and missions associated with the datasets.

Fig. 7 and Fig. 8 illustrate the research outputs for datasets produced by the Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2) project [20]. In the KG, MERRA-2 is represented as a model and categorized under the 'Platform' hierarchy within the GCMD Science Keywords. MERRA-2, produced by NASA's Global Modeling and Assimilation Office, is a NASA atmospheric reanalysis of the modern satellite era. It consists of over 100 long-term global reanalysis datasets that assimilate space-based observations of atmospheric chemistry components. With over 40 years of temporal coverage, MERRA-2 datasets are highly popular for applied research on atmospheric and ocean phenomena as well as air quality.

Since MERRA-2 products were published in 2017, over 2,400 publications have been collected citing these datasets. Querying our KG for publications across the entire MERRA-2 project or platform provides valuable insights for the dataset creators on how their products are being used in research. The histogram in Fig. 7 shows the distribution of applied research areas in publications citing MERRA-2 datasets. It reveals that 609, or 25%, of the collected MERRA-2 dataset-citing publications are not focused on applied research. These publications are primarily related to the validation and development of MERRA-2 datasets.

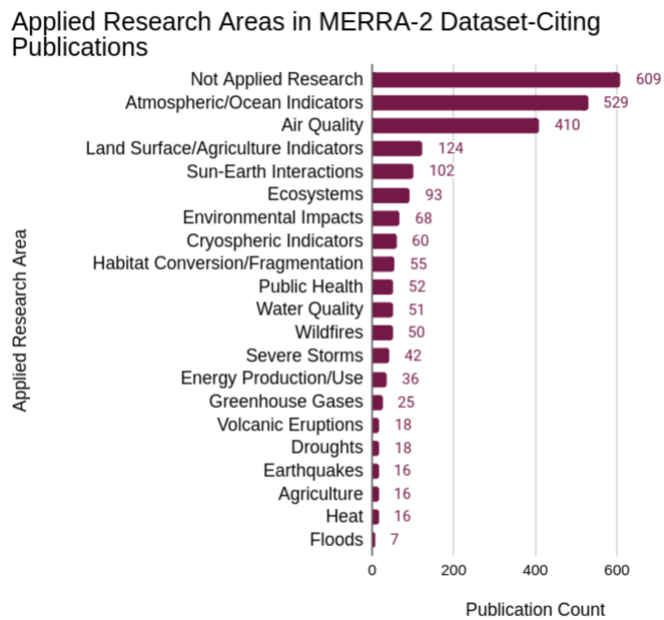

Fig. 7. The distribution of applied research areas in 2,400 MERRA-2 dataset-citing publications.

The most popular applied research areas for MERRA-2 datasets, such as "Atmospheric/Ocean Indicators" (22% of all MERRA-2 publications) and "Air Quality" (17% of all MERRA-2 publications), align with expectations given the focus of MERRA-2 data on atmospheric chemistry and dynamics. However, the histogram also highlights unexpected applied research areas, such as "Cryospheric Indicators", "Water Quality", and "Agriculture", which may not be directly related to the core content of the datasets. The KG enables dataset providers to identify and examine these publications to better understand how the data were used in those areas of research.

The stacked histogram in Fig. 8 highlights the usage of the 20 most cited individual MERRA-2 datasets in applied research. For each dataset, citation counts of that dataset are stacked according to assignment of citing publication to a specific applied research area. This visualization provides valuable insights for MERRA-2 data providers, helping them make informed decisions about future product development. By analyzing the popularity of specific dataset variables, as well as their arrangements by time frequency, horizontal and vertical groupings, and other factors, providers can prioritize enhancements or modifications. Additionally, this histogram serves as a useful tool for researchers, allowing them to evaluate which MERRA-2 products are most commonly utilized in similar applications, aiding in their decision-making when selecting datasets for their own research.

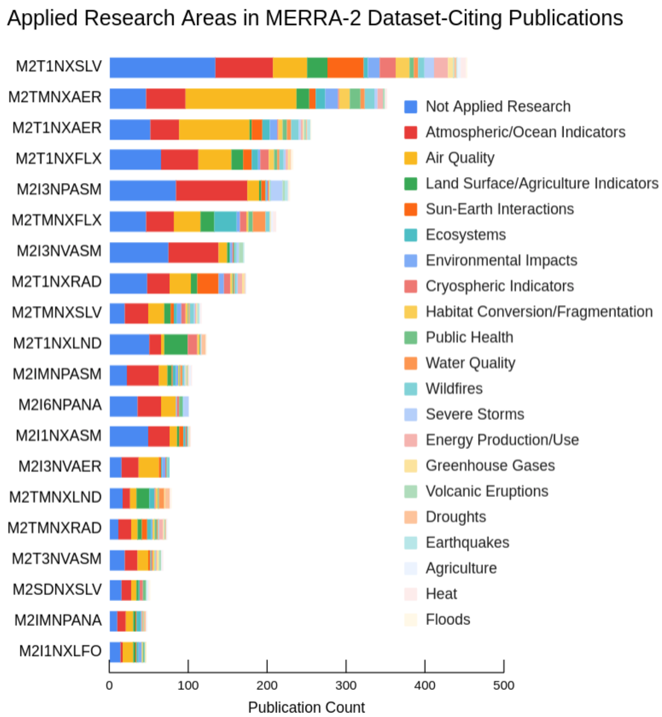

Fig. 8. Stacked histogram of MERRA-2 dataset-citing publications' applied research areas, arranged by 20 individual most cited MERRA-2 datasets.

IV. Operational Publication Management System

The developed dataset-publications KG enhanced with LLM content classification is utilized at one of the NASA DAACs, Goddard Earth Sciences Data and Information Services Center (GES DISC), for publication management and navigation. GES DISC maintains an extensive library of dataset-citing publications, currently containing over 12,000 citations. The details about the compilation of this library can be found in [21] and the GES DISC citation library can be found at [22].

All collected publications and dataset DOIs that they cite are integrated into a KG, where dataset DOIs are matched to the current CMR data collections. Corresponding CMR data collection metadata is retrieved from the CMR for creating KG nodes for dataset attributes such as instruments, mission,

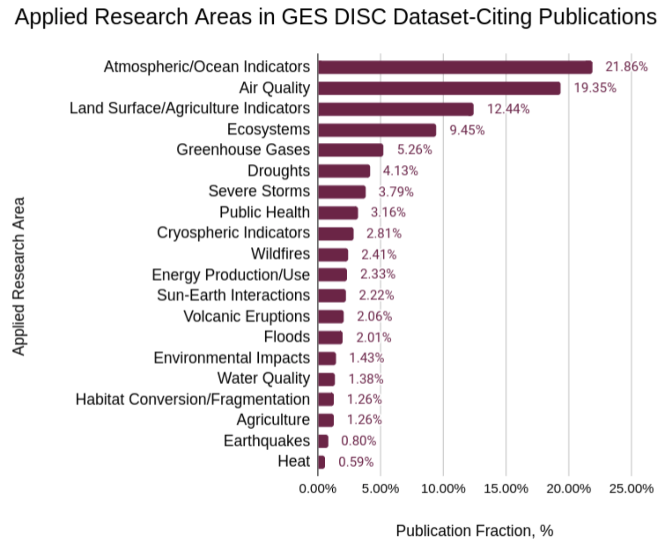

Fig. 9. Distribution of applied research areas in GES DISC dataset-citing publications.

projects, and science keywords with all of these attributes using the GCMD controlled vocabulary. This KG is continuously updated to add new dataset DOIs and versions as well as publications. The latest copy of this KG can be retrieved from (Mehrabian, 2024). The applied research areas for the GES DISC publications derived from LLM application are added to this KG also in the form of GCMD keywords. The distribution of applied research areas for the GES DISC dataset-citing publications is shown in Fig. 9. Comparing the results of Fig. 9 and Fig. 3, the latter for overall EOSDIS datasets, it can be seen that majority of GES DISC publications fall into 'Atmospheric/Ocean Indicators' and 'Air Quality' which is consistent with the majority of the GES DISC datasets containing data related to Atmospheric Composition, Water and Energy Cycles, and Climate Variability.

The collected publications' bibliography metadata, along with their derived applied research areas and linkage to the datasets, are presented in the GES DISC web interface in two ways. Each dataset DOI has a landing page, a web page where the DOI is resolved by the DOI server. On this page, all publications collected for all versions of the dataset throughout its lifetime are displayed. Fig. 10 shows an example of the dataset DOI landing page for the 'GLDAS_NOAH025_M' dataset, version 2.1 [23]. This page includes a description of the dataset and two types of references: 'Data Collection References' which pertain to the creation and/or validation of the dataset itself, and 'Related Publications' which are references to publications that have used the dataset. The provided disclaimer states, "The majority of publications using GES DISC data do NOT include the data version id. As such, publications below may not use the most recent processing version of the data." This disclaimer ensures that readers do not assume a specific version of the dataset was used in the cited papers, as some papers may not explicitly indicate the dataset version.

Users can consult the 'Version History' tab on the dataset landing web page to view the timeline of dataset versions. Each publication reference includes URLs to DSL pages of linked datasets. In Fig. 10, one 'Related Publication' is shown with the applied research area 'Wildfires,' classified by the fine-tuned INDUS LLM. This reference also lists two 'Related Data Collections' of the 'GLDAS_NOAH025_M' dataset, both currently available in the CMR. The URLs of these collections link directly to the landing pages of each data collection DOI.

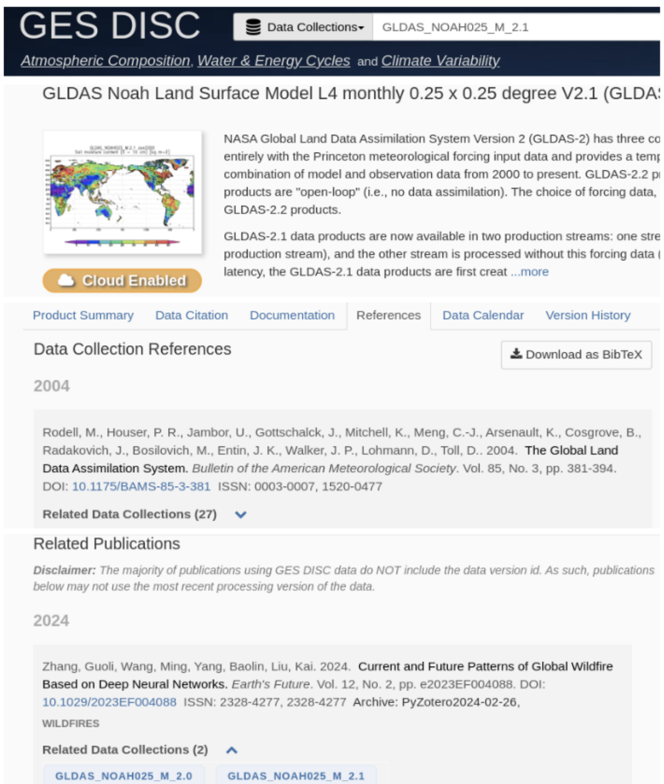

Fig. 10. DOI landing page for GLDAS Noah Land Surface Model L4 monthly 0.25 x 0.25 degree (GLDAS_NOAH025_M) V2.1 dataset [23].

At the GES DISC website, document citations are also available on a dedicated publications search interface, as shown in Fig. 11. Similar to the dataset landing pages, each reference includes URLs to the corresponding dataset landing pages. To facilitate browsing by applied research area, the publications search interface features a facet displaying twenty applied research areas from which users can select.

V. Conclusions

We presented a design of a KG augmented with fine-tuned LLM outputs, enabling seamless integration of dataset's application research areas for enhanced dataset discovery and search. Additionally, we highlighted the importance of research contributions from prior dataset versions for improving the search and discovery of publicly available versions of datasets and maintaining the provenance of research outputs

Fig. 11. Snapshot of GES DISC Publications interface with selected applied research area 'Droughts'.

throughout the dataset's life cycle. We also demonstrated how our operational publication management system leverages the KG-LLM design to improve navigation of over 12,000 dataset-citing publications at GES DISC, one of NASA's Earth science data archives.

As shown in the analysis of MERRA-2 research outputs, our KG augmented with LLM workflow has proven to be an invaluable tool for automating the analysis of dataset-citing publications, providing significant benefits to both data providers and users. Queries to the KG enable quick access to publications associated with entire projects, instruments, or missions, as these elements are all interconnected within the graph. By combining research publications throughout a dataset's lifetime, we can extract temporal trends in dataset usage and gain a comprehensive understanding of its impact in various application areas.

Based on our experience, we urge data generators and providers to ensure the provenance of dataset versions and store this provenance information in machine-accessible systems, such as dataset registries, that facilitate automated tracking of data provenance.

VI. Data and Code Availability

The GES DISC dataset-publication KG is available in [24]. The finetuned NASA IMPACT INDUS LLM is available at https://huggingface.co/arminmehrabian/nasa-impact-nasa-smd-ibm-st-v2-classification-finetuned. EOSDIS and GES DISC publications libraries are available from [18] and [22], correspondingly.

Acknowledgment

We greatly appreciate and acknowledge the contribution of Chris Gerasimov for his assistance with paper formatting.

The authors also extend their sincere appreciation to the anonymous reviewers for their thoughtful comments and suggestions, which have significantly improved this manuscript.

References

[1] J. Behnke, A. Mitchell, and H. Ramapriyan, "NASA's Earth observing data and information system–near-term challenges", *Data Science Journal*, 18, pp. 40-40, 2019, doi: 10.5334/dsj-2019-040.

[2] SPD-41a - Scientific Information Policy for NASA's Science Mission Directorate, 2022. Accessed July 15, 2024 [Online.] Available: https://science.nasa.gov/researchers/open-science/science-information-policy/

[3] CODATA-ICSTI - Task Group on Data Citation Standards and Practices, "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data", *Data Science Journal*, 12, pp. 75, 2013, doi: 10.2481/dsj.OSOM13-043

[4] M. Zhao, E. Yan, and K. Li, "Data set mentions and citations: A content analysis of full-text publications", *Journal of the Association for Information Science and Technology*, 69(1), 32–46, 2018, doi: 10.1002/asi.23919

[5] I. Gerasimov, B. KC, A. Mehrabian, J. Acker, and M. P. McGuire, "Comparison of datasets citation coverage in Google Scholar, Web of Science, Scopus, Crossref, and DataCite", *Scientometrics*, (2024), doi: 10.1007/s11192-024-05073-5

[6] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying Large Language Models and Knowledge Graphs: A Roadmap", *IEEE Transactions on Knowledge and Data Engineering*, 36(7), pp. 3580–3599, 2024, doi: 10.1109/TKDE.2024.3352100

[7] X. Ma, "Knowledge graph construction and application in geosciences: A review", *Computers & Geosciences*, 161, 105082, 2022, doi: 10.1016/j.cageo.2022.105082

[8] D. Brickley, M. Burgess, and N. Noy, "Google Dataset Search: Building a search engine for datasets in an open Web ecosystem", *The World Wide Web Conference*, pp. 1365–1375, 2019, doi: 10.1145/3308558.3313685

[9] M. A. Parsons, R. Duerr, and O. Godoy, "The evolution of a geoscience standard: An instructive tale of science keyword development and adoption", *Geoscience Frontiers*, 14(5), p. 101400, 2023, doi: 10.1016/j.gsf.2022.101400

[10] R. G. Raskin, and M. J. Pan, "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)", *Computers & Geosciences*, 31(9), pp. 1119–1125, 2005, doi: 10.1016/j.cageo.2004.12.004

[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, Improving language understanding by generative pre-training, 2018. Accessed July 15, 2024 [Online.] Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", (arXiv:1810.04805). arXiv, 2019, doi: 10.48550/arXiv.1810.04805

[13] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text", In K. Inui, J. Jiang, V. Ng, and X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3615–3620). Association for Computational Linguistics, 2019, doi: 10.18653/v1/D19-1371

[14] B. Bhattacharjee, A. Trivedi, M. Muraoka, M. Ramasubramanian, T. Udagawa, I. Gurung, R. Zhang, B. Dandala, R. Ramachandran, M. Maskey, K. Bugbee, M. Little, E. Fancher, L. Sanders, S. Costes, S. Blanco-Cuaresma, K. Lockhart, T. Allen, F. Grezes, … T. Lee, "INDUS: Effective and Efficient Language Models for Scientific Applications", (arXiv:2405.10725). arXiv, 2024, doi: 10.48550/arXiv.2405.10725

[15] Common Metadata Repository, 2019. Accessed July 19, 2024 [Online.] Available: https://www.earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/cmr

[16] J. Brase, "DataCite - A Global Registration Agency for Research Data", *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, Beijing, 2009, pp. 257-261, 2009, doi: 10.1109/COINFO.2009.66.

[17] DataCite Schema, 2024. Accessed July 15, 2024 [Online.] Available: https://schema.datacite.org/meta/kernel-4.5/

[18] EOSDIS Publications Library, 2024. Accessed July 15, 2024 [Online.] Available: https://www.zotero.org/groups/4567966/eosdis_dci/library

[19] B. Smith, S. Adusumilli, B. M. Csathó, D. Felikson, Fricker, H. A., Gardner, A., Holschuh, N., Lee, J., Nilsson, J., Paolo, F. S., Siegfried, M. R., Sutterley, T., and the ICESat-2 Science Team. ATLAS/ICESat-2 L3A Land Ice Height, Version 6 [Dataset]. Boulder, Colorado USA, NASA National Snow and Ice Data Center Distributed Active Archive Center, 2023, doi: 10.5067/ATLAS/ATL06.006

[20] MERRA-2 - Project Description at GES DISC, 2024. Accessed July 15, 2024 [Online.] Available: https://disc.gsfc.nasa.gov/information/mission-project?title=MERRA-2

[21] I. Gerasimov, A. Savtchenko, J. Alfred, J. Acker, J. Wei, and B. KC, "Bridging the Gap: Enhancing Prominence and Provenance of NASA Datasets in Research Publications", *Data Science Journal*, 23(1), (2024), doi: 10.5334/dsj-2024-001

[22] GES DISC Publications Library, 2024. Accessed July 15, 2024 [Online.] Available: https://www.zotero.org/groups/2395775/ges_disc/library

[23] H. Beaudoing and M. Rodell, "GLDAS Noah Land Surface Model L4 monthly 0.25 x 0.25 degree V2.1," [Dataset], Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), 2020, doi: 10.5067/SXAVCZFAQLNO

[24] A. Mehrabian, I. Gerasimov, "NASA GES-DISC Knowledge Graph for Link Prediction" (1.0) [Dataset]. Zenodo. 2024, doi: 10.5281/zenodo.11492533