Enhancing NASA Earth Science Data Discovery from Scientific Publications

AGU Fall 2020
December 1-17, 2020
Irina.Gerasimov@nasa.gov

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

Irina Gerasimov[1,2], Andrey Savtchenko[1,2], Jim Acker[1,2], Jerome Alfred[1,2], and Jennifer Wei[1]
[1]Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA [2]ADNET Systems Inc., Lanham, MD, USA

Creating searchable publication and data linked library

- GES DISC data archive contains around ~1300 publicly available data collections and thousands of papers are being published based on those datasets.
- These scientific papers contain valuable information about: studied phenomena, research methods applied to the datasets, groups of the datasets used together and particular dataset measurements. User exploring the data center datasets can use this information for better understanding which datasets are best fit for their applications. Having publications information translated and stored in searchable format can facilitate usage-based dataset discovery and enhance **Findability, Accessibility, Interoperability and Reusability** principles of data management.
- Our approach to enhance usage-based dataset discovery consists of:
  - **Collecting publications citations that use GES DISC datasets**. This is an ongoing work that takes place at GES DISC. Citations are collected from the data science teams and online sources (Google Scholar, Scopus, Web of Science, DataCite) via data collection DOI and keywords search.
  - **Uniform term extraction** from publication title and abstracts using Semantic Web for Earth and Environment Technology (SWEET) and Global Change Master Directory (GCMD) ontologies.
  - **Indexing citations with:** GES DISC dataset and variable names, locations and time ranges of the research, instruments and missions, science keywords, phenomena and properties.

Term extraction from text - Leveraging Earth Science ontologies

Research citations are collected from hundreds of different publishing sources, some of them have keywords and most do not. Keywords are free text assigned by individuals. Indexing all citations by uniform approach provides single set of keywords helping organize and search citations in uniform way: by measurement, phenomena, location, variables and datasets.

GCMD ontology -- locations, instrument, missions and science keywords.
- Covers all GES DISC instruments and missions.
- High level location coverage: countries and geographical features
- Science keywords -- decent coverage of Earth Science dataset measurements and high level variables. Used for GES DISC dataset metadata. It is a mix of phenomena, realms and measurements.

SWEET ontology - phenomena, property, process, material, ream, state, representation, human.
- Explicit coverage of Earth Science phenomena, properties, realms and materials
- Some terms erms on their own often do not carry much content (e.g. system, variability, event, etc.) but separate terms are easier detected in the text.
- Combination of terms more comprehensively reflects content of the text.

Difference and similarity of GCMD and SWEET ontology terms extracted from publication text:

GCMD science keyword: planetary boundary layer height
SWEET Realm SWEET Property

GCMD science keyword: none
planetary boundary layer SO2 column concentration
SWEET Realm SWEET Material SWEET Property

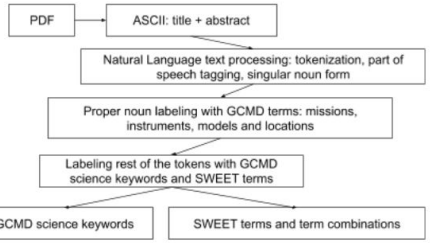
GCMD science keyword: methane
methane seasonal variation
SWEET Material SWEET state SWEET Phenomena

GCMD science keyword: methane
methane total column abundance
SWEET Material SWEET state SWEET Representation SWEET Property

Term extraction from publications

As a use case, a collection of publications that utilize popular Web service, NASA Giovanni (https://giovanni.gsfc.nasa.gov/), developed at GES DISC was analyzed. These publications use Giovanni service in various ways: data visualization and analysis, data download, sample image snapshots, etc. In addition, Giovanni services not only GES DISC datasets but also data from the other NASA data archives. With Giovanni service being in place since 2000, total of 2,300 publications were collected so far, out of which ~1000 were analyzed for publication years 2016-2020. Publication processing consisted of the following steps:

- **Extract** text excerpts that mention Giovanni service:
  - Using Giovanni database and text determine which variables were used.
  - Based on variable, instrument/mission/model names and research coverage determine names of the data collections the variables were derived from.
- **Process** titles and abstracts to retrieve:
  - Filter out from the text names of missions, instruments, models, and locations.
  - Extract GCMD keywords and SWEET terms. Create combinations of SWEET terms.

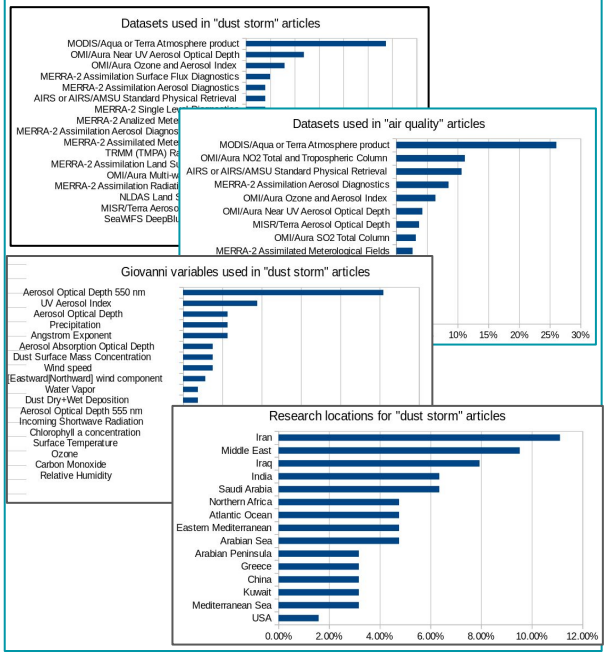


Te
Ti comparative analysis of Stromboli, Batu Tara and Tinakula volcanoes.
In addition to bibliographic tags collected from the source and stored in citation management system, following tags are added:

Giovanni variables	Cloud Fraction
Data collections:	OMI/Aura total column SO2 gridded product, MODIS/Aqua or Terra Atmospheric product
Cited datasets	OMI/Aura total column SO2 gridded product (10.5067/A
Locations	Indonesia
Research coverage	2000 - 201
GCMD keywords	Emission
SWEET term combinations	depth acc eruption, n period, pe SO2 mass intensity in

Data discovery use case

Examples below demonstrate types of the information that could be retrieved from the publication texts. Figures 1 and 2 present the proportion of various datasets used in the articles that mention "dust storm" and "air quality" phenomena in their abstracts and titles. Figure 3 presents the proportion of the Giovanni variables and Figure 4 shows the proportion of locations of articles that mention "dust storm".



Future Work

- Work with GCMD and SWEET communities to add terms that belong to these ontologies categories and necessary for reflecting text content.
- Add terms to the citation management system and implement advanced search over these term categories.
- Build a knowledge graph based on the retrieved terms to connect citations, datasets and terms to facilitate data discovery.