Linking Publications Knowledge and Data in Earth Sciences

ESIP Summer 2020
July 14-24, 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

Enabling Usage-Based Dataset Discovery

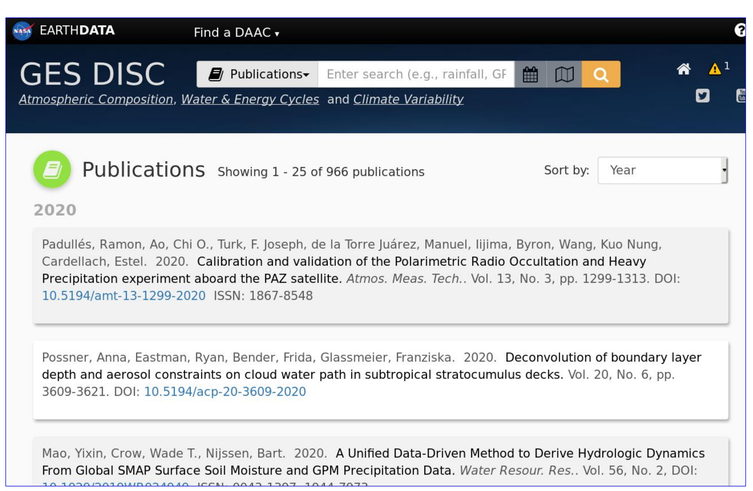
- GES DISC data archive contains around ~1300 publicly available data collections and thousands of papers are being published based on those datasets.
- These scientific papers contain valuable information about where, when and how the datasets were used for the research. Having this information translated and stored in searchable format can facilitate usage-based dataset discovery and enhance **FAIR** data management in two ways:
  - Making data **Findable** - enabling data discovery through connecting publications knowledge with the data.
  - Making data **Reusable** - associating prior knowledge with data used to produce and thus letting future scientists to reproduce this knowledge.
- Our approach to enable usage-based dataset discovery consists of:
  - **Collecting publications citations that use GES DISC datasets**. This is a recently started ongoing work that takes place at GES DISC.
  - **Creating a knowledge base** that stores publication content along with associated datasets in a searchable form. Since publications are written in free text, straight forward text-to-database translation is not possible. Our approach to this is mapping paper text to a dictionary that consists of Earth science domain ontology terms. This makes a knowledge base content uniform across all publications.

Collecting Publications Citations at GES DISC

The publications citations at GES DISC are collected from several sources:

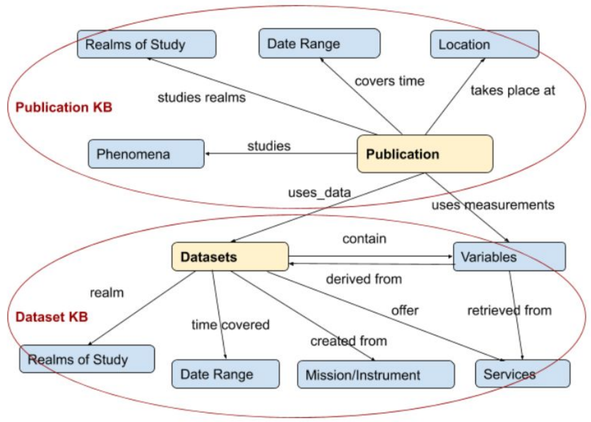
- Scopus, DataCite and Web of Science APIs allow automated or semi-automated citation retrieval using dataset DOIs.
- Many of GES DISC data providers maintain collections of paper citations that use their data. In addition, the providers supply GES DISC with the most important citations of papers about dataset algorithms and validation.
- Collecting citations that mention using GES DISC Giovanni service for data analysis and/or download.
- Establishing various subscriptions from online libraries to harvest citations that mention GES DISC data.

The citations that do not have automatically assigned dataset DOIs are manually reviewed to determine names of the datasets that were used in the research described by the paper.
To-date GES DISC citation collection contains around 1000 citations that are associated with particular datasets and this collection keeps growing.



Creating Publication-Dataset Knowledge Base

- A research publication contains information about used datasets, measurements, locations, phenomena, dates of conducted research, all together formulating a **publication Knowledge Base (KB)**.
- A **dataset Knowledge Base** consists of dataset instruments and missions, variables, covered time range, discipline, science keywords, as well as available services.
- Publications KB extends dataset KB by adding information about how this dataset was used by researchers. Together they form a **Publication-Dataset KB** which can be represented as **a map of linked concepts** of phenomena, realms, variables, time ranges, locations and datasets:



Major challenge is extracting uniform KB terms from publications because publications are written in free text.

Approach: using domain ontology terms as a dictionary.

Use case: Giovanni publications

"Giovanni" publications authors use GES DISC Giovanni service (https://giovanni.gsfc.nasa.gov/) either for data visualization and analysis and/or for data download. Most of these publications study various applications many of which can be described in terms of phenomena.

Extracting publication KB terms from Giovanni's publications consist of the following steps:

- **Extract** text excerpts that mention Giovanni service along with used variables, platforms, instruments and other sources.
- **Derive** Giovanni variable names and datasets those variables were derived from using text excerpts, content of Giovanni database and heuristics.
- **Retrieve** date ranges and places of the subject of study.
- **Extract** SWEET ontology Phenomena and Realm category terms from publication title and abstract.
- **Review** extracted terms to retain only the ones that reflect publication content.

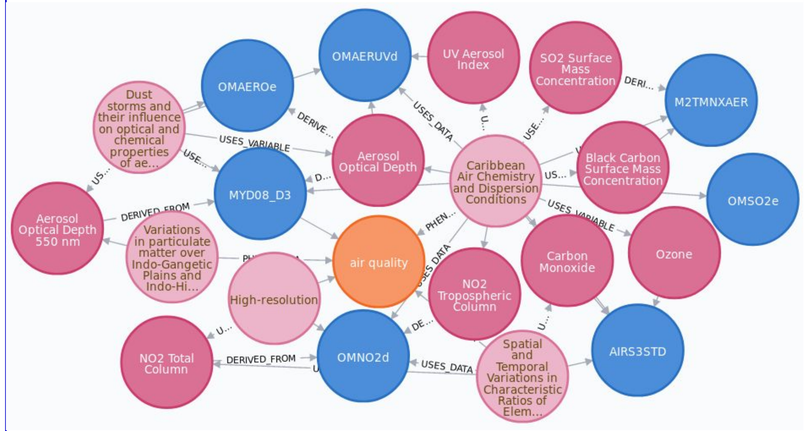
Processing 186 publications "Giovanni" from 2017 produced: 113 datasets, 75 variables, 191 phenomena, 72 realms, 160 places and 29 dates (for start and end)

Future Work

- Automate dataset name and term extraction from the publications using ML tools.
- Complete creating dataset-publication KB for Giovanni publications.
- Create dataset-publication KB for entire GES DISC publications library.
- Utilize created publication-dataset KB for usage-based dataset discovery.

Giovanni Publication - Dataset KB

Datasets and variables used in publications studying "Air Quality" phenomena:



Legend:

- Publication node with the publication title
- Variable node with Giovanni variable name
- Phenomena node (SWEET term)
- Dataset node with the dataset short name
- Place node with the place name

Phenomena (above) and places (below) studied in publications that use OMI/Aura NO2 Cloud-Screened Total and Tropospheric Column L3 Global Gridded 0.25 degree x 0.25 degree (OMNO2d) dataset: