Improving Earth Science dataset search with publications content via Knowledge Graph linkage

ESIP Summer 2021
July 19-23, 2021
kristina.a.stoyanova@nasa.gov

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

Kristina Stoyanova[1,2], Irina Gerasimov[1,2], Armin Mehrabian[1,2], Jennifer Wei[1], and Mohammad Khayat[1,2]
[1]Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA [2]ADNET Systems Inc., Lanham, MD, USA

Abstract and Purpose

The NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) archives a large number of Earth observational datasets. Thousands of the publications are created each year based on these datasets. The content of these publications can be used for discovery of the datasets based on the characteristics of applicational research. We leverage the content of these publications to retrieve the information about phenomena and domains where measurements from the datasets were utilized through linking these publications and dataset in Knowledge Graph. We retrieve phenomena and domain information using SWEET (Semantic Web for Earth and Environmental Terminology) ontology and produce the set of keywords that are linked to the datasets. Further, we evaluate this link strength according to the frequency of dataset usage in the papers mentioning these keywords. We demonstrate how this linkage can improve dataset search by comparing the search results obtained from the Common Metadata Repository (CMR) search and publications based data.

CMR Search and Knowledge Graph Search

CMR Search:

- NASA stores databases of experiments and measurements from satellites in the Earth Observing System Data and Information System. When looking up datasets related to a word, CMR free text search goes through all collection metadata, including science and ancillary keywords, and abstracts, to find related datasets. A dataset is returned if any part of it's collection metadata contains the search term.
- The main issue is there are many search terms which the collection returns nothing because the metadata does not have those terms. We are seeking to fix this issue.

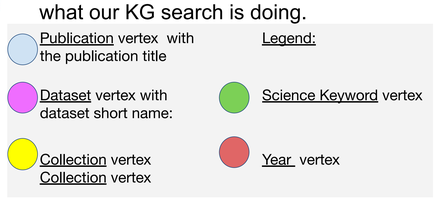
Knowledge Graph (KG) Search

- GES DISC maintains citation management system, Zotero, where it collects publications related to GES DISC datasets
- For the search we used a collection of ~1200 papers from 2016 to 2021 referencing NASA Giovanni service that provides visualization and analysis for the most popular GES DISC datasets.
- Thus, if a term appears in a publication, it can be linked to the datasets that publication uses. That is the relationship that we want in the knowledge graph.

Terms Creating Publication-Dataset Knowledge Graph (KG) Base

Sample Gremlin Query Graph for a publication:

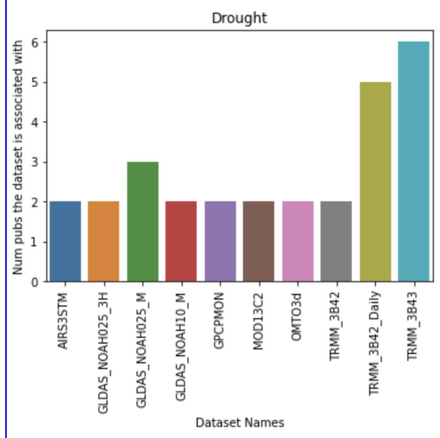
- Create relevant vertices
  - Ex: Publications, Datasets, Science Keywords
- Edges connect vertices
  - Ex: CreatedBy Edges
- Our KG abstract and title search provides an insight how the full knowledge graph can help us to improve the search.
- The publication vertex may have an attribute of a title or abstract that contains an ontology term, which can then connect that ontology term to a dataset. Which is what our KG search is doing.



SWEET Ontology

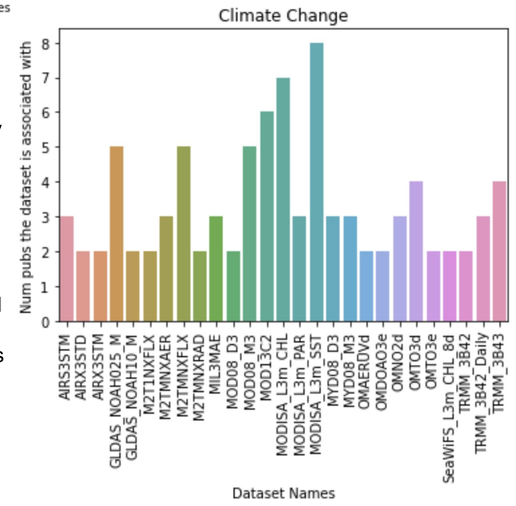
- An ontology of the Earth science concepts - we used it as a dictionary of terms describing various phenomena.
- We will use the SWEET ontology as a dictionary of earth science terms that scientists might look up when searching datasets.
- We chose to look at the terms for:
  - Phenomena Atmosphere Precipitation ('thunderstorm', 'tornado', 'tropical storm', 'hurricane')
  - Phenomena Environmental Impact ('spill', 'toxicity', 'water pollution', 'water quality')
  - Phenomena Planetary Climate ('microclimate', 'global change', 'drought', 'heat island')
- We will compare CMR vs Knowledge Graph search results on the same SWEET terms

Dataset and term co-appearances in publications titles and abstracts



- KG on title and abstracts for the Term "Drought" from Phenomena Planetary Climate.
  - From 2016 - 2021 giovanni reviewed, 19 Publications contained this term.
  - 28 unique datasets associated with these publications.
- Frequency of dataset co-appearance with the term is the measure of association strength between term and the dataset

- Term "Climate Change" from Phenomena Planetary Climate
- 50 Publications
- 65 unique datasets associated with these publications
- From 2016 - 2021
- Enabling usage based discovery: search for datasets in paper titles and abstracts by data usage terms.

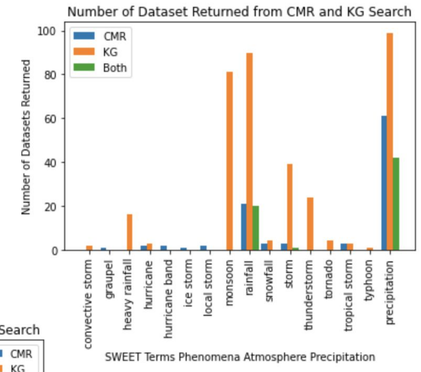


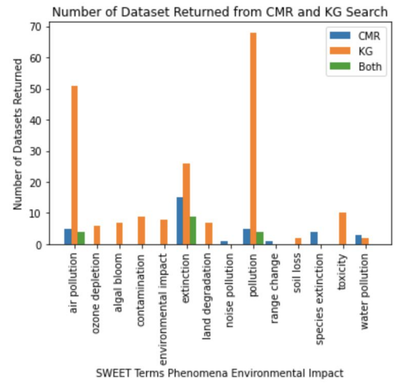
KG search returns not only the number of unique datasets in publications that have the term in their title or abstract, but also the number of times each dataset are used in multiple publications. The count of number times a dataset is used in a publication can be used as weights in the graph for usage based dataset discovery.

SWEET Search Results: CMR vs KG

Average amount Datasets returned for each SWEET category

- We compared CMR and the KG on 48 SWEET terms
- Overall the KG returned more datasets than CMR
- For most of the terms KG returns a set of datasets that include all the ones from CMR

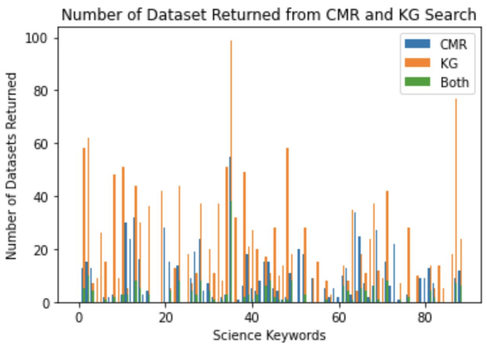




- The knowledge graph title and abstract search captures more term to dataset relationships that can enhance CMR search
- Applying this KG to the CMR search can return results for words that we were previously not able to query on CMR

Science Keyword Search: CMR vs KG

- We also compared the KG search and CMR search on 90 Scientific Keywords from the KG, which are words scientists have created to describe CMR datasets.
- CMR search normally works with science keywords, it is expected that CMR does well, but KG managed to return many different datasets



Average KG: 17.8 datasets
Average CMR: 19.7 datasets
On Average the KG returned 90% of the datasets that CMR returned.

Outcomes and Future Work

- Searching through Publication Titles and Abstracts for ontology terms and then returning the corresponding datasets shows significant search improvement over normal CMR search.
- Co-appearance of terms and datasets in publications allow us to weigh the term-to-dataset connection and help to rank the search results.
- Our full Knowledge Graph will be similar to this publications search and even more informative as we have other kinds of relationships that can affect the search.