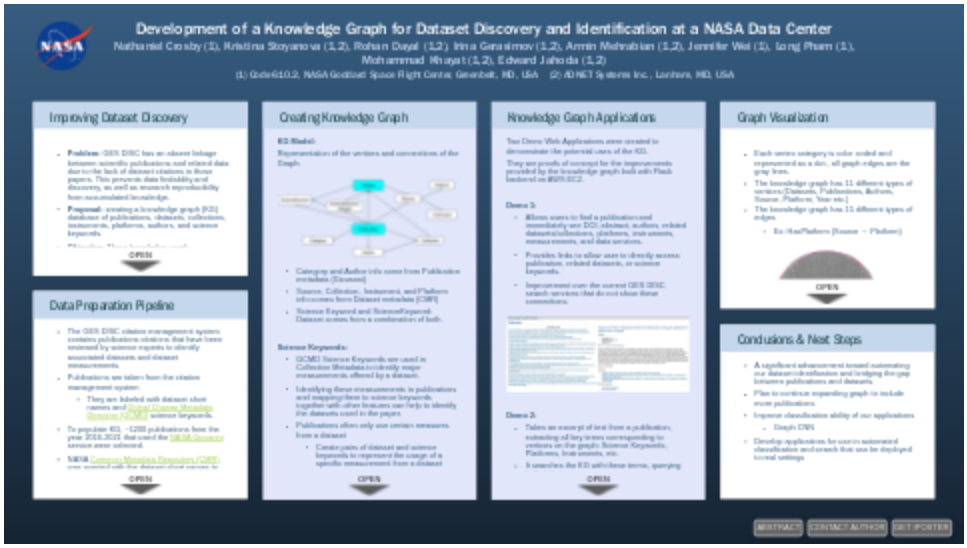
Development of a Knowledge Graph for Dataset Discovery and Identification at a NASA Data Center



Nathaniel Crosby (1), Kristina Stoyanova (1,2), Rohan Dayal (1,2), Irina Gerasimov (1,2), Armin Mehrabian (1,2), Jennifer Wei (1), Long Pham (1), Mohammad Khayat (1,2), Edward Jahoda (1,2)

(1) Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA (2) ADNET Systems Inc., Lanham, MD, USA

PRESENTED AT:

IMPROVING DATASET DISCOVERY

- **Problem:** GES DISC has an absent linkage between scientific publications and related data due to the lack of dataset citations in those papers. This prevents data findability and discovery, as well as research reproducibility from accumulated knowledge.

- **Proposal:** creating a knowledge graph (KG) database of publications, datasets, collections, instruments, platforms, authors, and science keywords.

- **Objective:** These knowledge graph relationships can help to identify datasets in the paper texts, discover the datasets through publication search, and improve search for the datasets in the data center.

DATA PREPARATION PIPELINE

- The GES DISC citation management system contains publications citations that have been reviewed by science experts to identify associated datasets and dataset measurements.

- Publications are taken from the citation management system

  - They are labeled with dataset short names and Global Change Metadata Directory (GCMD) (https://gcmdservices.gsfc.nasa.gov/KeywordViewer/) science keywords.

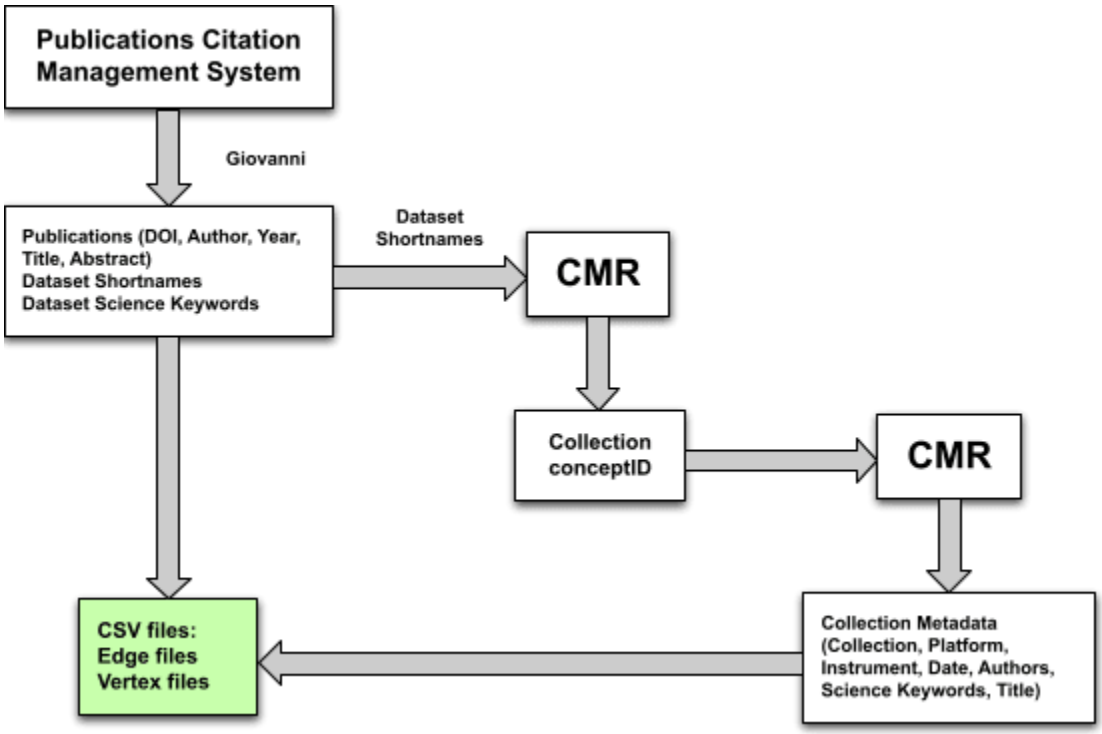
- To populate KG, ~1200 publications from the year 2016-2021 that used the NASA Giovanni (https://giovanni.gsfc.nasa.gov/) service were selected.

- NASA Common Metadata Repository (CMR) (https://cmr.earthdata.nasa.gov/search) was queried with the dataset short names to gather information about collections currently available for the public search.

- Collected metadata for publications and datasets is broken into edge and vertex files, in that format required by AWS Neptune.

  - The vertex files contain the publication, dataset, collection, etc. information

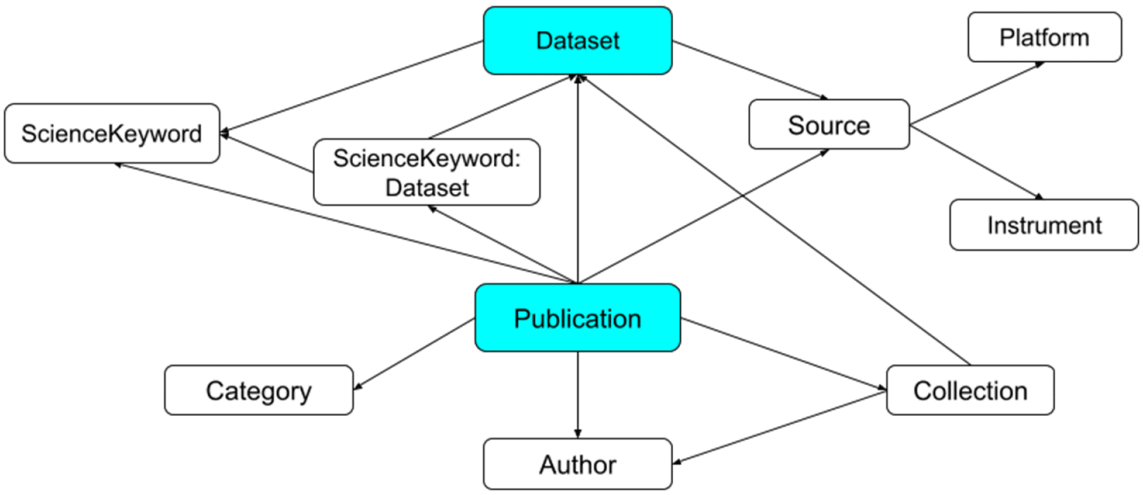
  - The edge files create a directed relationship between two vertices such as publication related to a dataset.



CREATING KNOWLEDGE GRAPH

KG Model:

Representation of the vertices and connections of the Graph:



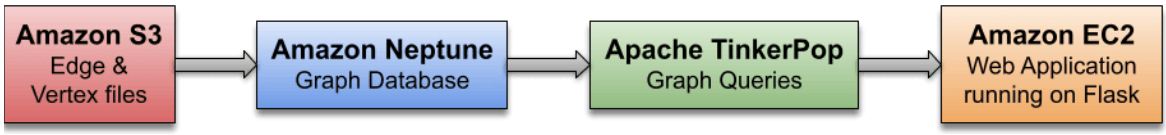
- Category and Author info come from Publication metadata (Giovanni)
- Source, Collection, Instrument, and Platform info comes from Dataset metadata (CMR)
- Science Keyword and ScienceKeyword: Dataset comes from a combination of both.

Science Keywords:

- GCMD Science Keywords are used in Collection Metadata to identify major measurements offered by a dataset.
- Identifying these measurements in publications and mapping them to science keywords together with other features can help to identify the datasets used in the paper.
- Publications often only use certain measures from a dataset
  - Create pairs of dataset and science keywords to represent the usage of a specific measurement from a dataset
  - Allows additional KG connections and better understanding which data was used for research.
- Science keywords in our KG follow GCMD keyword hierarchy
  - Allow for keyword search through graph traversal.

AWS Framework:

- AWS Neptune allows creation of graph databases.
- Gremlin Graph Traversal Language can quickly and easily query for specific vertices and edge connections.
- To create a graph in Neptune, an AWS S3 Bucket must store the vertex and edge CSV files created by our data pipeline.
- AWS also allows straightforward development of applications and services that use the graph database.



KNOWLEDGE GRAPH APPLICATIONS

Two Demo Web Applications were created to demonstrate the potential uses of the KG.

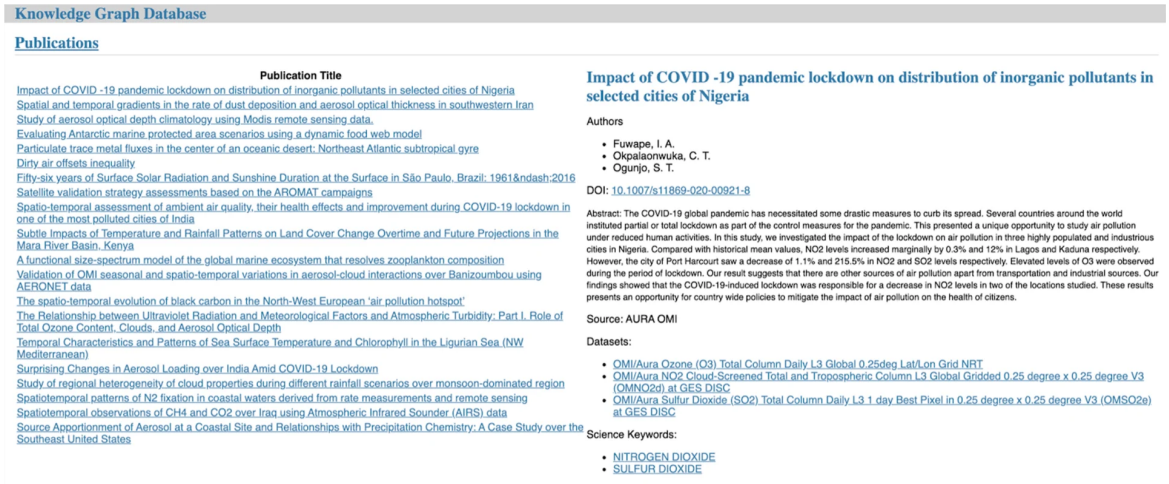
They are proofs of concept for the improvements provided by the knowledge graph built with Flask backend on AWS EC2.

Demo 1:

- Allows users to find a publication and immediately see DOI, abstract, authors, related datasets/collections, platforms, instruments, measurements, and data services.

- Provides links to allow user to directly access publication, related datasets, or science keywords.

- Improvement over the current GES DISC search services that do not show these connections.



Demo 2:

- Takes an excerpt of text from a publication, extracting all key terms corresponding to vertices on the graph: Science Keywords, Platforms, Instruments, etc.

- It searches the KG with these terms, querying to find datasets.

  - If none of those key terms are found, it matches words in the titles and abstracts

- After querying, it returns the most likely datasets and science keywords associated with that publication.

  - This is based on which datasets and keywords appear most in our queries results and well as which are more popular in general (most connections)

- Important step toward automating our dataset classification.

Publications Knowledge Graph Database

Publications Dataset Science Keyword Search

Furthermore, the ozone monitoring instrument (OMI) time series of area-averaged methane (CH 4) mole fraction in the air for March–June 2020 over 20–21° N, 85–86° E shows no substantial change in CH 4 concentration during the observation period (Fig. S3).

Submit

Science Keywords:

- METHANE
- ATMOSPHERIC OZONE
- NITROGEN DIOXIDE
- AEROSOL EXTINCTION
- UV AEROSOL INDEX
- REFLECTANCE
- SULFUR DIOXIDE

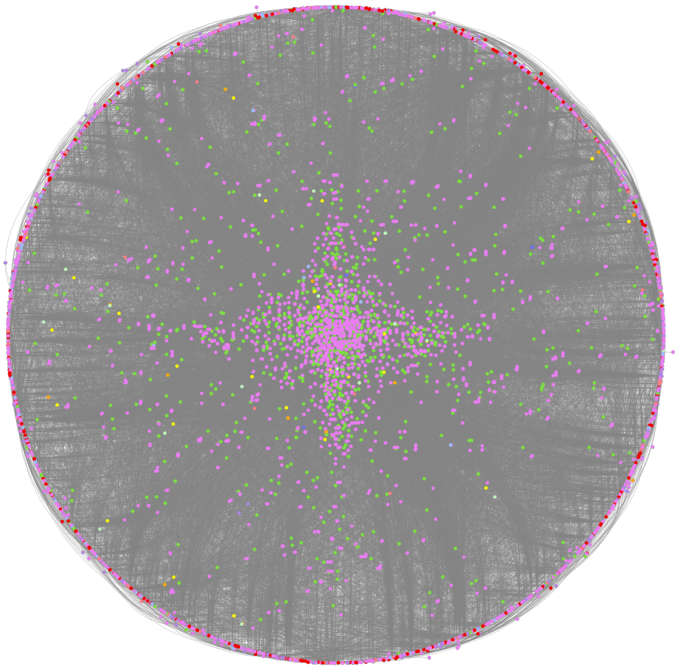
Dataset Short Names:

- OMTO3d : OMI/Aura TOMS-Like Ozone, Aerosol Index, Cloud Radiance Fraction L3 1 day 1 degree x 1 degree V3 (OMTO3d) at GES DISC
- OMNO2d : OMI/Aura NO2 Cloud-Screened Total and Tropospheric Column L3 Global Gridded 0.25 degree x 0.25 degree V3 (OMNO2d) at GES DISC
- OMAERUVd : OMI/Aura Near UV Aerosol Optical Depth and Single Scattering Albedo L3 1 day 1.0 degree x 1.0 degree V3 (OMAERUVd) at GES DISC
- OMSO2e : OMI/Aura Sulfur Dioxide (SO2) Total Column Daily L3 1 day Best Pixel in 0.25 degree x 0.25 degree V3 (OMSO2e) at GES DISC
- OMTO3e : OMI/Aura TOMS-Like Ozone and Radiative Cloud Fraction L3 1 day 0.25 degree x 0.25 degree V3 (OMTO3e) at GES DISC
- OMDOAO3e : OMI/Aura Ozone (O3) DOAS Total Column Daily L3 1 day 0.25 degree x 0.25 degree V3 (OMDOAO3e) at GES DISC
- OMAEROe : OMI/Aura Multi-wavelength Aerosol Optical Depth and Single Scattering Albedo L3 1 day Best Pixel in 0.25 degree x 0.25 degree V3 (OMAEROe) at GES DISC
- OMUVBd : OMI/Aura Surface UVB Irradiance and Erythemal Dose Daily L3 Global Gridded 1.0 degree x 1.0 degree V3 (OMUVBd) at GES DISC
- OMNO2 : OMI/Aura Nitrogen Dioxide (NO2) Total and Tropospheric Column 1-orbit L2 Swath 13x24 km V003 (OMNO2) at GES DISC
- OMHCHOd : OMI/Aura Formaldehyde (HCHO) Total Column Daily L3 Weighted Mean Global 0.1deg Lat/Lon Grid V003 (OMHCHOd) at GES DISC
- OMSO2G : OMI/Aura Sulphur Dioxide (SO2) Total Column Daily L2 Global Gridded 0.125 degree x 0.125 degree V3 (OMSO2G) at GES DISC
- OMAERO : OMI/Aura Multi-wavelength Aerosol Optical Depth and Single Scattering Albedo 1-orbit L2 Swath 13x24 km V003 (OMAERO) at GES DISC
- OMNO2G : OMI/Aura NO2 Total and Tropospheric Column Daily L2 Global Gridded 0.25 degree x 0.25 degree V3 (OMNO2G) at GES DISC
- OMAERUV : OMI/Aura Near UV Aerosol Optical Depth and Single Scattering Albedo 1-orbit L2 Swath 13x24 km V003 (OMAERUV) at GES DISC
- OMI_MINDS_NO2 : OMI/Aura NO2 Tropospheric, Stratospheric & Total Columns MINDS 1-Orbit L2 Swath 13 km x 24 km V1 (OMI_MINDS_NO2) at GES DISC
- OMI_MINDS_NO2d : OMI/Aura NO2 Tropospheric, Stratospheric & Total Columns MINDS Daily L3 Global Gridded 0.25 degree x 0.25 degree V1 (OMI_MINDS_NO2d) at GES DISC
- OMTO3 : OMI/Aura Ozone(O3) Total Column 1-Orbit L2 Swath 13x24 km V003 (OMTO3) at GES DISC
- OMHCHO : OMI/Aura Formaldehyde (HCHO) Total Column 1-orbit L2 Swath 13x24 km V003 (OMHCHO) at GES DISC
- OMSO2 : OMI/Aura Sulphur Dioxide (SO2) Total Column 1-orbit L2 Swath 13x24 km V003 (OMSO2) at GES DISC
- OMUVB : OMI/Aura Surface UV Irradiance 1-orbit L2 Swath 13x24 km V003 (OMUVB) at GES DISC

GRAPH VISUALIZATION

- Each vertex category is color coded and represented as a dot., all graph edges are the gray lines.

- The knowledge graph has 11 different types of vertices (Datasets, Publications, Authors, Source, Platform, Year etc.)
- The knowledge graph has 15 different types of edges
    - Ex: HasPlatform (Source → Platform)



- Below is a publication vertex with all immediate connections.
- The paper is connected to its authors, science keywords, datasets, category, sources, collections used, and science keyword dataset pairs.


science...
science...
dataset
year
dataset
dataset
category
collection
collection
publica...
collection
science...
author
source
science...
science...
author
author
science...

CONCLUSIONS & NEXT STEPS

- A significant advancement toward automating our dataset identification and bridging the gap between publications and datasets.
- Plan to continue expanding graph to include more publications
- Improve classification ability of our applications
  - Graph CNN
- Develop applications for use in automated classification and search that can be deployed to real settings

ABSTRACT

The NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) archives and distributes hundreds of Earth Science data collections to the public. These collections are used in research, resulting in the publication of thousands of scientific papers each year. As new users come to GES DISC for data, it is important for them to understand how prior research used the data. To help researchers, a knowledge graph (KG) was designed and implemented to connect publication citations with dataset metadata. The relationships created in the graph have the potential to allow the Web applications that utilize this information to directly connect the publication to the GES DISC datasets and services. These relationships are demonstrated using a web application prototype. In addition, the graph can also make connections between publications, datasets, and measurements based on the mentions of datasets and their attributes in the publications. To demonstrate this capability, a web application was created that takes the excerpt from the publication and returns a most likely dataset and measurement pairing, ranking the results based on how often these datasets and measurements were used in prior publications.