# Creating a knowledge graph to connect scientific publications and datasets for improving discovery of GES DISC's data and services.

ESIP Summer 2021
July 19-23, 2021
nathaniel.r.crosby@nasa.gov

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

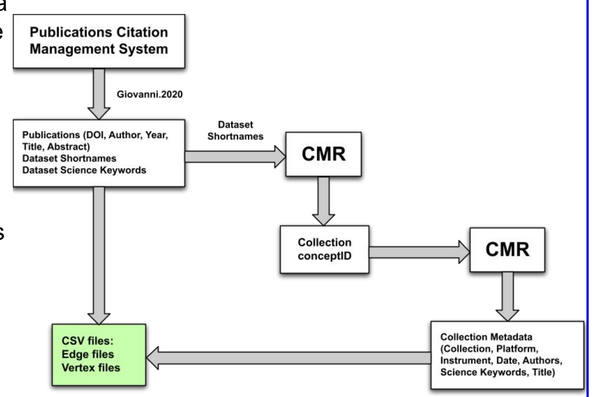
**Nathaniel Crosby[1], Rohan Dayal[1,2], Kristina Stoyanova[1,2], Irina Gerasimov[1,2], Armin Mehrabian[1,2], Jennifer Wei[1], Long Pham[1], and Mohammad Khayat[1,2]**
[1]Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA [2]ADNET Systems Inc., Lanham, MD, USA

## Improving Dataset Discovery

- Persistently absent linkage between scientific publications and related data due to the lack of dataset citations in those papers prevents data **findability** and **discovery**, as well as research **reproducibility** from accumulated knowledge.
- A solution to these issues is creating a knowledge graph database of publications, datasets, collections, instruments, platforms, authors, and science keywords.
- 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.

## KG Data Preparation Pipeline

- GES DISC citation management system contains publications citations that have been reviewed by the science experts to identify associated datasets and dataset measurements.
- Publications are labeled with dataset short names and Global Change Metadata Directory (GCMD) science keywords.
- To populate initial KG, ~250 publications from the year 2020 that used NASA Giovanni service were selected.
- NASA Common Metadata Repository (CMR) was queried by the dataset short names to gather information about collections currently available for the public search.
- Collected metadata is broken into edge and vertex files. The vertex files contain the publication, dataset, collection, etc. information while the edge files create a directed relationship between two vertices such as publication related to a dataset.

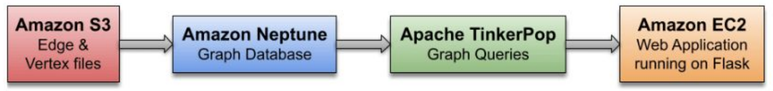



## Science Keywords

- 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
  - 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.
  - Use GCMD UUID as KG keyword id for compatibility.
- Mapping content of the paper to particular science keyword and dataset pairs is a tedious process.

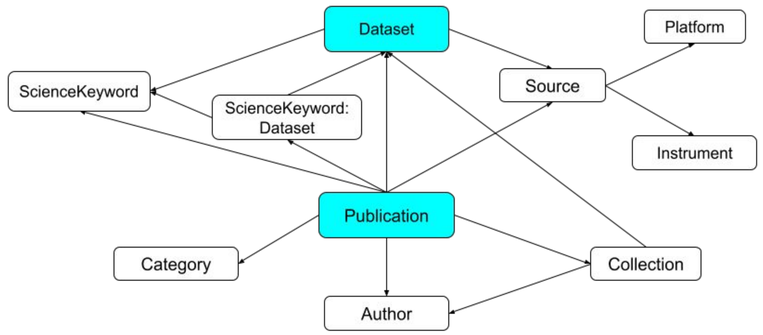
## AWS Framework

- AWS Neptune
  - AWS service that 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.
  - This includes our web application.




## Creating Knowledge Graph (KG)

- Upload Vertex and Edge CSV files into AWS S3 bucket
- Load CSV files from AWS S3 bucket into Neptune
- Use Gremlin to query Neptune database
- Graph vertex and edge types follow the diagram below:



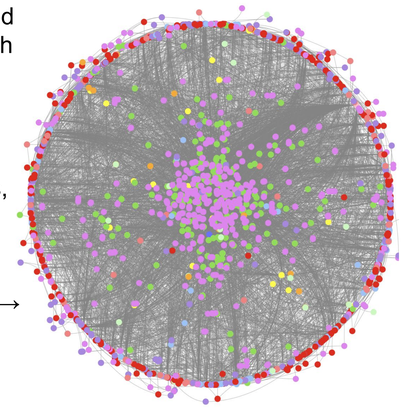

## Current and Future KG Work

- Expand the knowledge graph to include more publications and additional information about collections and papers.
- Utilize the content of the knowledge graph database, e.g. publications titles and abstracts to enhance the GES DISC dataset search:
  - Stoyanova, K. Gerasimov, I., Mehrabian, A., Wei, J, Khayat, M. (2021). *Improving Earth Science dataset search with publications content via Knowledge Graph linkage*, ESIP Summer meeting, 2021.
- Learn new dataset features from the paper texts and add to the graph database:
  - Dayal, R. Gerasimov, I., Mehrabian, A., Wei, J, Khayat, M. (2021). *Automated classification of scientific publications linked to GES DISC datasets*, ESIP Summer meeting, 2021.
- Use the KG to identify the datasets mentioned in the papers:
  - Traverse KG using features extracted from the papers.
  - Create a graph convolutional neural network (GCNN).

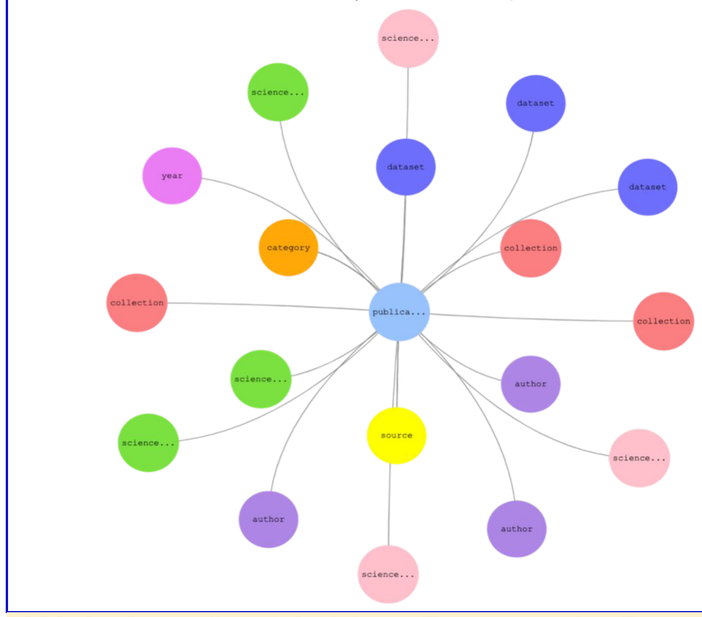
## Visualizing Graph

Entire Knowledge Graph:

- Each vertex category is color coded and represented as a dot., all graph edges are the gray lines.
- Space Complexity:
  - 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: (createdBy (Publications → Authors),
    - HasPlatform (Source → Platform)



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




## Web App for Linking Research, Data and Services

- The Web Application is built with Flask and runs on AWS.
  - A proof of concept for the improvements provided by the knowledge graph.
- The web app allows users to find a publication and immediately see related data collections, platforms, instruments, measurements, and data services.
- Provides links to allow user to directly access any of these items.
- This is an improvement over the current GES DISC services that do not show these connections.
- This is only a preliminary example of what the KG can accomplish.