Holistic Data Discovery: Navigating Human Health, Food, Environment, and Climate

27th Conference of Atmospheric Science Librarians International
AMS, 2024

Irina Gerasimov, Binita KC, Armin Mehrabian, Jerome Alfred, Andrey Savtchenko, James Acker, Mohammad Khayat, Jennifer Wei

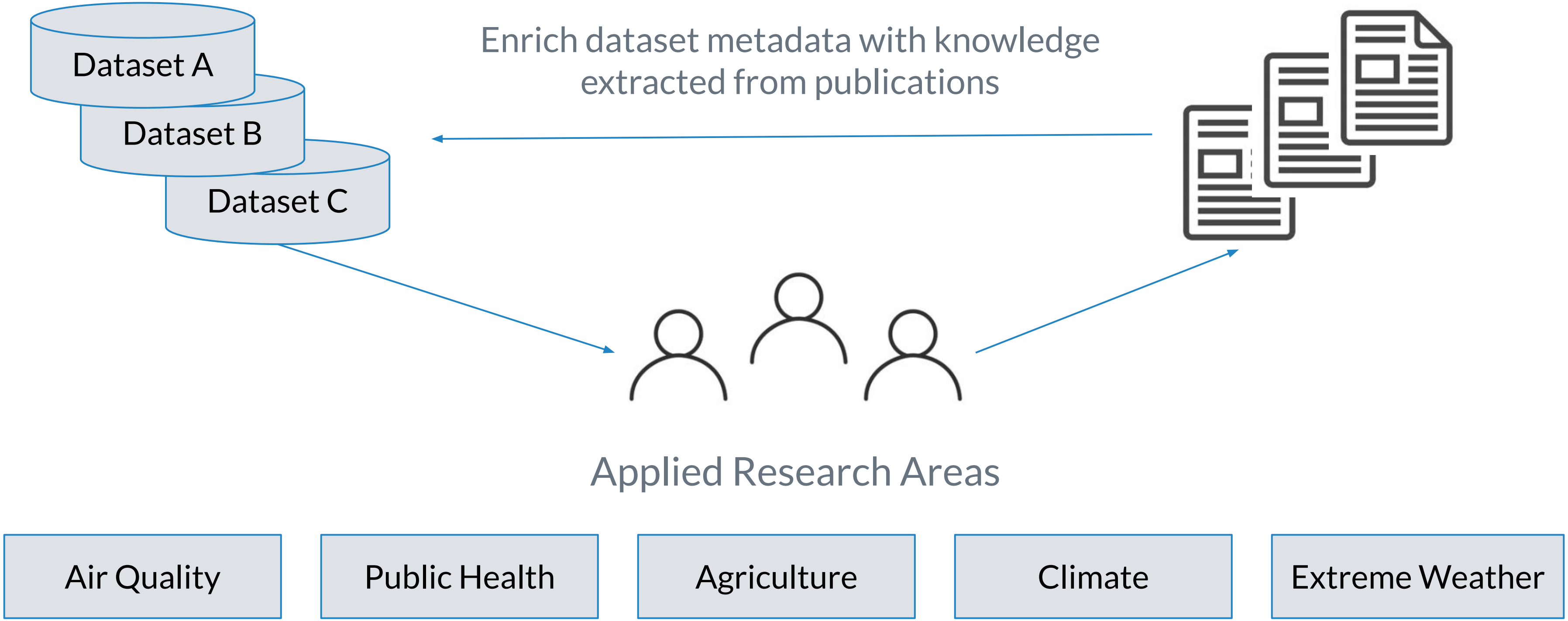
Code 619, NASA Goddard Space Flight Center, Greenbelt, MD, USA

ADNET Systems Inc., Lanham, MD, USA

Dataset Discovery Challenge

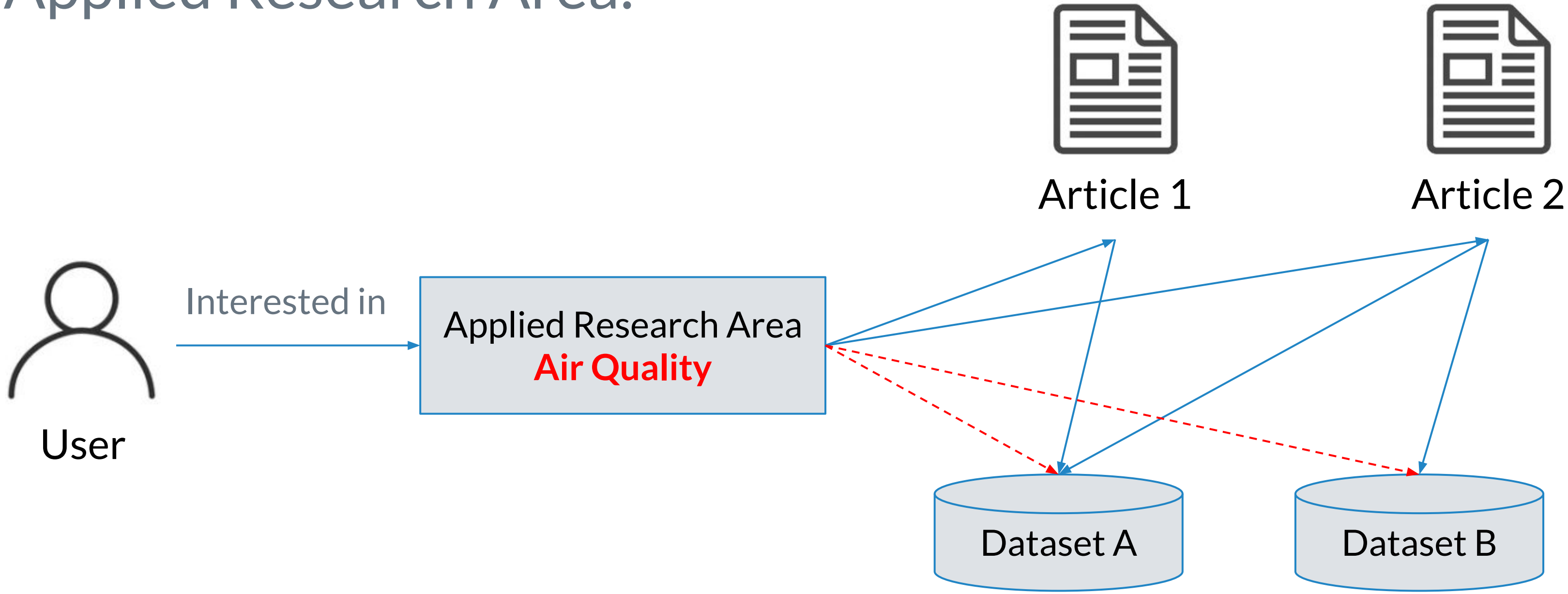
- The Dataset Discovery Challenge arises from the vast array of datasets provided by large data centers, such as NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC). These centers offer a multitude of datasets spanning diverse disciplines, including Atmospheric Composition, Water & Energy Cycles, and Climate Variability.
- For non-expert users, the process of navigating through these datasets is often hindered by the complexity of the dataset metadata. This complexity encompasses various aspects, such as the type of measurement, spatial and temporal resolutions, data formats, the time span covered by the data, and geographical coverage, among others.

Enable Data Discovery through Published Research



Approach

Allow dataset discovery based on user interest in specific Applied Research Area:



Benefits

When a user accesses the dataset accompanied by its publication, they gain several valuable insights, including:

- Demonstrative application of the dataset in a field of applied research relevant to the user's area of interest.
- A real-world instance showcasing the utilization of the dataset in a scholarly article.
- Detailed insights into the particular application of the dataset in applied research.
- A comprehensive guide outlining the usage of the dataset, including the selection of variables, preprocessing steps, and the time span covered.
- In-depth analysis and visual representations derived from the dataset data.

Example

Publication: Katpatal, Y.B., Patel, V.K. & Londhe, D.S. Impact of COVID-19 on spatio-temporal variation of aerosols and air pollutants concentration over India derived from MODIS, OMI and AIRS. *Spat. Inf. Res.* **31**, 637–651 (2023). DOI: 10.1007/s41324-023-00530-4

GES DISC datasets used in publication: OMTO3d_003, OMNO2d_003, OMDOAO3_003, OMSO2e_003, AIRS3STD_006.

Publication Applied Research Areas:

AIR QUALITY, PUBLIC HEALTH

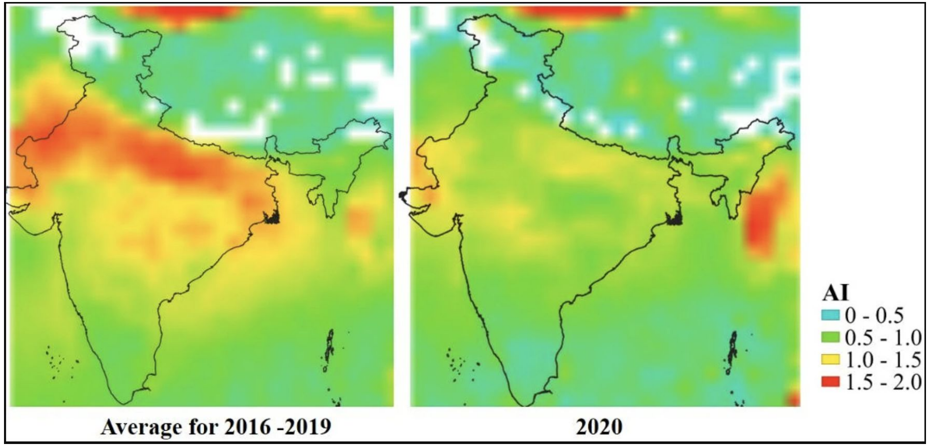


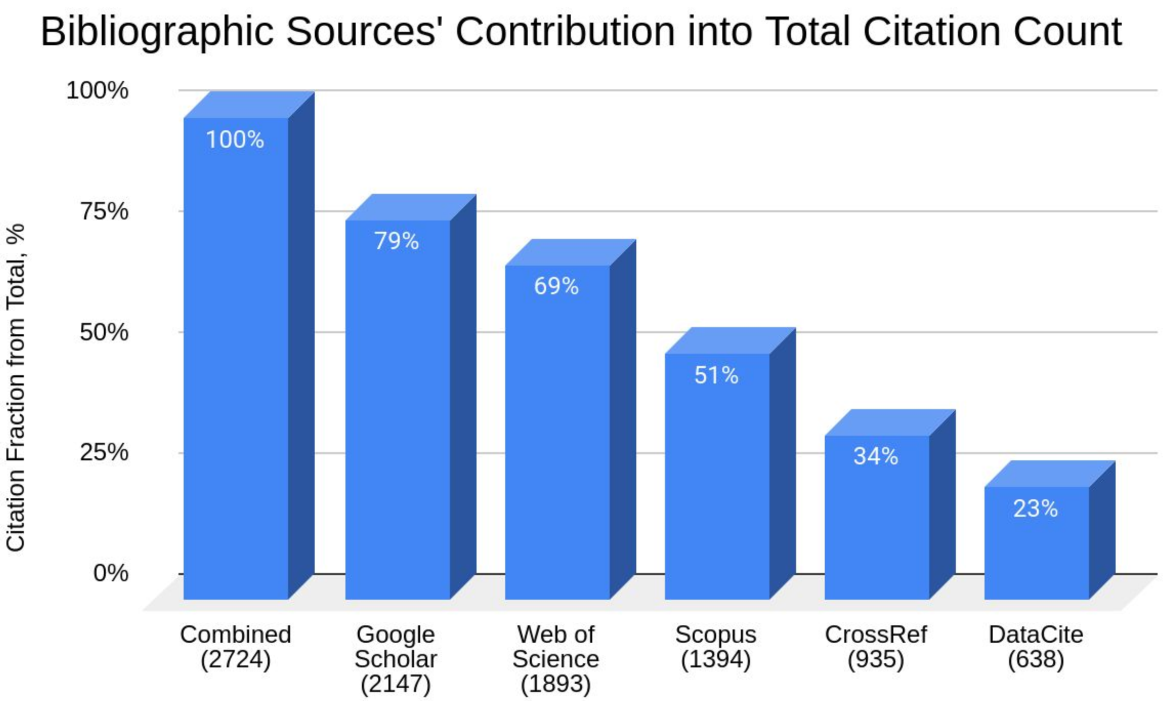
Figure from publication produced from the OMTO3d_003 data: Variation of OMI [Ozone Monitoring Instrument] Aura UV AI over India during 25 March to 3 May 2020 (lockdown period) and average of 2016–2019 for the same period.

Challenge - Poor Dataset Citation

Poor citing of datasets in articles makes finding articles challenging.

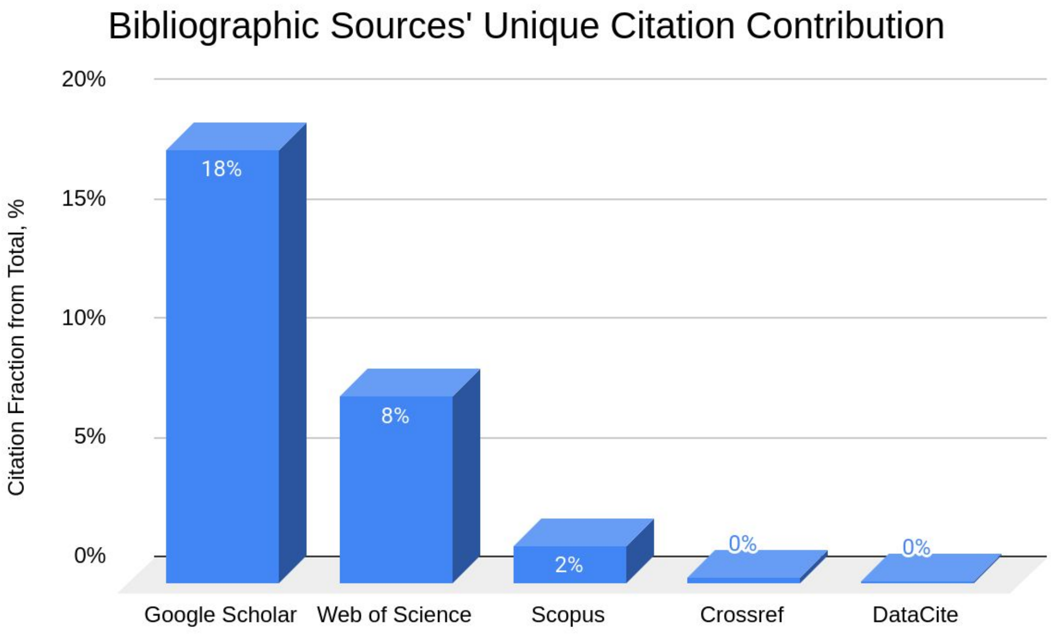
We have to search multiple bibliographic sources for the dataset DOIs.

The figure at right presents the search results on ~1,500 dataset DOIs for articles published in the last decade



Challenge - Poor Dataset Citation

- While the largest contributor of unique citations is Google Scholar, it lacks the citation metadata.
- While Crossref contains majority of publications DOIs - it has very limited data on dataset to publication linkage.



Challenge - No uniform metadata

As we retrieve metadata for publications, we get keywords when available.

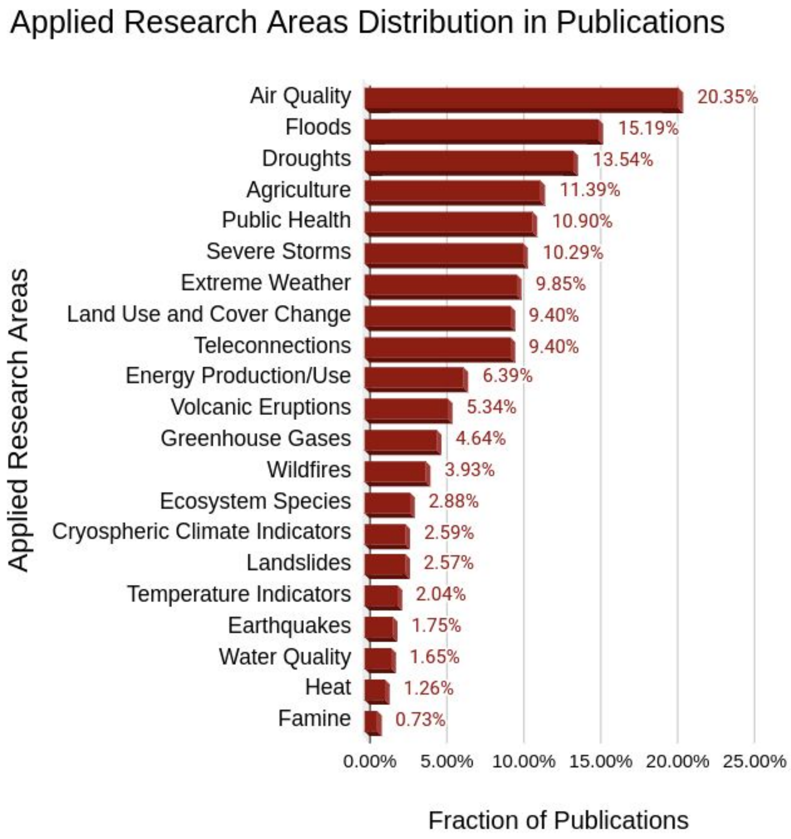
- The keywords are retrieved for 50% of publications.
- The retrieved keywords are not interchangeable between publications.
- The retrieved keywords do not necessarily describe applied research areas of publications.

A mechanism for the uniform assignment of applied research areas is needed.

Methodology

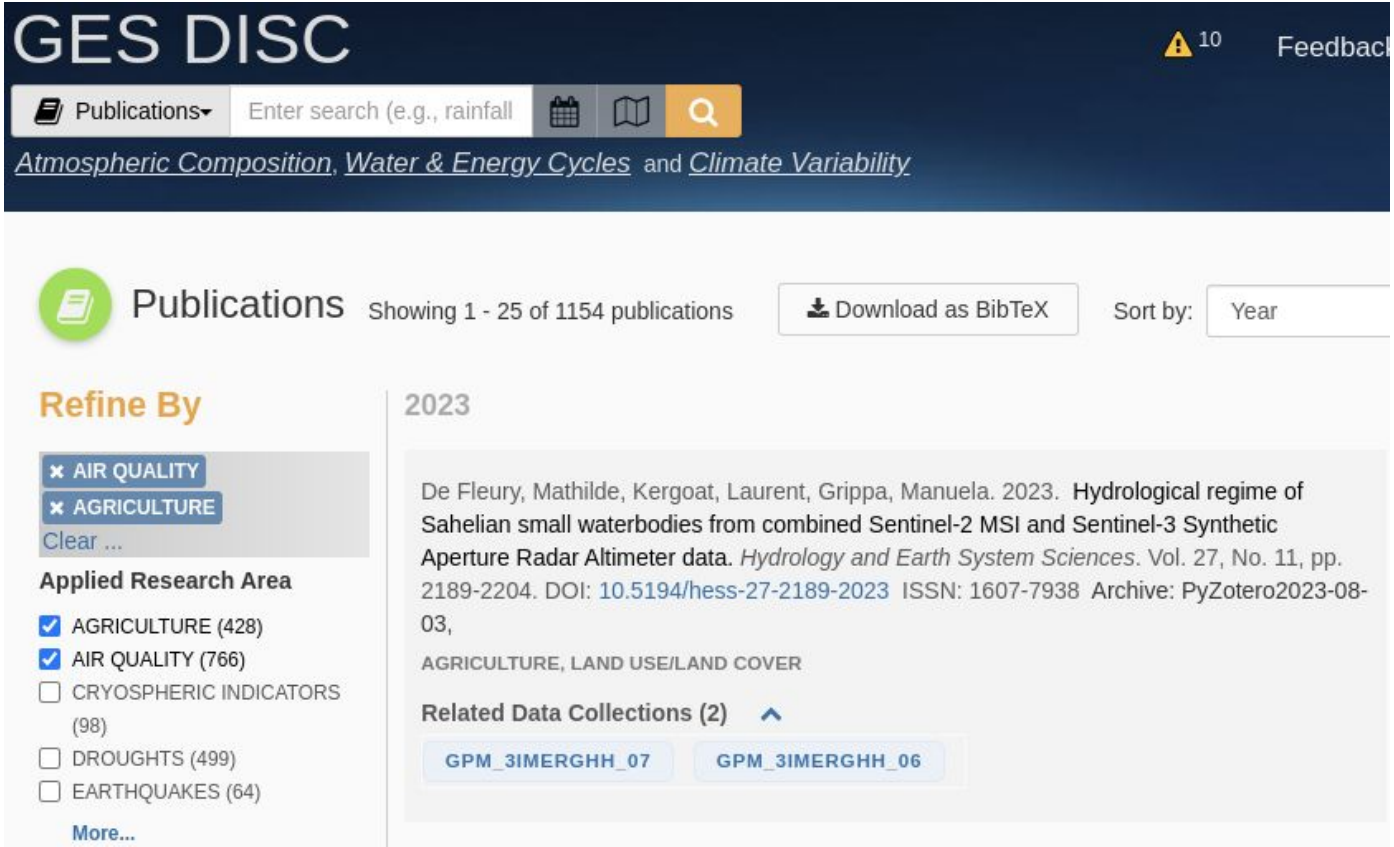
- Rely on content of the publication title and abstract to assign the publication to one of several applied research areas.
- Characterize each research area with the set of 10-20 terms.
- Preprocess publication abstracts using Natural Language techniques and extract terms.
- Based on the Term Frequency Inverse Document Frequency (TF-IDF) technique, TF-IDF vectors are generated for each publication, resulting in TF-IDF weights for each found search term.
- Each article is assigned to one or more Applied Research Areas based on the thresholds for the search terms' TF-IDF weight sums.

Results



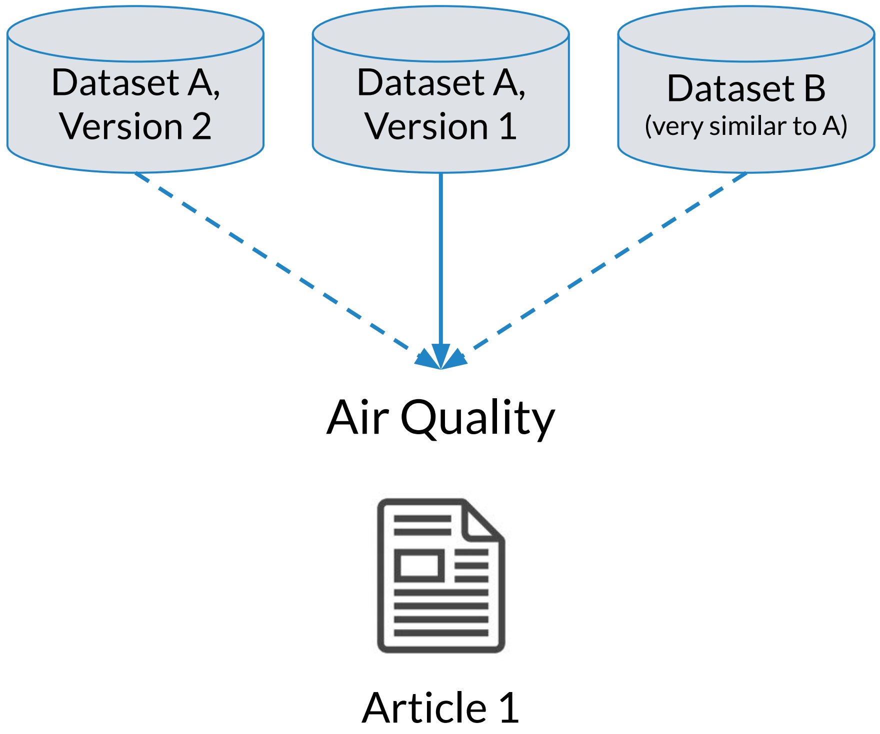
Approximately one-half of collected publications can be attributed to one or more Applied Research Areas

GES DISC Publications Search



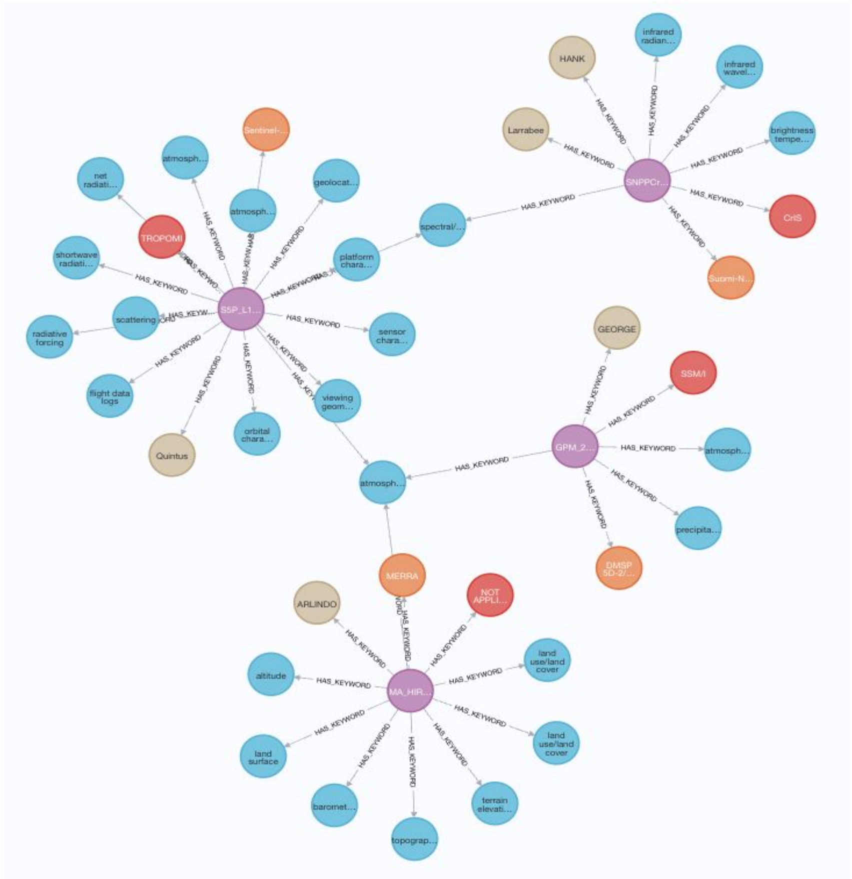
The GES DISC Publications Web Interface provides direct linkage between publications and datasets and allows Applied Research Area publications browsing.

Improving Dataset Discovery



- Assigning applied research areas to the datasets helps to discover the datasets, forgoing publications browsing.
- New versions of datasets can inherit the applied research areas of the older dataset versions, thus speeding up discovery of those datasets.
- The datasets that are similar to the ones that already have assigned applied research areas can be offered for usage as well.

Link Prediction



Link Prediction is a Machine Learning algorithm that can be applied to assign the applied research areas to the datasets that were not used in the publications, but which are very similar to the datasets that were used.

Please see a poster presented at AMS student conference: [Enhancing Dataset Discovery with Knowledge Graph Link Prediction Techniques].

Knowledge graph generated using GES DISC dataset metadata, with "Dataset" nodes in purple, "Keyword" nodes in blue, "Investigator" nodes in beige, "Platform" nodes in orange, and "Instrument" nodes in red.

Future Automated Information Extraction

GESDISC NASA

Large Language Models are powerful tools that allow information to be extracted from the publications. Using GPT-3.5 to extract information from abstracts of publications linked to GES DISC datasets yielded:

- Years of research coverage (70% of abstracts).
- Geographic regions (97% of abstracts).
- Studied variables (100% of abstracts).


```
"title": "Long-term spatio-temporal trends in atmospheric aerosols and trace gases over Pakistan using remote sensing",
"year": "2023",
"abstract": "One of today's most important environmental problems is air pollution augmentation. Air pollution is getting worse over time and hurts human health. For the current study, various polar orbiting satellites were utilized to collect data on PM2.5, SO2, AOD, CO, and ozone over Pakistan between January 2005 and December 2021. According to the spatial distribution results, these characteristics have high values throughout central Punjab, western Baluchistan, central Sindh, and Khyber Pakhtunkhwa. The seasonal variation in PM2.5, SO2, AOD, CO, and ozone was calculated using monthly data. The greatest value for PM2.5 is 8.7 108 kg/m3 during the monsoon season, while the highest value for SO2 is 1.4 105 kg/m2 during the winter. Over Punjab, Sindh, Baluchistan, KPK, and Gilgit, AOD was between 0.7 and > 1.0, CO was 127.2 ppb, and ozone was 330.7 DU. Furthermore, we create correlation maps of AOD, CO, SO2, PM2.5, and ozone and evaluate their relationship of high and low values across Pakistan. We looked into the 0.99 correlation between AOD and PM2.5, the strongest ever recorded. Despite this, we look at time series graphs to show the rising and falling pattern of these parameters from January 2005 to December 2021. We also used tables to determine the relative change in Multan, Lahore, Karachi, Peshawar, Quetta, Rawalpindi, Faisalabad, Hyderabad, Gujranwala, and Abbottabad in Pakistan from January 2005 to December 2021.",

"GPT vars": "Measurements found: PM2.5, SO2, AOD, CO, ozone",
"GPT dates": "Earliest year of research: 2005\nLatest year of research: 2021",
"GPT places": "Pakistan, central Punjab, western Baluchistan, central Sindh, Khyber Pakhtunkhwa, Punjab, Sindh, Baluchistan, KPK, Gilgit, Multan, Lahore, Karachi, Peshawar, Quetta, Rawalpindi, Faisalabad, Hyderabad, Gujranwala, Abbottabad."
```

Questions? Comments?

Contact Irina Gerasimov at Irina.Gerasimov@nasa.gov