Linking Science Data and Research to Enable Data and Knowledge Discovery

ESIP Winter 2021
January 26-29, 2021
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], 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

Creating publications-and-data linked library

- NASA Earth Science data archives contain thousands of publicly available data collections and thousands of research papers are being published yearly based on those data.
- These scientific papers carry valuable information about: studied phenomena, research methods applied to the datasets, groups of the datasets used together and particular dataset measurements. User exploring the data center datasets can use this information for better understanding which datasets are best fit for their applications. Having publications knowledge translated and stored in searchable format can facilitate usage-based dataset discovery and enhance **Findability, Accessibility, Interoperability and Reusability** (**FAIR**) principles of data management (Wilkinson, 2016).
- Goddard Earth Sciences Data and Information Services Center (GES DISC) approach to enhance usage-based dataset discovery consists of:
  - **Collecting publications citations that use GES DISC datasets**. This is an ongoing work that takes place at GES DISC. Citations are collected from the data science teams and online sources (Google Scholar, Scopus, Web of Science, DataCite) via data collection DOI and keywords search.
  - **Uniform term extraction** from publication title and abstracts using Semantic Web for Earth and Environment Technology (SWEET) and Global Change Master Directory (GCMD) ontologies.
  - **Indexing citations with:** GES DISC dataset and variable names, targeted locations, spatial and time ranges of the research, instruments and missions, science keywords, phenomena and properties.

Term extraction from text - Leveraging Earth Science ontologies

Research citations in GES DISC library are collected from hundreds of different publishing sources. Some of these citations have keywords and most do not. Keywords are the free text assigned by individuals. Indexing all citations by a single approach provides consistent set of keywords helping organize and search citations in uniform way: by measurement, phenomena, location, variables and datasets. Earth Science ontologies help extract the following terms:

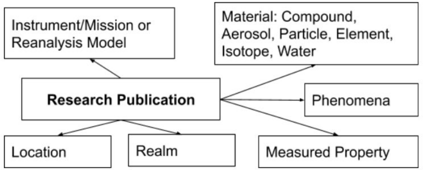
GCMD ontology - location, instrument, mission and science keywords:

- Covers majority of NASA Earth Science instruments and missions.
- Provides high level location coverage: countries and geographical features.
- Provides names for Earth Science dataset high level variables in a form of science keywords.

SWEET ontology - phenomena, property, process, material, realm, state, representation, human.

- Provides explicit coverage of Earth Science phenomena, properties, realms and materials.
- Large variety of terms allow for extracting many terms from the text. However many terms do not carry much content on their own. For example, "system", "variability", "event".

Based on these terms, the atmospheric science research paper content can be presented using the following publication knowledge concept map:



Instrument/Mission or Reanalysis Model

Material: Compound, Aerosol, Particle, Element, Isotope, Water

Research Publication

Phenomena

Location

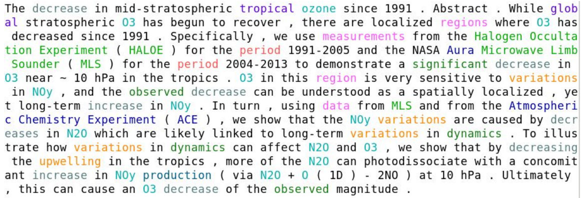
Realm

Measured Property

Example of term extraction from scientific publications

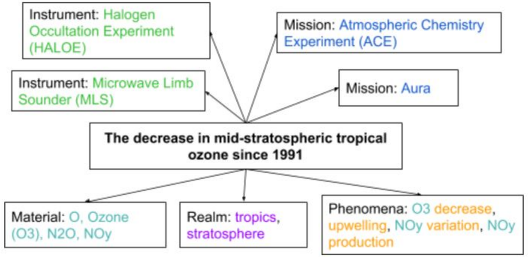
- **Case study**: Analyze publications collected by the MLS science team at JPL.
- **Challenges:**
  - The MLS science team has collected over a **1,000** research publications citations that utilize Aura/MLS data either in application research, re-analysis, or validating other datasets.
  - GES DISC archives **~50** MLS datasets and strives to guide users to the datasets most relevant to their research.
- **Goal:** Match users interests with published research
- **Approach:** Publication citations were processed to extract ontology-based terms containing mission and instrument names, atmospheric composition constituents, phenomena and realms.

MLS - Microwave Limb Sounder (Aura satellite, https://mls.jpl.nasa.gov/)
JPL - Jet Propulsion Lab



```
The decrease in mid-stratospheric tropical ozone since 1991 . Abstract . While glob
al stratospheric O3 has begun to recover , there are localized regions where O3 has
 decreased since 1991 . Specifically , we use measurements from the Halogen Occulta
tion Experiment ( HALOE ) for the period 1991-2005 and the NASA Aura Microwave Limb
 Sounder ( MLS ) for the period 2004-2013 to demonstrate a significant decrease in
O3 near ~ 10 hPa in the tropics . O3 in this region is very sensitive to variations
 in NOy , and the observed decrease can be understood as a spatially localized , ye
t long-term increase in NOy . In turn , using data from MLS and from the Atmospheri
c Chemistry Experiment ( ACE ) , we show that the NOy variations are caused by decr
eases in N2O which are likely linked to long-term variations in dynamics . To illus
trate how variations in dynamics can affect N2O and O3 , we show that by decreasing
 the upwelling in the tropics , more of the N2O can photodissociate with a concomit
ant increase in NOy production ( via N2O + O ( 1D ) - 2NO ) at 10 hPa . Ultimately
, this can cause an O3 decrease of the observed magnitude .
```

Extracted terms can be presented using the following concept map:



Instrument: Halogen Occultation Experiment (HALOE)

Mission: Atmospheric Chemistry Experiment (ACE)

Instrument: Microwave Limb Sounder (MLS)

Mission: Aura

The decrease in mid-stratospheric tropical ozone since 1991

Material: O, Ozone (O3), N2O, NOy

Realm: tropics, stratosphere

Phenomena: O3 decrease, upwelling, NOy variation, NOy production

As it can be seen in the map, the information about the instruments and missions is incomplete: there is an instrument, HALOE, without a mission and a mission, ACE without an instrument. This missing information can be augmented with information stored in GCMD dataset collection metadata that contain these instrument/mission pairs. The "phenomena" terms extracted from the text are combined with the "material" terms that appear together in the sentence because stand-alone phenomena terms such as "decrease", "variation" and "production" carry ambiguous meaning.

Data and knowledge discovery use case

- ~450 randomly selected articles collected by Aura/MLS science team were processed and ingested into a simple database.
- The database was queried for "Nitrogen Dioxide", or NO2, compound, one of the major air pollutants produced as a result of road traffic, other fossil fuel combustion processes and wildfires. Presence of NO2 in the air contributes to the formation and modification of other air pollutants, such as ozone and particulate matter.
- While Aura/MLS does not measure NO2, our query returned 32 articles because NO2 takes part in many atmospheric chemistry processes and it is studied along with many other species that are derived from Aura/MLS data.
- Generated statistics of terms appearing in articles that mention NO2:

Compounds studied along with NO2:

- 43% Ozone (O3)
- 40% Chlorine Nitrate (ClONO2)
- 40% Nitric acid (HNO3)
- 34% Hydrochloric acid (HCl)
- 28% Carbon Monoxide (CO)
- 25% Chlorine Monoxide (ClO)
- 22% Nitric Oxide (NO)
- 22% Aerosols

Instrument/Mission:

- 60% Microwave Limb Sounder/Aura
- 22% Ozone Monitoring Instrument/Aura
- 19% Michelson Interferometer for Passive Atmospheric Sounding/ENVISAT
- 16% Fourier Transform Spectrometer/Atmospheric Chemistry Experiment
- 13% Measurements of Pollution In The Troposphere/Terra
- 13% Scanning Imaging Absorption Spectrometer for Atmospheric Chartography/ENVISAT

Locations:

- 41% Polar
- 22% Arctic
- 13% Antarctic
- 9% China
- 3% each: South Korea, Bay of Bengal, India, Greenland, Japan, Russia

Phenomena:

- 19% Emission (ozone, fire, biomass burning, thermal)
- 19% Vortex (sunlight, chlorine, winter, spring, temperature)
- 16% Chlorine activation
- 15% Pollution(air, troposphere, urban, atmosphere, anthropogenic)
- 6% Ozone depletion

Lessons learned and future Work

- Variety of extracted terms allow search for citations by species, instruments/missions, realms, locations and phenomena. Quantity of extracted phenomena terms is high and needs further concept modeling work.
- Since average article uses datasets from more than one instrument/mission pair, extracted terms for instrument/mission and materials/measured properties are not sufficient to determine which measurements were taken from which instrument. As articles rarely provide citations of used datasets, additional article text processing is needed to determine the exact datasets used in the paper.
- The approach described in this work can be applied to term extraction of other articles in Earth Sciences domain. The GES DISC library currently contains over 5,000 citations referencing GES DISC datasets. We plan to extract terms from these articles to provide enhanced data discovery.