GES DISC use of dataset DOIs for Searching Research Citations

Irina Gerasimov, Jerome Alfred, Andrey Savtchenko, and Jennifer Wei

NASA Goddard Space Flight Center
Code 610.2

Dataset Science Impact

- Dataset science impact = number of journal or conference publications that **use** this dataset.
- Can only be automatically computed if the dataset reference contains the dataset DOI.
- **Fact**: < 1,000 papers harvested so far from online sources for the GES DISC's >1,500 public collections.
- Science teams collect papers related to their datasets using 20's century methods such as **keyword search**.
- Papers collected by science teams in most cases cannot be directly linked to individual datasets and some of the papers do not use those datasets but rather refer to related science.

Online Research Citations Sources - Scopus

Pros: API: https://dev.elsevier.com/.

- Python-based API wrappers
- Input: dataset DOI and publication date range

Latest DOI search retrieval for GES DISC collections returns only 630 citations for 2013-2021 (GES DISC started assigning DOIs to datasets in 2012)

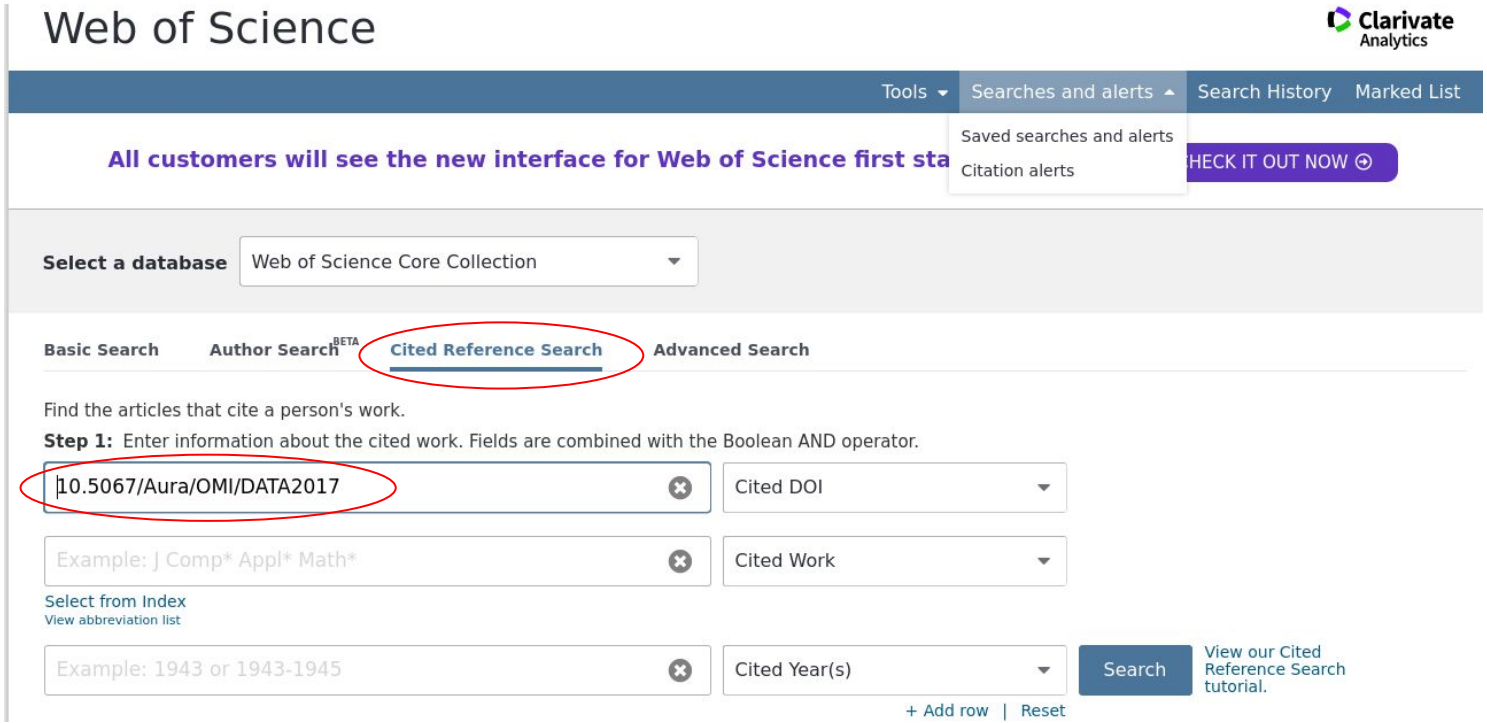
API usage example:

https://api.elsevier.com/content/search/scopus?query=ALL:'+doi+'&date=2013-2021&APIKey='+key

Cons: covers only Elsevier publications

Online Research Citations Sources - Web of Science

https://apps.webofknowledge.com/ **Cons**: No API



Online Research Citations Sources - Google Scholar

Google Scholar search produces the most number of citations.

Citations include peer-reviewed articles, conferences, books, thesis, articles from foreign publishing sources.

Cons: No API

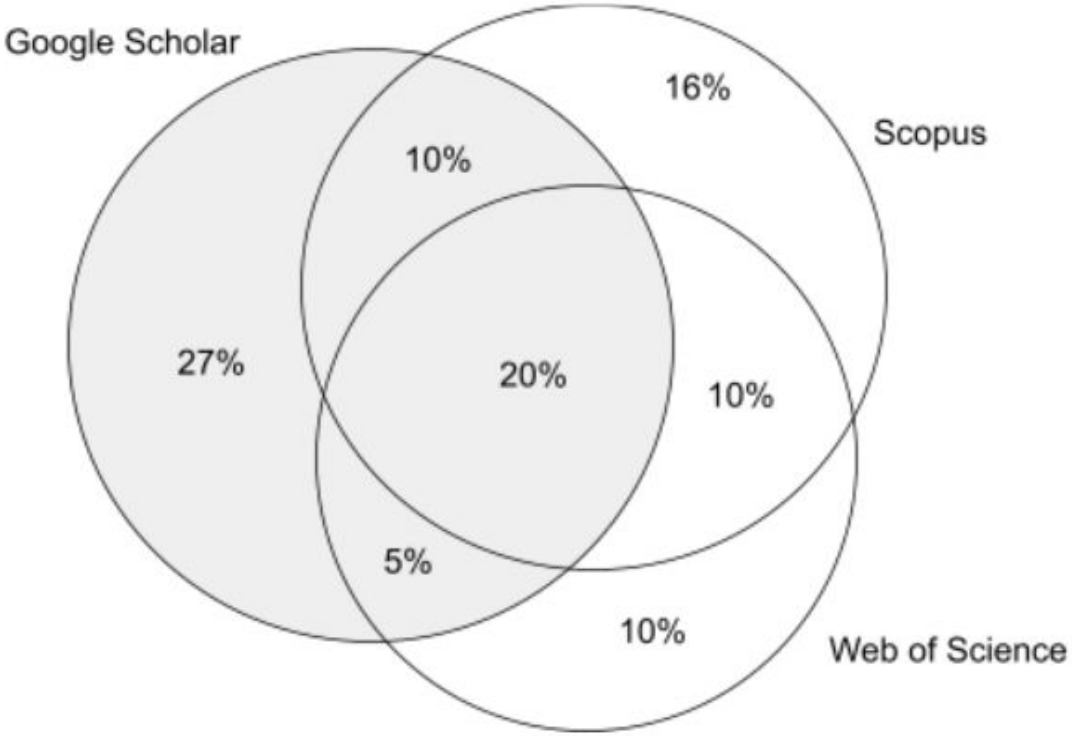
Online Research Citations Sources - Comparison

10.5067/Aura/OMI/DATA2017

- OMI/Aura Nitrogen Dioxide (NO2) Total and Tropospheric Column 1-orbit L2 Swath 13x24 km V003 (OMNO2)
- Temporal coverage since 2004-10-01

Source	Citation Count
Scopus	25
Web of Science	21
Google Scholar	44

Online Research Citations Sources - Comparison



Percentages of ***overlapping*** and ***unique*** citations retrieved by dataset DOIs from Google Scholar, Web of Science, and Scopus for publications dated **2019** that cite the GES DISC data collections.

Online Research Citations Sources - DataCite

Pros: API https://api.datacite.org/dois/10.5067/Aura/OMI/DATA2017

Cons: returns meager number of citations, above URL returns 0 citations!

for 2019 DataCite API returned only 3% of all citations for the combined GES DISC dataset DOI search in DataCite, Scopus, Web of Science, and Google Scholar

Relies on Crossref Event data search which did not work properly for long time and now become to look promising.

Crossref Event Data API Example:

https://api.eventdata.crossref.org/v1/events?mailto=irina.gerasimov@nasa.gov&obj-id=10.5067/Aura/OMI/DATA2017 returns 11 results

Online Research Citations Sources - Data Providers

Data providers perform various searches of online resources to find the publications that use their data, e.g.:

- OCO (https://ocov2.jpl.nasa.gov/publications/)
- Aura MLS (https://mls.jpl.nasa.gov/publications.php)

For the year **2019** we searched Google Scholar, Web of Science and Scopus for OCO-2 and Aura MLS datasets DOIs.

The fraction of papers that were returned by DOI search vs papers collected by science teams is:

- **8%** for OCO-2 datasets
- **17%** for Aura MLS datasets

Dataset identification in research publications

Citations collected by data providers generally are not directly linked to specific datasets and some of the publications may not use the datasets at all.

Tools are needed to link these publications and datasets.

At GES DISC Natural Language Processing (NLP) pipeline that processes publications texts was created to extract:

- Missions/instruments
- Citations
- Dataset short names
- Sentences with dataset terms

These metadata facilitate ***semi-automated*** dataset identification.

Thank you!

Please provide your comments and feedback to

Irina.Gerasimov@nasa.gov