



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

The screenshot shows the Web of Science interface. At the top right is the Clarivate Analytics logo. Below it is a navigation bar with 'Tools', 'Searches and alerts', 'Search History', and 'Marked List'. A dropdown menu for 'Searches and alerts' is open, showing 'Saved searches and alerts' and 'Citation alerts'. A purple button says 'CHECK IT OUT NOW'. Below the navigation bar is a message: 'All customers will see the new interface for Web of Science first starting on 11/15/2017'. A 'Select a database' dropdown is set to 'Web of Science Core Collection'. The search options are 'Basic Search', 'Author Search^{BETA}', 'Cited Reference Search' (circled in red), and 'Advanced Search'. Below this is the instruction: 'Find the articles that cite a person's work. Step 1: Enter information about the cited work. Fields are combined with the Boolean AND operator.' There are three search rows. The first row has '10.5067/Aura/OMI/DATA2017' (circled in red) in the 'Cited DOI' field. The second row has 'Example: J Comp* Appl* Math*' in the 'Cited Work' field. The third row has 'Example: 1943 or 1943-1945' in the 'Cited Year(s)' field. A 'Search' button is at the bottom right of the search area. Below the search area are links for '+ Add row' and 'Reset'. A link for 'View our Cited Reference Search tutorial.' is also present.



Online Research Citations Sources - Google Scholar

The screenshot shows a Google Scholar search interface. At the top left is the Google Scholar logo. The search bar contains the query "10.5067/AURA/OMI/DATA2017" and a search button. Below the search bar, it indicates "Page 4 of about 43 results (0.02 sec)". On the left, there are filters for "Articles" and a time range selector with options: "Any time", "Since 2021", "Since 2020", "Since 2017", and "Custom range...". The main search result is for the article "Impact of COVID-19 on the Air Quality over China and India Using Long-term (2009-2020) Multi-Satellite Data" by M Soni, S Verma, H Jethava, and S Payra, published in Aerosol and Air Quality Research in 2021. The article is available as a PDF from aaqr.org. The citation count is 5, and it is listed in the Web of Science database.

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

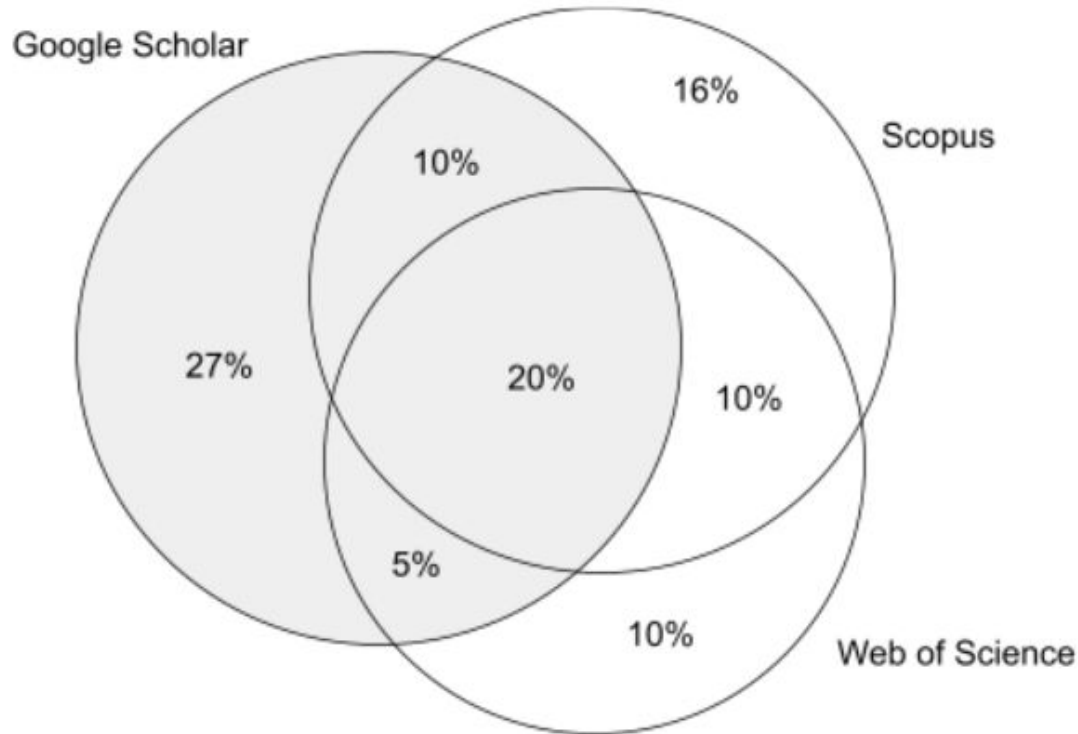
10.5067/Aura/OMI/DATA2017

- 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/doi/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