TRUSTWORTHY REPLICA DATA REPOSITORIES

CHALLENGES AND SOLUTIONS

AUTHORS

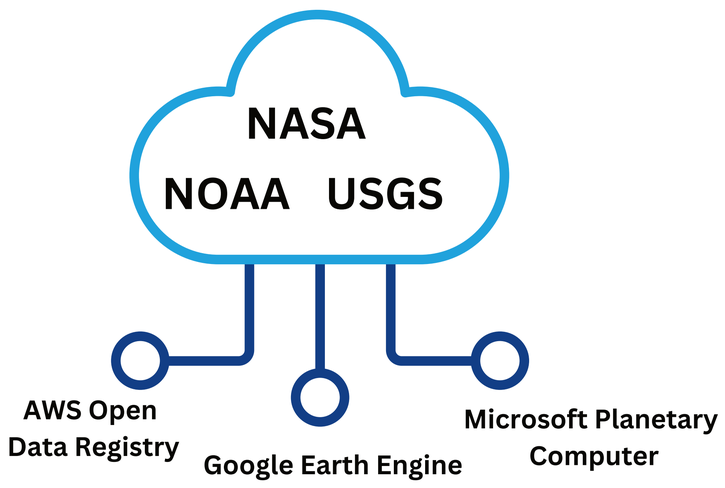
Mahabal Hegde[1], Simon Ilyushchenko[2], and Brianna Rita Pagán[1,3]

AFFILIATIONS

1 - NASA Goddard Spaceflight Center
2 - Google
3 - ADNET Systems Inc.

BACKGROUND

Earth system science is transitioning to the cloud. There are a **growing number of cloud providers redistributing** data collections from a **trusted data source.**



CHALLENGES

Many times the data is transformed and provided in a '**fit for a purpose**' format.

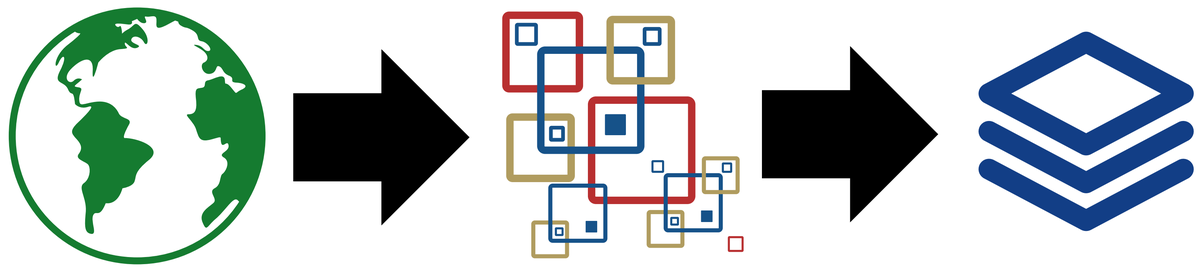


Figure 1: A representation of processing analysis ready data (*Landsat Analysis Ready Data Coming Soon*)

A **lack of open data governance** for **mirror data stores** can result in distributed data that potentially is:

- INCOMPLETE
- OUTDATED
- INACCURATE

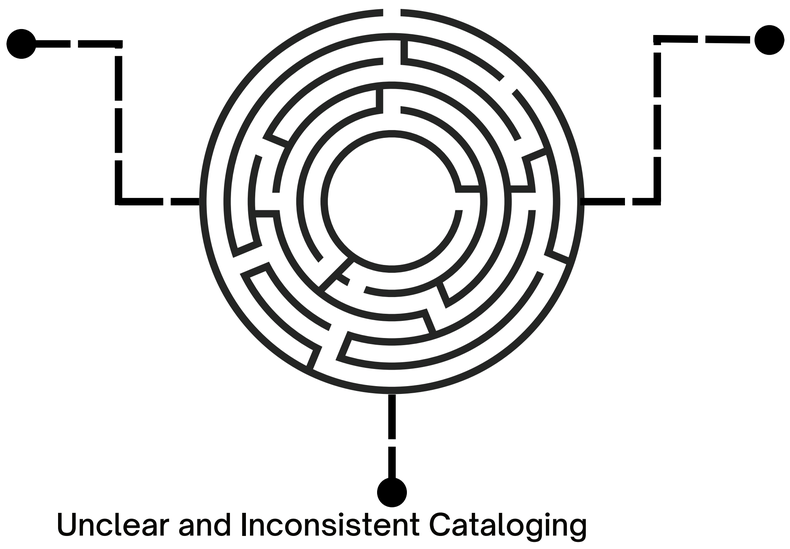
CASE STUDIES

Multiple Trusted Sources

As there can be many scientific uses for satellite data, it's common to have joint missions which allow **multiple agencies to leverage the same data** for different use cases. One example is the **NASA-NOAA Joint Polar Satellite Systems (JPSS) mission** which is responsible for the Visibility Infrared Imaging Radiometer Suite (VIIRS) sensor on two satellites.

VIIRS data is distributed by both NOAA and NASA. However, **which product is the 'correct' one depends on your use case**. Data for VIIRS is available in multiple resolutions. Other datasets might be offered in different cadences or processing levels.

To further complicate provenance, **numerous different data providers at NASA distribute VIIRS data**. Upon further investigating, we **uncovered additional apparent distribution endpoints which are not documented in NASA's Common Metadata Repository (CMR).** While advanced catalogs do exist like CMR or the new Copernicus dataspaces, both are easily overloaded with inefficient query calls.



Unclear and Inconsistent Cataloging

While trusted data providers might provide official catalogues to help simplify data acquisition, there are numerous cases where **outdated or inaccurate endpoints are accidentally exposed or uncovered by the public, and more alarmingly by data redistributors**. We identified numerous examples where initiatives like Google Earth Engine or pangeo-forge **unintentionally used the wrong endpoints** to copy data to their systems. These can be hard to detect as many have very official URLs for example:

- https://gimms.gsfc.nasa.gov/VIIRS/
- https://acd-ext.gsfc.nasa.gov/anonftp/toms/omi/data/
- https://www.ncei.noaa.gov/data/global-precipitation-climatology-project-gpcp-daily/

Unofficial Redistributions

As visualized above, **there is a need in the scientific community for datasets to be transformed for specific purposes or harmonized** between multiple datasets. When trusted data sources do not provide such products, data redistributors will.

One example is when the Pangeo-Forge project moved the wrong version of Global Precipitation Climatology Project (GPCP) data to a cloud-native format. An error was detected, ultimately leading back to an unofficial data endpoint that was used. Luckily, Pangeo-Forge recipes are **open source which enables such errors to be reported and resolved**.

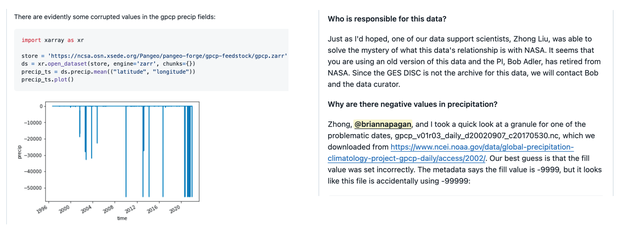


Figure 2: Online correspondence showing dataset issues.

Google Earth Engine provides merged products of data, which can be more susceptible to error and without proper data provenance. An example is a merged ozone product (TOMS/OMI) which l**acks an asset property that allows users to know which products are used**.

SOLUTIONS

IMPROVE CATALOGING

Trusted data sources need to set more unified cataloging requirements. Better diligence is needed to prevent unofficial endpoints to be exposed to the public. **Data providers should also enable more straightforward dataset validity indicators** to boost user confidence when searching for data.

THIRD-PARTY VERIFICATION

Third-party entities should exist to **audit open data governance policies as well as monitor dataset health.** Such guidelines are increasingly needed not just for datasets but also for **computations completed across different platforms**. Data redistributers should also work more closely with the trusted data sources to ensure data accuracy.

INCENTIVIZE ACCURACY

Bug bounty programs are offered by organizations or software developers where **individuals can receive compensation for reporting bug**s. Google, Microsoft and even the United States. Department of Defense have started using such programs. Similar efforts could be **supported for the geospatial community** as a whole.

REFERENCES

- *Landsat Analysis Ready Data Coming Soon* | Landsat Science. 25 Sept. 2017, https://landsat.gsfc.nasa.gov/article/landsat-analysis-ready-data-coming-soon/. Accessed 4 Dec. 2023.

CONTACT

mahabaleshwa.s.hegde@nasa.gov
simonf@google.com
brianna.r.pagan@nasa.gov