

Operating Mirror Data Stores

Mahabal Hegde

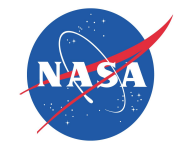
NASA Goddard Earth Science Data and Information Services Center
Greenbelt, MD



Mirror Data Stores

A data store that is derived from the authoritative source which might

1. be stored in a format different than the original
2. be hosted using a different resource (e.g., AWS S3, Azure Blob)
3. have been transformed



Mirror Data Stores: Business Case

Data access in a format and location friendly for consumption

Advantages:

- Better read performance = faster analysis + new possibilities
- Better user experience = reduced frustration
- **Reduced cost in developing services; datacenter-hosted services benefit too!**

Reality:

- No one size fits all solution :(



Mirror Data Stores: Challenges

Common data consumer concerns:

- Is this data copy from a trusted source?
- What is the provenance?
- Is the data verified + verifiable?
- Is the data complete?
- Were the data values modified?
- How to cite the data? Licenses etc.,

If a data store can answer these, how to provide evidence?



Mirror Data Stores: Classic Case

A person or an organization

1. Acquires data from its official source
2. Reformats or transforms it for their use
3. Publishes data to open catalogs
4. Moves on to other interesting things



Mirror Data Stores: Proliferation



* [Can icon](#) created by [Freepik - Flaticon](#)



Mirror Data Stores: They do it; why can't we?



They provide unbroken chain of comparison linking to the highest standard!

Image Credits: NIST, Ruler icons created by [Freepik - Flaticon](#)



Mirror Data Stores: Trusted Source

Key Characteristics:

- Expert guidance on data
- Continuity of access
- Data integrity and authenticity
- Long term preservation

Standards and Certifications:

- [CoreTrustSeal](#) Trustworthy Data Repositories (by DSA + WDS)

A Best Practice:

- Maintainers of data stores display the the accreditation certificate/seal and link to evidence of compliance.



Mirror Data Stores: Provenance

Minimal Information:

1. Process Identifier
2. Input
3. Time when the input was sourced

Standards to consider:

1. Processing Graphs (<https://api.openeo.org/v/0.3.0/processgraphs/>)
2. PROV (<https://www.w3.org/TR/prov-overview/>)

A Best Practice:

- Consider provenance to be part of the data store and embed with the data.

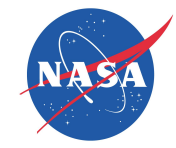


Mirror Data Stores: Verification

The question to answer: does the data store contain correct numbers?

Candidate Best Practices:

1. Provide checkers to verify data for select cases similar to unit tests for code
2. Have frequent verification of data stores and publish the last verification time



Mirror Data Stores: Completeness

The question to answer: does the data store contain all the data?

Candidate Best Practices:

1. Provide metadata about freshness (e.g., “Last updated at”)
2. Provide easy to verify completeness checker.

From the provenance, trace the origin and compare using:

- Data update/insert time
- Hash of input catalog
- Root hash of input data (hash tree)

Refer to Data store's provenance.



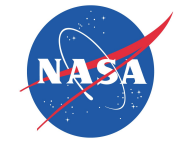
Mirror Data Stores: Licensing and Citation

1. Use metadata in data registries (e.g., AWS Open Data registry)
2. Evangelize use of DOIs



Mirror Data Stores: What is next?

Any interest in ESIP Cloud Computing Cluster to help standardize or come up with guidelines/best practices?



Questions?



EO Process Graphs

```
{
  "process_id": "append_to_store",
  "data": [
    {
      "process_id": "get_data_from_opendap",
      "url": "https://gpm1/3B-HHR.MS.MRG.3IMERG.20210929-S000000-E002959.0000.V06B.HDF5.nc4?precipitationCal[0:1:0][0:1:3599][0:1:1799]",
      "sourced_at": "2022-01-01T00:00:00Z"
    },
    {
      "process_id": "get_data_from_opendap",
      "url": "https://gpm1/3B-HHR.MS.MRG.3IMERG.20210929-S003000-E005959.0030.V06B.HDF5.nc4?precipitationCal[0:1:0][0:1:3599][0:1:1799]",
      "sourced_at": "2022-01-01T00:00:00Z"
    }
  ]
}
```

