

## PRACTICE PAPER

# Recommendations to Improve Downloads of Large Earth Observation Data

Rahul Ramachandran<sup>1</sup>, Christopher Lynnes<sup>2</sup>, Kathleen Baynes<sup>2</sup>, Kevin Murphy<sup>3</sup>, Jamie Baker<sup>4</sup>, Jamie Kinney<sup>4</sup>, Ariel Gold<sup>4</sup>, Jed Sundwall<sup>4</sup>, Mark Korver<sup>4</sup>, Allison Lieber<sup>5</sup>, William Vambenepe<sup>5</sup>, Matthew Hancher<sup>5</sup>, Rebecca Moore<sup>5</sup>, Tyler Erickson<sup>5</sup>, Josh Henretig<sup>6</sup>, Brant Zwiefel<sup>6</sup>, Heather Patrick-Ahlstrom<sup>6</sup> and Matthew J. Smith<sup>6</sup>

<sup>1</sup> NASA Marshall Space Flight Center, US

<sup>2</sup> NASA Goddard Space Flight Center, US

<sup>3</sup> NASA Headquarters, US

<sup>4</sup> Amazon Web Services, US

<sup>5</sup> Google, US

<sup>6</sup> Microsoft, US

Corresponding author: Rahul Ramachandran ([rahul.ramachandran@nasa.gov](mailto:rahul.ramachandran@nasa.gov))

With the volume of Earth observation data expanding rapidly, cloud computing is quickly changing the way these data are processed, analyzed, and visualized. Collocating freely available Earth observation data on a cloud computing infrastructure may create opportunities unforeseen by the original data provider for innovation and value-added data re-use, but existing systems at data centers are not designed for supporting requests for large data transfers. A lack of common methodology necessitates that each data center handle such requests from different cloud vendors differently. Guidelines are needed to support enabling all cloud vendors to utilize a common methodology for bulk-downloading data from data centers, thus preventing the providers from building custom capabilities to meet the needs of individual vendors.

This paper presents recommendations distilled from use cases provided by three cloud vendors (Amazon, Google, and Microsoft) and are based on the vendors' interactions with data systems at different Federal agencies and organizations. These specific recommendations range from obvious steps for improving data usability (such as ensuring the use of standard data formats and commonly supported projections) to non-obvious undertakings important for enabling bulk data downloads at scale. These recommendations can be used to evaluate and improve existing data systems for high-volume data transfers, and their adoption can lead to cloud vendors utilizing a common methodology.

**Keywords:** Earth Observation Data; Large Data Transfers; Cloud; Best Practices

## 1. Introduction

### 1.1 Purpose

With the volume of Earth observation data expanding rapidly, cloud computing is quickly changing the way these data are processed, analyzed, and visualized. The cloud infrastructure provides the flexibility to scale up (for both storage and computation) to large volumes of data and to efficiently process high velocity data streams. Collocating freely available Earth observation data on a cloud infrastructure may create opportunities for innovation and value-added data re-use that were unforeseen by the original data provider (or data center). These innovations could spur new industries and applications and spawn scientific pathways previously undiscoverable due to data volume and computational infrastructure limitations. Various cloud vendors are now

regularly accessing data from multiple data centers, and they can be viewed as a new class of data consumer, with distinct search, access, and use patterns which differentiate them from researchers (the typical data users).

NASA, in collaboration with Amazon, Google, and Microsoft, propose a set of recommendations to enable efficient transfer of Earth observation data from existing data systems to a cloud computing infrastructure. The purpose of these recommendations is to guide all data providers in evaluating existing data systems and improving any issues uncovered to enable efficient search, access, and use of large volumes of data. Employing these recommendations as an integrated set of guidelines can allow cloud vendors to utilize a common methodology for bulk-downloading data from data providers, thus preventing the providers from building custom capabilities to meet the needs of individual vendors. The adoption of these recommendations as a set of guidelines will benefit a new class of data consumer interested in moving large volumes of data from a data center to any other location. In the remainder of the paper, the term “user” refers to this new class of data consumer, namely cloud vendors who want to provide data as a resource to their customers but also applies to researchers needing to leverage cloud computing or any distributed high-performance computing environment.

## 1.2 Scope

Barriers to data acquisition as well as technical challenges prevent data from being utilized to their full potential (Overpeck, et al., 2011). In order to reduce these challenges, the U.S. Group on Earth Observations (USGEO) drafted the *Common Framework for Earth-Observation Data*, which provides Federal agencies with recommended standards on various facets of data management (National Science and Technology Council Committee on Environment, Natural Resources, and Sustainability, U.S. Group on Earth Observations Subcommittee Data Management Working, 2016). These data management facets include data search and discovery services, data access services, data documentation, and compatible formats and vocabularies. The common framework addresses issues associated with these facets by endorsing specific standards and protocols and listing recommended methods and practices. The ultimate intent of these endorsements is to aid in the creation of value-added products such as data portals, visualizations, and other tools. The recommendations presented in this paper are limited to Earth observation data and are made within the context of the *Common Framework for Earth-Observation Data*. The common framework adopts the definition of Earth observation data used in the 2013 *National Strategy for Civil Earth Observations* (Executive Office of the President, 2013):

[We] use the terms “data” and “Earth observations” interchangeably to mean geo – referenced digital information about Earth, including the observations, metadata, imagery, derived products, data-processing algorithms (including computer source code and its documentation), and forecasts and analyses produced by computer models. Non-digital data, published papers, preserved geological or biological samples, or other media that have not been digitized are not included in this definition. . .’

This paper is a result of the recommendations distilled from use cases provided by three cloud vendors (Amazon, Google, and Microsoft) and used to assess NASA’s existing data system. The provided use cases described how Earth observation data would be used by cloud vendors and also documented the vendors’ experiences when interacting with data systems at different Federal agencies and organizations. Because the lessons learned from these use cases may be applicable to data management at other agencies and organizations, the resulting recommendations have been synthesized in this document. This paper thus intends to provide an integrated set of guidelines for evaluating and improving existing data systems in the context of high-volume data transfers. This paper further recommends that others follow the same guidelines to facilitate efficient access and improve data usability.

## 2. Data Search and Discovery Services

These services enable search and discovery of relevant datasets from different distributed data archives. The recommendations listed below focus on automating this process to enable bulk downloads and to preserve synchronization of data holdings.

## 2.1 Search and Discovery

### Recommendations:

- Provide a single, authoritative catalog for federated data systems that contains all data holdings.
- Ensure the catalog contains a complete and accurate representation of the data holdings at any time and is consistently updated.
- Provide a complete dataset metadata listing, or catalog file, for download in a simple machine-readable format such as JSON (2017). The catalog file should:
  - List all available data products.
  - List all data granules for each data product with data access URLs.
  - Utilize an efficient hierarchy where needed to minimize paging.

### Rationale:

In order for users to efficiently search for and discover data, the authoritative metadata catalog must be accurate and up-to-date. Data that cannot be discovered through search via the catalog do not exist. In addition, cloud vendors utilize large automated systems to query catalogs to obtain all the metadata information needed to bulk download data, and they require a scalable method to fetch the entire/complete metadata catalog that is also easy to traverse. This could be the API so long as it can handle the large traffic loads. A machine-parsable catalog file that lists all dataset metadata is a simple, scalable, and robust mechanism that can address this need. This catalog file should include high-level information for each data product such as the product name, dataset description, reference documentation links, and any other relevant information users need when selecting a product to use. A complete listing of the data granules for each data product with access URLs is required in order to download the product. Finally, utilizing a hierarchy ensures that the catalog file does not grow unwieldy and makes parsing efficient.

## 2.2 Data Notification Services

### Recommendations:

- Develop a lightweight push notification feed to notify users of new datasets and granules and version updates. The notification system should:
  - Be based on an HTTP post request to a URL that is registered to a user.
  - Distribute a notification message that is machine parsable.
- Allow users to subscribe to a notification feed for individual datasets.

### Rationale:

A push mechanism places control of the curated archives directly in the hands of the data center and eliminates polling latency (i.e., the latency introduced by the rate at which users are able to monitor catalog files to detect new or updated data). A push mechanism also has the significant advantage of low latency for near-real-time data. An effective notification message provides enough detail on a dataset update to allow the user to immediately download and ingest the new or updated data if desired.

## 3. Data Access Services

While data access services provide users with various methods of retrieving data with a range of functionality such as subsetting, these specific recommendations focus on access from different types of data servers to support bulk downloads.

### Recommendations:

- Provide data access using HTTP/HTTPS or FTP file downloads without restrictions or with basic authentication headers. Avoid stateful authentication mechanisms that require the user to enter through a login form each time.
- Allow parallel downloads. If there must be a limit on parallel downloads, ensure that the limit is high enough to completely download large archives in a reasonable amount of time. Avoid requirements that all connections in a session come from the same IP address. Designate a technical point of contact to address questions concerning download limits (see Section 5.1).

**Rationale:**

A system acquiring data may not be a simple computer or virtual machine (VM) running conventional scripts. For example, cloud vendors have a whole distributed system for managing the retrieval of data from partners at scale. This system does not interact very well with stateful sessions, which generally operate under the assumption that there is a single machine at the other end of the line.

## 4. Data Documentation Services

Data documentation services are needed to ensure that the data are used correctly. As such, these recommendations include ensuring that data transferred remain in sync, there is no data corruption during the download process, and the data are read correctly by the analysis/visualization software.

### 4.1 Metadata

#### 4.1.1 Data Version Information

**Recommendations:**

- Provide metadata fields that flag when a dataset is complete and also flag when the dataset has changed.
- Ensure the version number in metadata fields is modified whenever the data have been updated.

**Rationale:**

It is important for data users to know when a particular dataset is complete and when it has changed. Proper versioning not only ensures that users do not unknowingly use stale data but also supplies users with the ability to quickly determine whether or not they have the most current version of the data.

#### 4.1.2 Data Integrity

**Recommendations:**

- Ensure metadata records include fields for both the file size and a checksum (preferably MD5) for each granule file.
  - Checksums can be used to track updates to granule files.
- Provide a complete listing of files for datasets that include a variable number of tiles so that a user can determine which tiles are legitimately missing.

**Rationale:**

The data catalog should include enough information to allow the user to verify whether they have obtained an accurate and complete copy of the data. The file size, in conjunction with a strong checksum, can serve as an initial verification during the download stage. Strong checksums prevent unnecessary downloading and reprocessing of unchanged files after a version update. In addition, a list of legitimately missing tiles clarifies for a user whether all available data has been downloaded and whether data should ever be provided for a given tile.

### 4.2 Data

#### 4.2.1 Data Formats

**Recommendations:**

- Utilize standard file formats and minimize the use of esoteric file formats.

**Rationale:**

Data provided in a standard file format do not require special readers and as such can be quickly and easily used. These standard formats also allow data to be imported into major GIS packages without additional effort.

#### 4.2.2 Data Projections

**Recommendations:**

- Utilize standard (common) map projections and coordinate systems where possible.
- Provide and support software tools to handle specialized projections and coordinate systems (GDAL, 2017; Proj4 2017, etc.).

**Rationale:**

The utilization of common coordinate systems and map projections minimizes the possibility of misuse of data and failure of tools during data use.

## 5. Other Support Services

These recommendations cover customer (user) support services needed specifically for customers utilizing large volumes of Earth observation data.

### 5.1 Technical Support

#### Recommendations:

- Provide a clear point of contact for technical support and other questions related to data access mechanisms and networking issues. Technical support must be able to communicate directly with cloud vendor network engineers to support robust high-bandwidth connections. Ensure TCP/IP configurations on servers are optimized to maximize transfer rates.
- Establish a communication channel such as a mailing list to provide updates on upcoming datasets, information about outages, and other relevant news related to data feeds.

#### Rationale:

A clear point of contact enables cloud vendors to distribute data with high reliability and ensures a timely response. A rapid response time is important due to the large number of end users that may be indirectly affected.

## 6. Conclusion

The recommendations presented in this paper are distilled from use cases provided by three cloud vendors (Amazon, Google, and Microsoft) and are based on their interactions with data systems at different Federal agencies and organizations. These specific recommendations range from obvious steps for improving data usability (such as ensuring the use of standard data formats and commonly supported projections), to non-obvious undertakings important for enabling bulk data downloads at scale. These recommendations identify the need for a single, authoritative catalog that contains all the metadata records for the data holdings of a federated data system, as well as the need for improved data management processes to ensure this catalog is consistently updated and the metadata records accurately represent the data holdings at any given time. While improving the overall existing metadata quality is always important for facilitating data search, access, and usability, these recommendations identify specific fields (e.g., data version and checksum) in the metadata record that are often overlooked. It is important for data centers to ensure that these fields have valid entries that are both correct and consistent. The need for new services such as automated data notifications, as well as new types of user support, are part of these recommendations.

These recommendations have been used as guidelines in evaluating NASA EOSDIS (2017), and in identifying gaps and areas of improvement to support transfers of large volumes of data. These guidelines are applicable to data centers at other agencies and organizations and can be used for evaluating and improving existing data systems in the context of high-volume data transfers, thereby improving overall interoperability and data usability.

## Competing Interests

The authors have no competing interests to declare.

## References

**Executive Office of the President** 2013 National Strategy for Civil Earth Observations, National Science and Technology Council. Available at: [https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/nstc\\_2013\\_earthobsstrategy.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/nstc_2013_earthobsstrategy.pdf).

**GDAL** 2017 Geospatial Data Abstraction Library. Available at: <http://www.gdal.org/>.

**NASA EOSDIS** 2017 Earth Observing System Data and Information System. Available at: <https://earthdata.nasa.gov/about>.

**National Science and Technology Council Committee on Environment, Natural Resources, and Sustainability and U.S. Group on Earth Observations Subcommittee Data Management Working** 2016 Common Framework for Earth-Observation Data, National Science and Technology Council. Available at: [https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/common\\_framework\\_for\\_earth\\_observation\\_data.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/common_framework_for_earth_observation_data.pdf).


**Json.org** 2017 Introducing JSON – Java Script Object Notation. Available at: <http://www.json.org/>.

**Overpeck, J T, Meehl, G a, Bony, S and Easterling, D R** 2011 Climate data challenges in the 21st century. *Science (New York, N.Y.)*, 331(6018): 700–702. DOI: <https://doi.org/10.1126/science.1197869>  
**Proj.4** 2017 proj.4. Available at: <http://proj4.org/>.

**How to cite this article:** Ramachandran, R, Lynnes, C, Baynes, K, Murphy, K, Baker, J, Kinney, J, Gold, A, Sundwall, J, Korver, M, Lieber, A, Vambenepe, W, Hancher, M, Moore, R, Erickson, T, Henretig, J, Zwiefel, B, Patrick-Ahlstrom, H and Smith, M J 2018 Recommendations to Improve Downloads of Large Earth Observation Data. *Data Science Journal*, 17: 2, pp. 1–6, DOI: <https://doi.org/10.5334/dsj-2018-002>

**Submitted:** 05 July 2017    **Accepted:** 17 November 2017    **Published:** 24 January 2018

**Copyright:** © 2018 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

**OPEN ACCESS** 