Data File Requirements and Recommendations for Data Producers

Version 2.3 (Last updated March 2023)

The following guidance is intended to preserve and describe NASA Earth Science data in ways that make them discoverable, accessible, understandable, and reproducible now and into the future.

Practices identified as “required” are system-level requirements which must be met for data to be ingested and archived in the NASA DAAC systems.

For additional guidance on the creation of NASA Earth Science data products, please refer to the *Data Product Development Guide for Data Producers*.

https://www.earthdata.nasa.gov/esdis/esco/standards-and-references/data-product-development-guide-for-data-producers

Contents

- Data Formats
- Filenames
- File Structure and Content
  - Variables
  - Time
  - Geolocation
  - ASCII
- Browse Files
- Data Creation Checklist

Data Formats

- Use a NASA Earth science standard data format
- Avoid proprietary formats
  - They're challenging for users now and may not be readable in the future
  - NASA standards do allow proprietary for raw/Level-0 data

Links to overviews and resources for NASA Earth science standard data formats

NASA Earthdata netCDF-4/HDF5 File Format
https://www.earthdata.nasa.gov/esdis/esco/standards-and-references/netcdf-4hdf5-file-format

NSIDC Guidelines for NetCDF Attributes
https://nsidc.org/sites/default/files/documents/other/nsidc-guidelines-netcdf-attributes.pdf

NASA Earthdata HDF5
https://www.earthdata.nasa.gov/esdis/esco/standards-and-references/hdf5

NASA Earthdata ASCII File Format Guidelines
https://www.earthdata.nasa.gov/esdis/esco/standards-and-references/ascii-file-format-guidelines-for-earth-science-data

NASA Earthdata GeoTIFF
https://www.earthdata.nasa.gov/esdis/esco/standards-and-references/geotiff

OGC GeoTIFF standard
http://docs.opengeospatial.org/is/19-008r4/19-008r4.html

Data Formats, continued

NetCDF4 & HDF5

- Good for multidimensional data
- Capable of holding rich metadata, being self-describing
- Interoperable with variety of computational platforms and protocols (e.g., OPeNDAP)

GeoTIFF

- User friendly format; most requested format by NSIDC users
- Widely interoperable with GIS, image processing, and map server applications

Shapefile

- Good for feature data (e.g., points, lines, polygons)

Type of data	Recommended formats
Tabular or site-based data	Delimited ASCII/CSV
	HDF5
	NetCDF4
Raster	GeoTIFF
	HDF5
	netCDF4
	Delimited ASCII/CSV
Vector	LAS 1.4 (LAZ)
	Shapefile
	Delimited ASCII/CSV
Photos/Movies	JPEG/MPEG

Filenames

- Filenames must be unique*
- Begin with data product identifier* (NSIDC DAAC provides identifier)
- Include ASCII characters only; no spaces*
- Cannot exceed 80 characters*
- File extension indicates data format*

- Include descriptive information such as:
  - Mission / campaign
  - Instrument
  - Measurement
  - Spatial (site, resolution)
  - Temporal (date, time)
  - Processing version

<ProductID>_<DescriptiveInfo>.<extension>

ICESat-2 example
ATL06_20220404104324_01881512_005_01.h5

Icebridge example
ILATM2_20090422_131620_smooth_nadir3seg_50pt.csv

MEaSUREs example
NSIDC-0630-EASE2_N3.125km-F18_SSMIS-2022166-37H-M-SIR-CSU_ICDR-v1.5.nc

*required

File Structure and Content: Variables

- Define unique, interpretable names for each variable
- Define units for each variable
- Define missing data as a numeric value outside the range of valid data values, and use consistently (e.g., -9999, and not NaN)

International Standards for Variable Names and Units

ISO 8601
Representation of dates and times

International System of Units

CF MetaData
CF Standard Names and Controlled Vocabulary

File structure and content: Time

- Define time standard and time zone used
  - Recommend using UTC
  - Timestamps may be reported as UTC decimal seconds from the time at which measurements began (commonly as seconds past midnight)
- Use standard date/time formats
  - Recommend using yyyymmdd, hhmmss, or yyyymmddTHHMMSS.SSSZ (ISO 8601 standard)

File Structure and Content: Geolocation

- Include appropriate geospatial coordinates for measurements in the data file
  - e.g., for files in a geodetic data product: longitude, latitude, and height (if applicable); or for files in a projected data product: easting, northing, and height (if applicable)
- Use a consistent coordinate format and define units
  - e.g., lat/lon decimal degrees, northing/easting meters, y/x meters, etc.
- Provide coordinate reference system (CRS)
  - e.g., via an EPSG code (see epsg.io), and/or the projection, horizontal datum, ellipsoid, and if applicable, the vertical datum.

File Structure and Content: ASCII

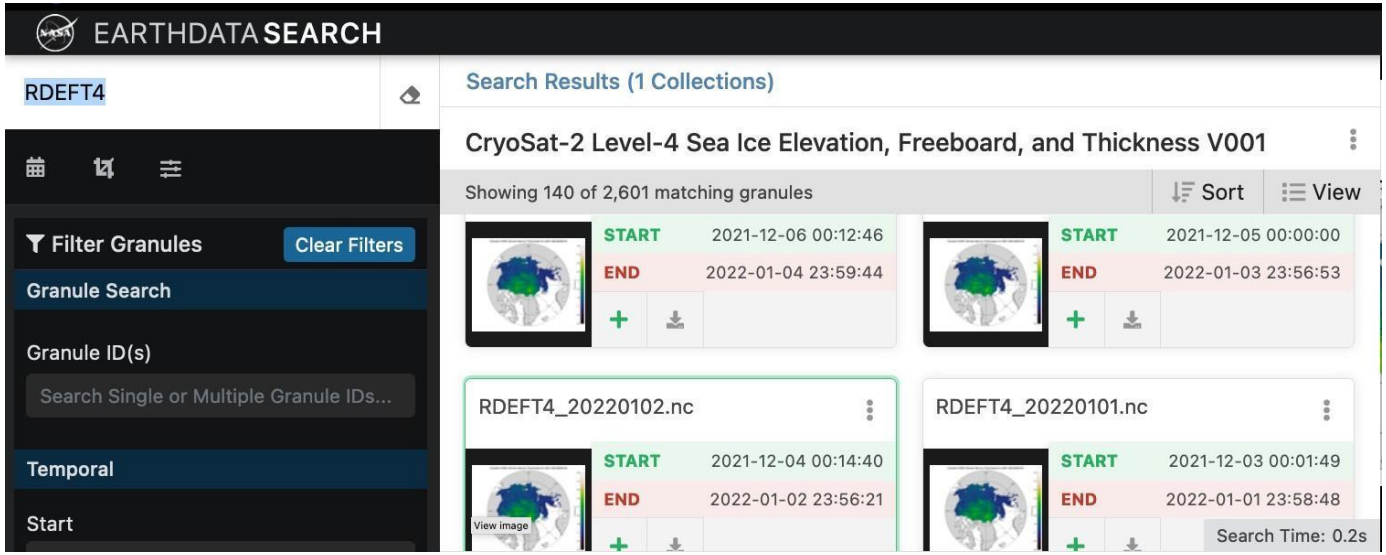
- Include separate header and data sections within files
  - Header needs to be clearly delineated from data rows (e.g., begin with #)
- Use consistent delimiter between data value
  - Visible characters are preferred (e.g., comma, semi-colon, colon, |)
- Separate rows with end-of-line character
  - Mac: CR
  - Unix: LF
  - Window: CR/LF
- Do not use empty lines or rows

Browse Files

Browse images are optional, but when included:

- They provide potential data users the ability to evaluate the data before downloading.
- They're ingested and archived with the data files.

- One browse image is ingested per data file.
- Preferred formats are: .jpeg or .png
- Should range from 200 to 600 pixels per image - smaller is better as long as quality does not suffer.



Example of browse images displayed for each data file in the search results of Earthdata Search.

Data Creation Checklist

- Data formats
  - Use a NASA Earth science data format
  - Avoid proprietary formats
- Filenames
  - Use descriptive, unique names*
  - Product identifier begins each filename*
  - ASCII characters only; no spaces*
  - Max 80 characters*
  - File extension indicates data format*
- File contents/structure: Variables
  - Unique, interpretable variable names
  - Define units for each variable
  - Consistently use a numeric* missing data value
- File contents/structure: Time
  - Define time standard and time zone used: recommend UTC
  - Use standard date/time formats
- File structure/content: Geolocation
  - Include geographic coordinates
  - Use a consistent coordinate format
  - Provide coordinate reference system
- File structure/content: ASCII
  - Include and delineate header from data section
  - Use consistent delimiter between data values: visible characters preferred
  - Separate rows with EOL
  - Do not use empty lines or rows
- Browse Files (optional)
  - One browse image per data file
  - Preferred file formats: .jpeg or .png
  - 200 to 600 pixels per image

*required