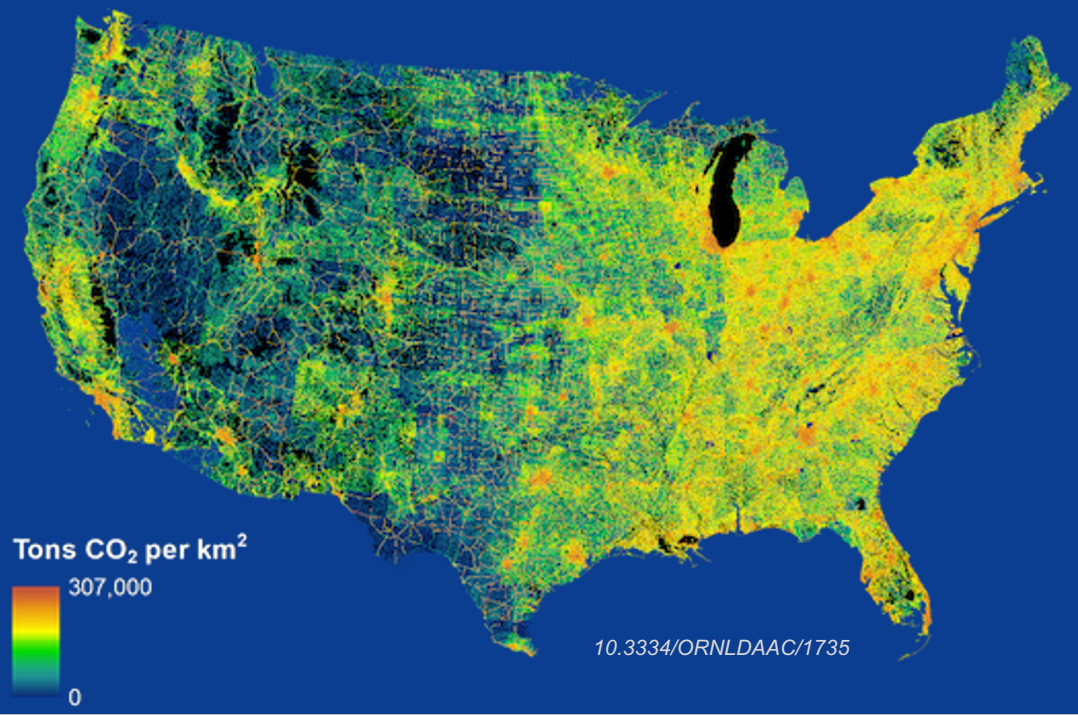
NACP 7th OSM Data Management Workshop

Yaxing Wei, Michele Thornton, Jessica Welch, Kay Shope, Tammy Walker

Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)

March 24th, 2021



The Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics operates under an interagency agreement between NASA and the U.S. Department of Energy

Workshop Agenda

I. Benefits of Data Management and Publication
   Yaxing Wei, 20 mins

II. Best Practices of Data Management
   Michele Thornton, 25 mins

III. Data Quality Assurance
   Jessica Welch, 30 mins

IV. Submitting Data to the ORNL DAAC
   Yaxing Wei, 15 mins

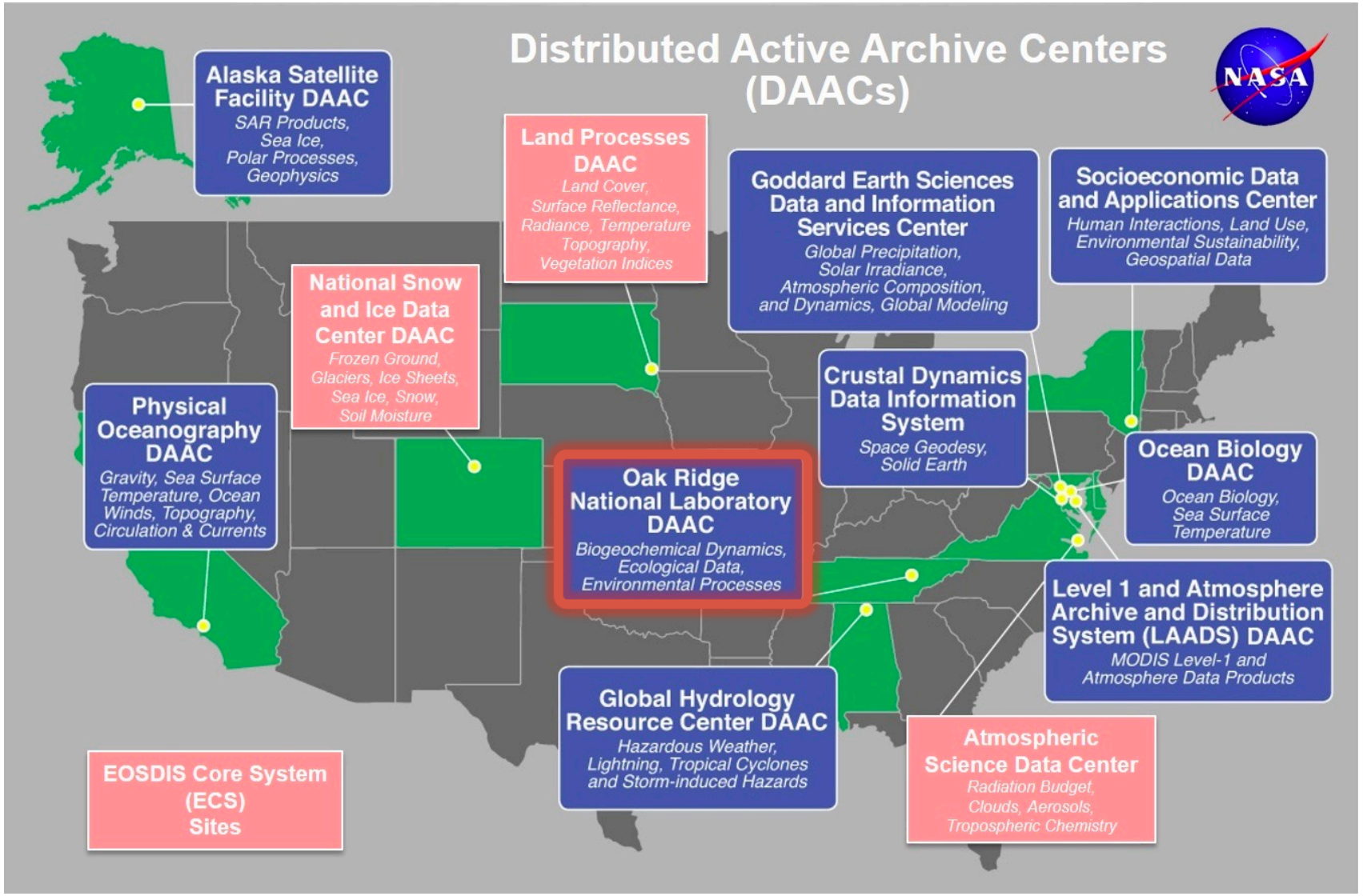
V. Question & Answer Session
   Kay Shope & Tammy Walker, 30 mins

Workshop Goals

Provide information that investigators can use to

- Ensure proper data curation
- Increase the usability of data
- Ensure long-term data preservation
- Encourage open science and reproducible research

The ORNL DAAC

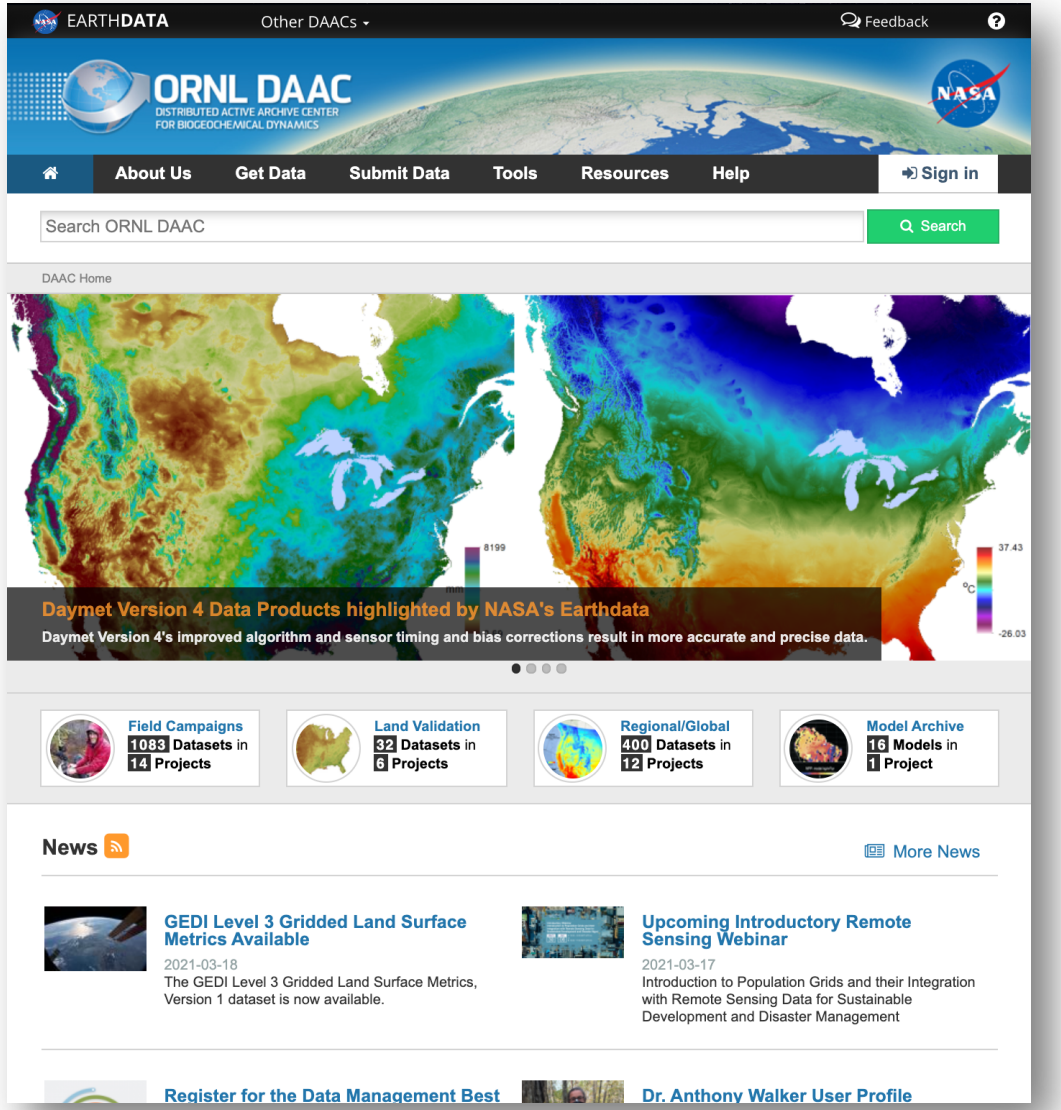


The ORNL DAAC

daac.ornl.gov/

- Established in 1993
- ORNL DAAC Mission

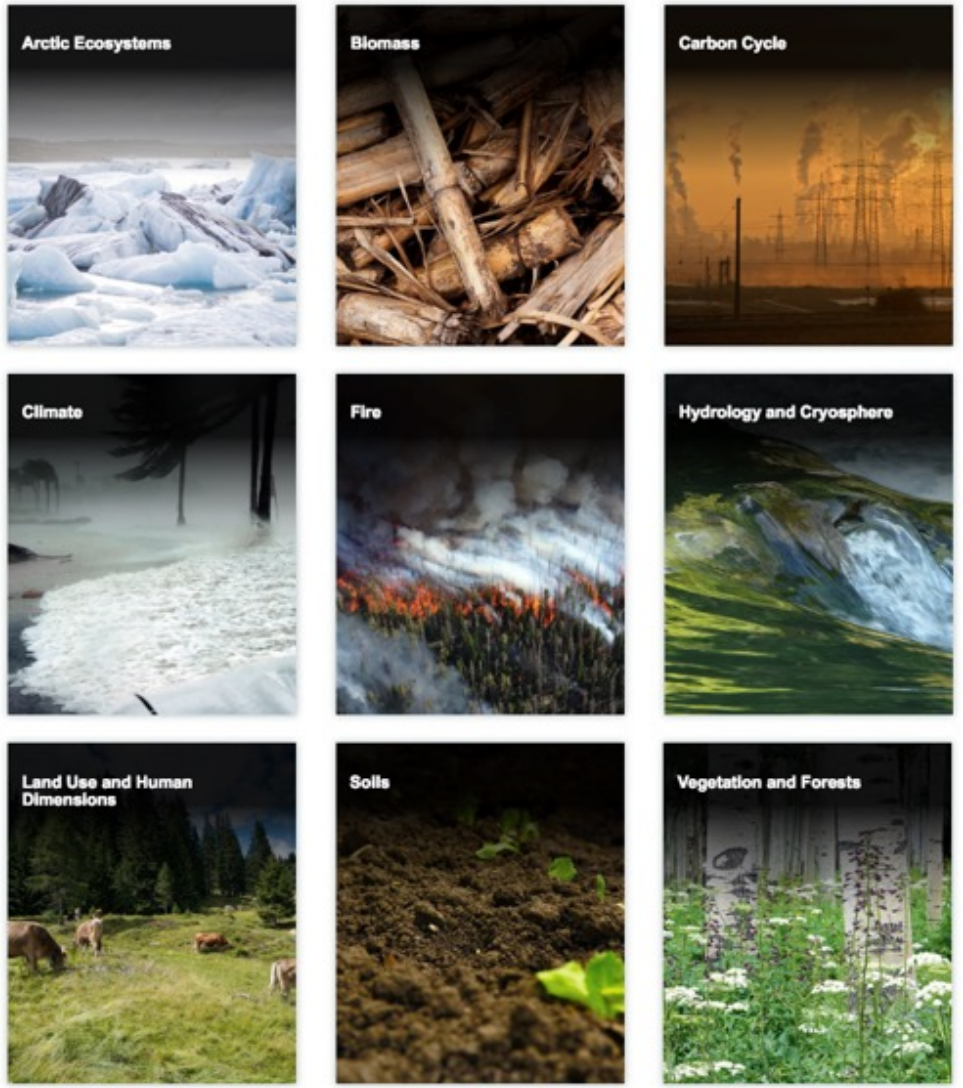
  Assemble, distribute, and provide data services for a comprehensive archive of terrestrial biogeochemistry and ecological dynamics observations and models to facilitate research, education, and decision-making in support of NASA's Earth Science.



Data at the ORNL DAAC

daac.ornl.gov/get_data/

- 1,500 datasets
- 30,000 users per year
- 4,200 data citations
- 9 Science Themes



ORNL DAAC
DISTRIBUTED ACTIVE ARCHIVE CENTER
FOR BIOGEOCHEMICAL DYNAMICS

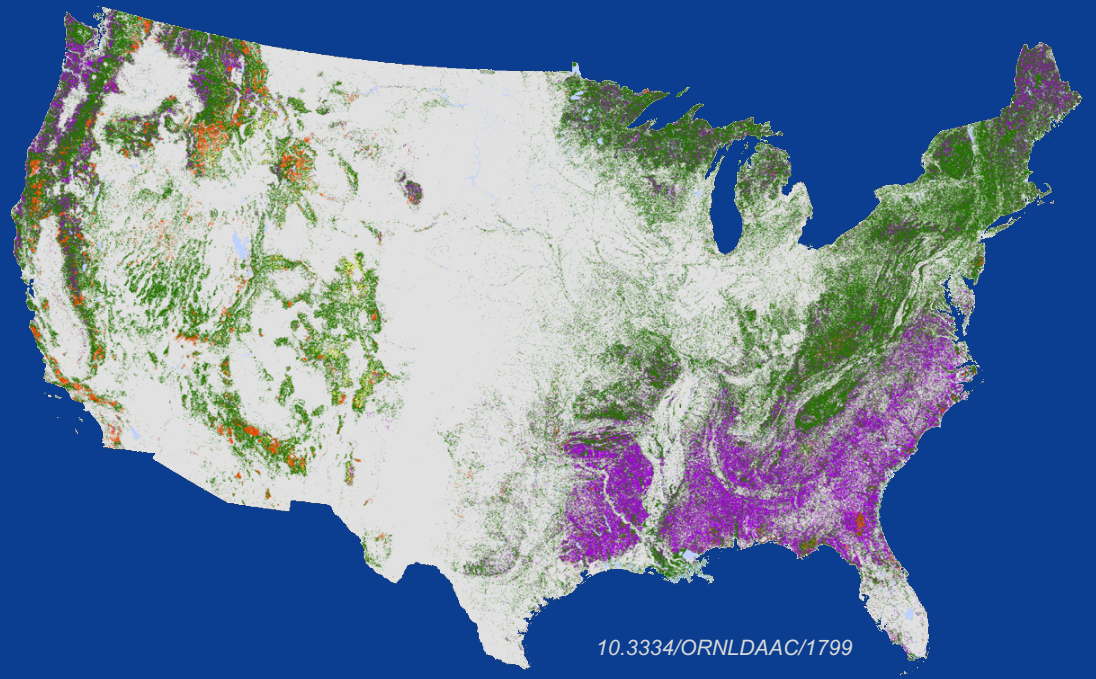
EOSDIS

Benefits of Data Management and Publication

Yaxing Wei

NACP 7th OSM Data Management Workshop

March 24th, 2021



The Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics operates under an interagency agreement between NASA and the U.S. Department of Energy

U.S. DEPARTMENT OF ENERGY

NASA

Explore Dataset Landing Pages (5 mins)

- Vulcan: High-Resolution Hourly Fossil Fuel CO2 Emissions in USA, 2010-2015, Version 3

  doi.org/10.3334/ORNLDAAC/1810

- GEDI L3 Gridded Land Surface Metrics, Version 1

  doi.org/10.3334/ORNLDAAC/1865

- Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3

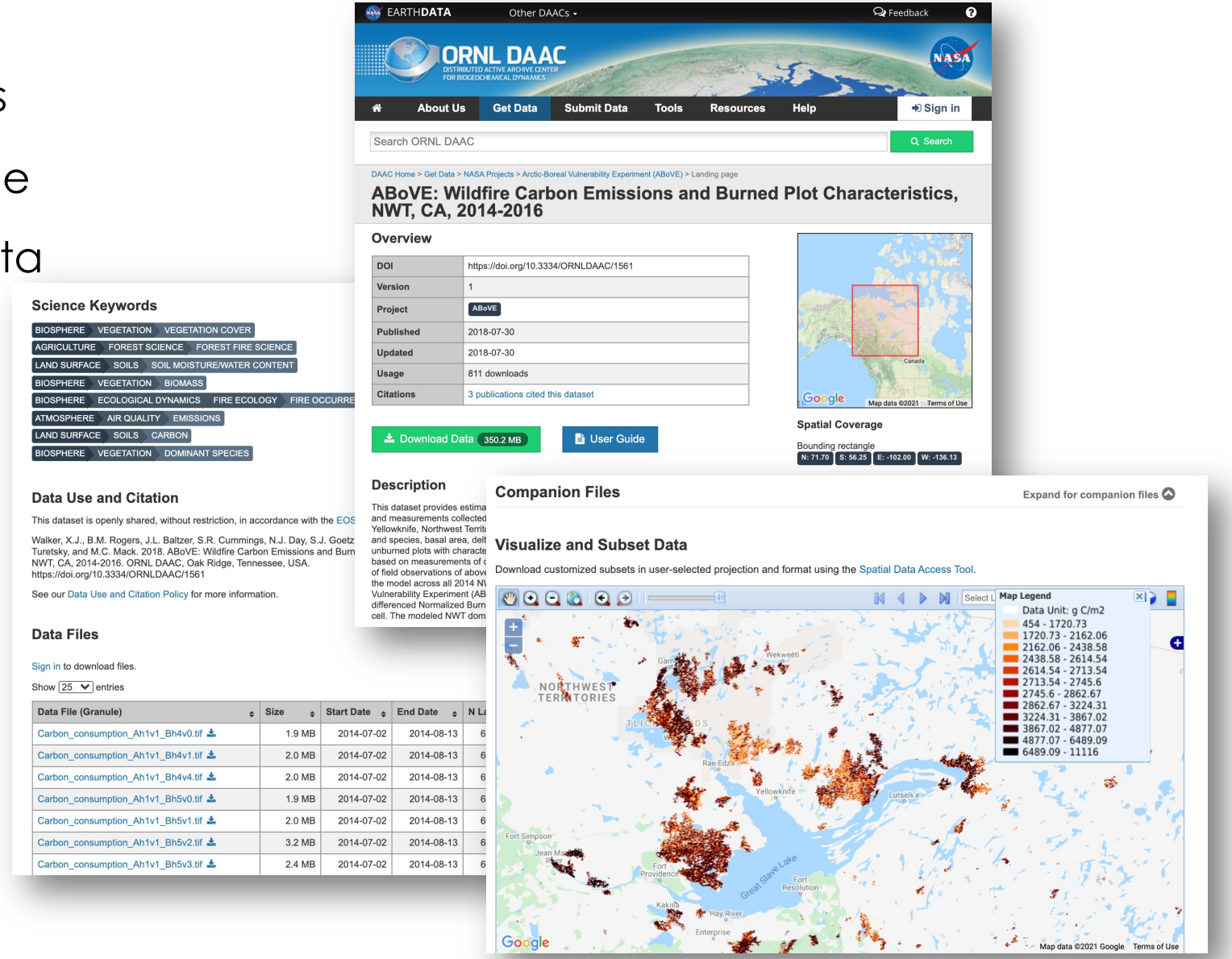
  doi.org/10.3334/ORNLDAAC/1328

- ABoVE: Wildfire Carbon Emissions and Burned Plot Characteristics, NWT, CA, 2014-2016

  doi.org/10.3334/ORNLDAAC/1561

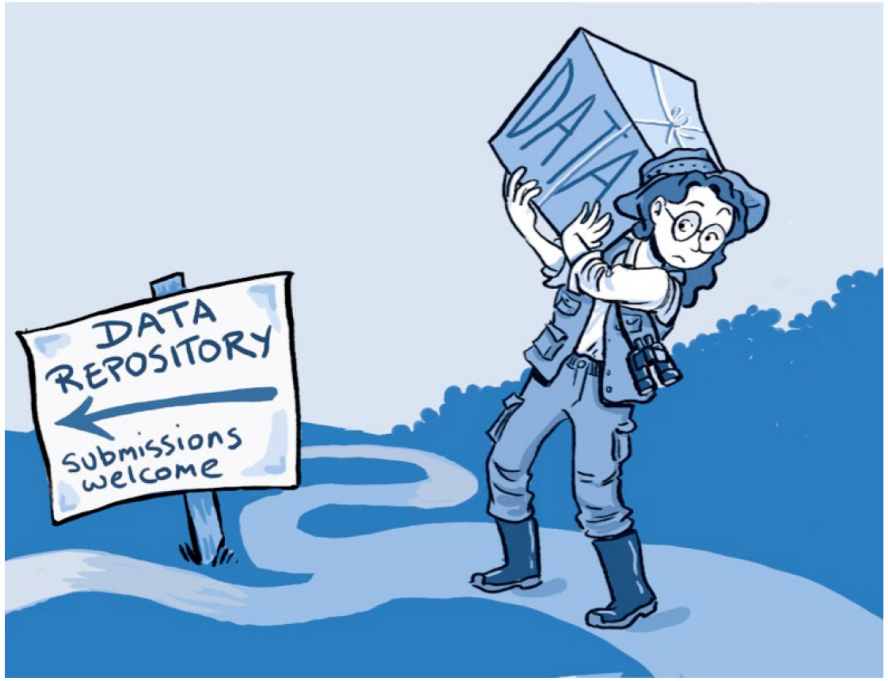
What are the benefits?

- Curated with standard formats
- Persistent dataset landing page
- Discoverable through metadata
- Release and revision dates
- Access to data files
- User Guide
- Visualization
- Web services / open access
- Subsetting / reprojecting
- Related tutorials
- Exposure



More Benefits

- Get credit for your effort
  - Digital Object Identifier and citation
- Share data for broader use
- Measure your impact
  - Citation and download statistics
- Increase access to federally-funded scientific research



From Roche DG, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779. DOI: 10.1371/journal.pbio.1001779

ORNL DAAC
DISTRIBUTED ACTIVE ARCHIVE CENTER
FOR BIOGEOCHEMICAL DYNAMICS

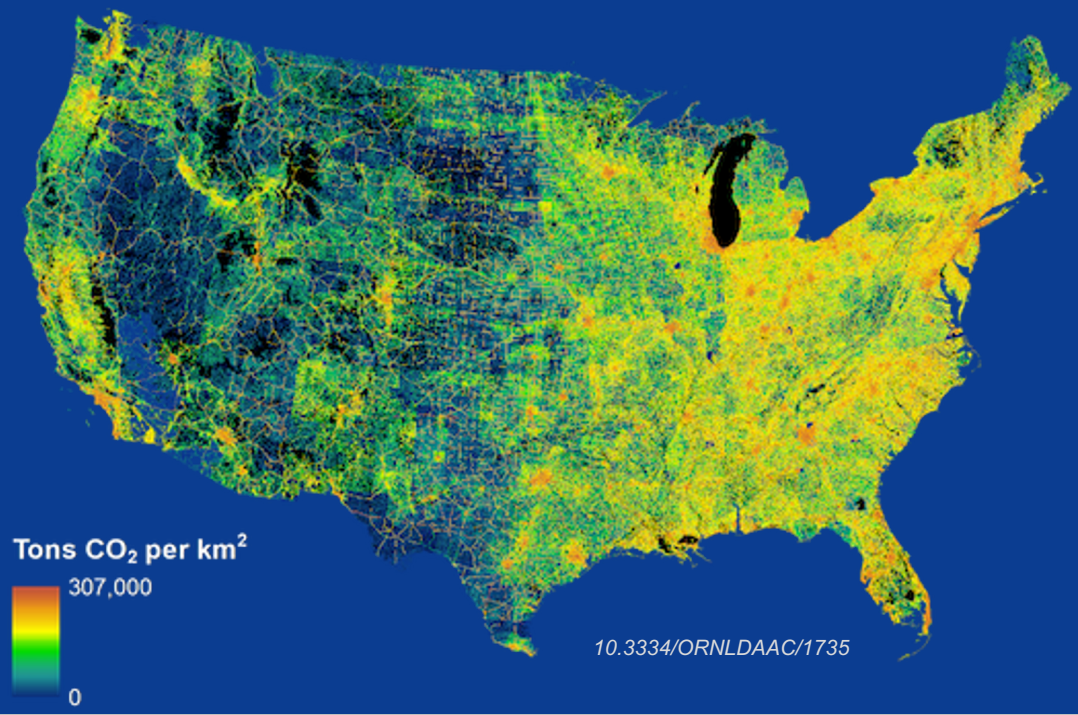
EOSDIS

Best Practices of Data Management

Michele Thornton

NACP 7th OSM Data Management Workshop

March 24th, 2021



The Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics operates under an interagency agreement between NASA and the U.S. Department of Energy

U.S. DEPARTMENT OF ENERGY

NASA

What is GOOD Data Management?

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process.

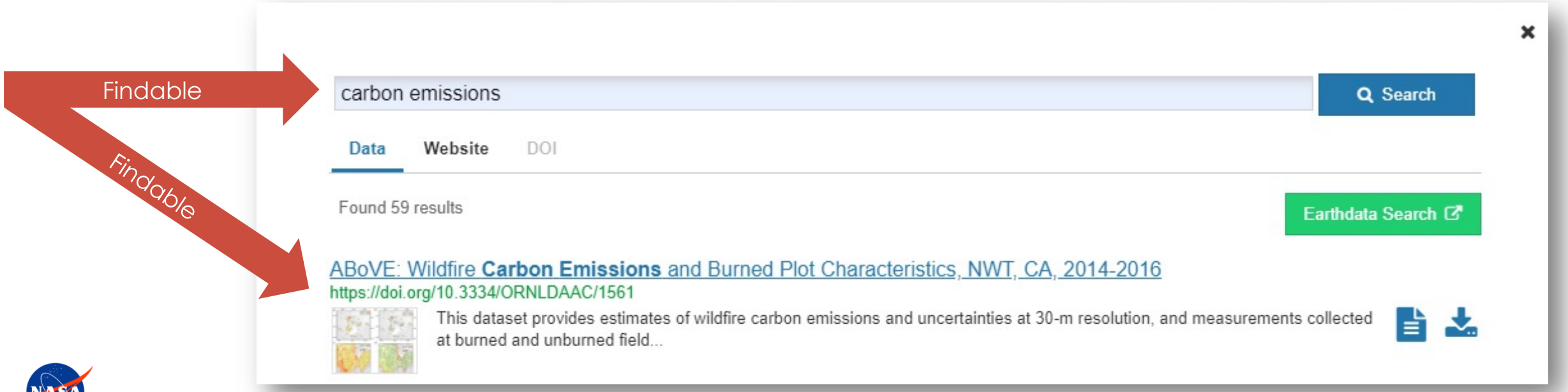
What is GOOD Data Management?

- For research: the *intentional* process of collecting, storing, processing, and protecting data
- For preservation: data are discoverable, accessible, understandable, and reusable now and into the future
- Good data management follows the **FAIR** data principles
  - **F =** Findable
  - **A =** Accessible
  - **I =** Interoperable
  - **R =** Reusable

What is GOOD Data Management?

FAIR data principles

- **F = Findable**
  - data are archived with a reputable data center
    - persistent DOI's
  - spatially and temporally defined / accurate
  - use compliant metadata / keywords



What is GOOD Data Management?

FAIR data principles

- **A = Accessible**
  - readily downloaded
  - open-standards access points

Accessible

Accessible

DAAC Home > Get Data > NASA Projects > Global Ecosystem Dynamics Investigation (GEDI) > Landing page

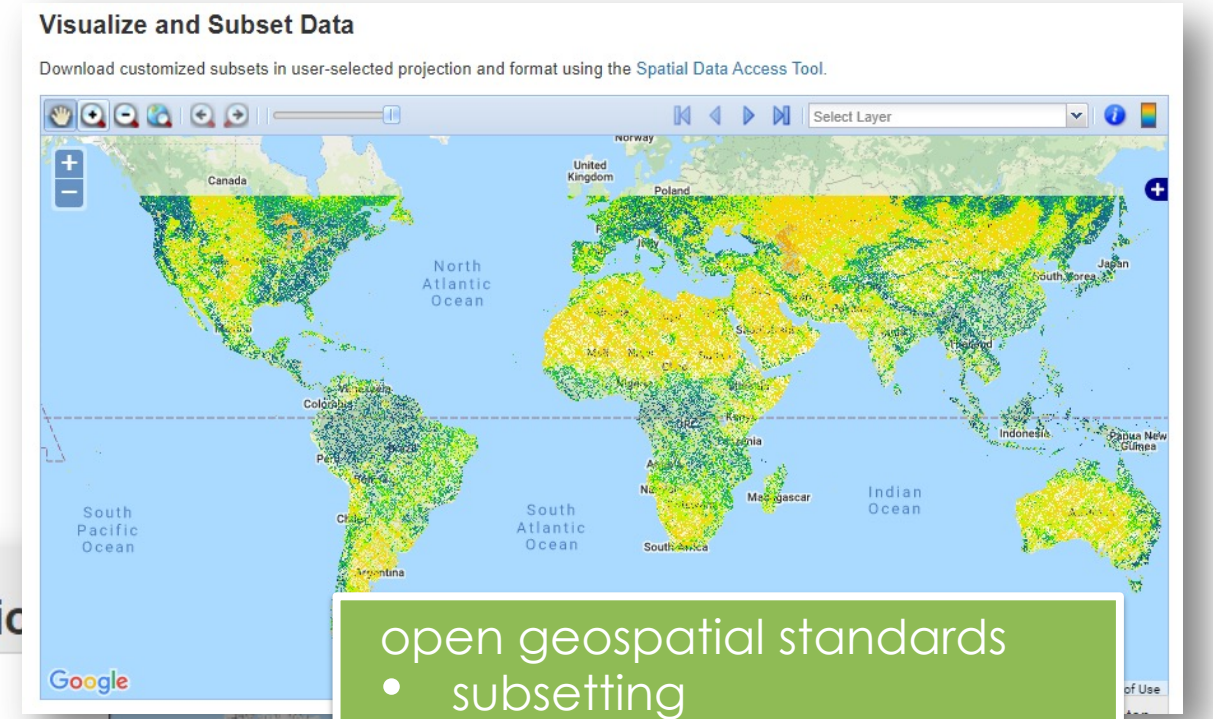
GEDI L3 Gridded Land Surface Metrics, Versic

Overview

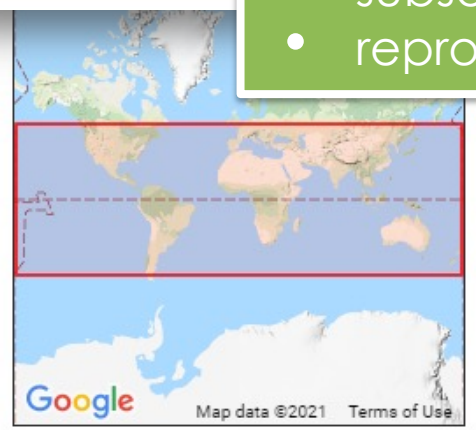
DOI	https://doi.org/10.3334/ORNLDAAC/1865
Version	1
Project	GEDI
Published	2021-03-15
Updated	2021-03-15
ge	326 downloads

Download Data 2.8 GB

User Guide



open geospatial standards
- subsetting
- reprojection



Spatial Coverage

What is GOOD Data Management?

FAIR data principles

- **I = Interoperable**
  - (meta)data can be integrated with other data
  - Data work with other applications and workflows for processing and analysis

What is GOOD Data Management?

FAIR data principles

- **R = Reusable**

DAAC Home > Get Data > NASA Projects > North American Carbon Program (NACP) > Landing page

Vulcan: High-Resolution Annual Fossil Fuel CO2 Emissions in USA, 2010-2015, Version 3

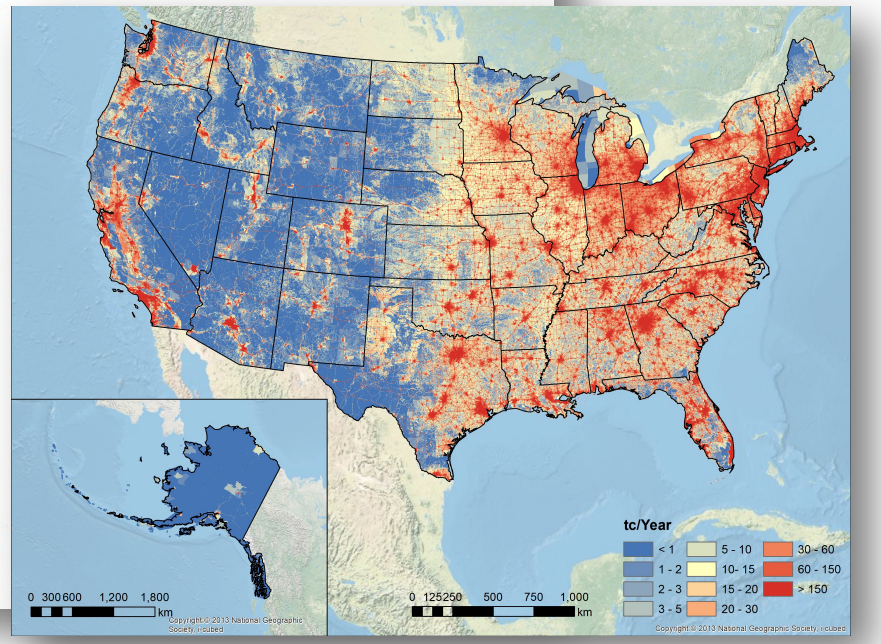
Overview

DOI	https://doi.org/10.3334/ORNLDAAC/1741
ersion	1
P	NACP
Publish	2020-10-01
Updated	2020-10-01
Usage	221 downloads

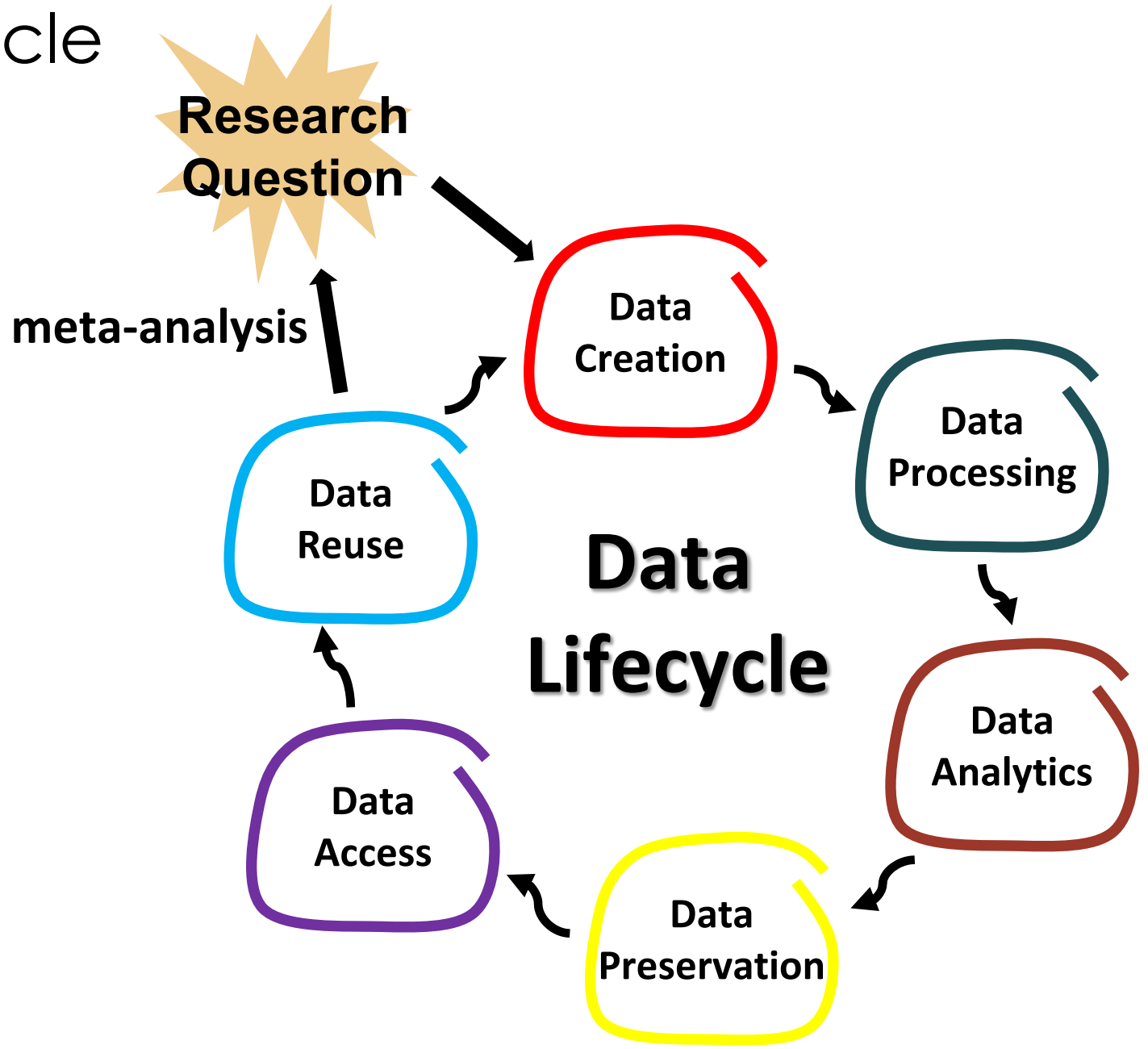
Reusable

Download Data 7.2 GB

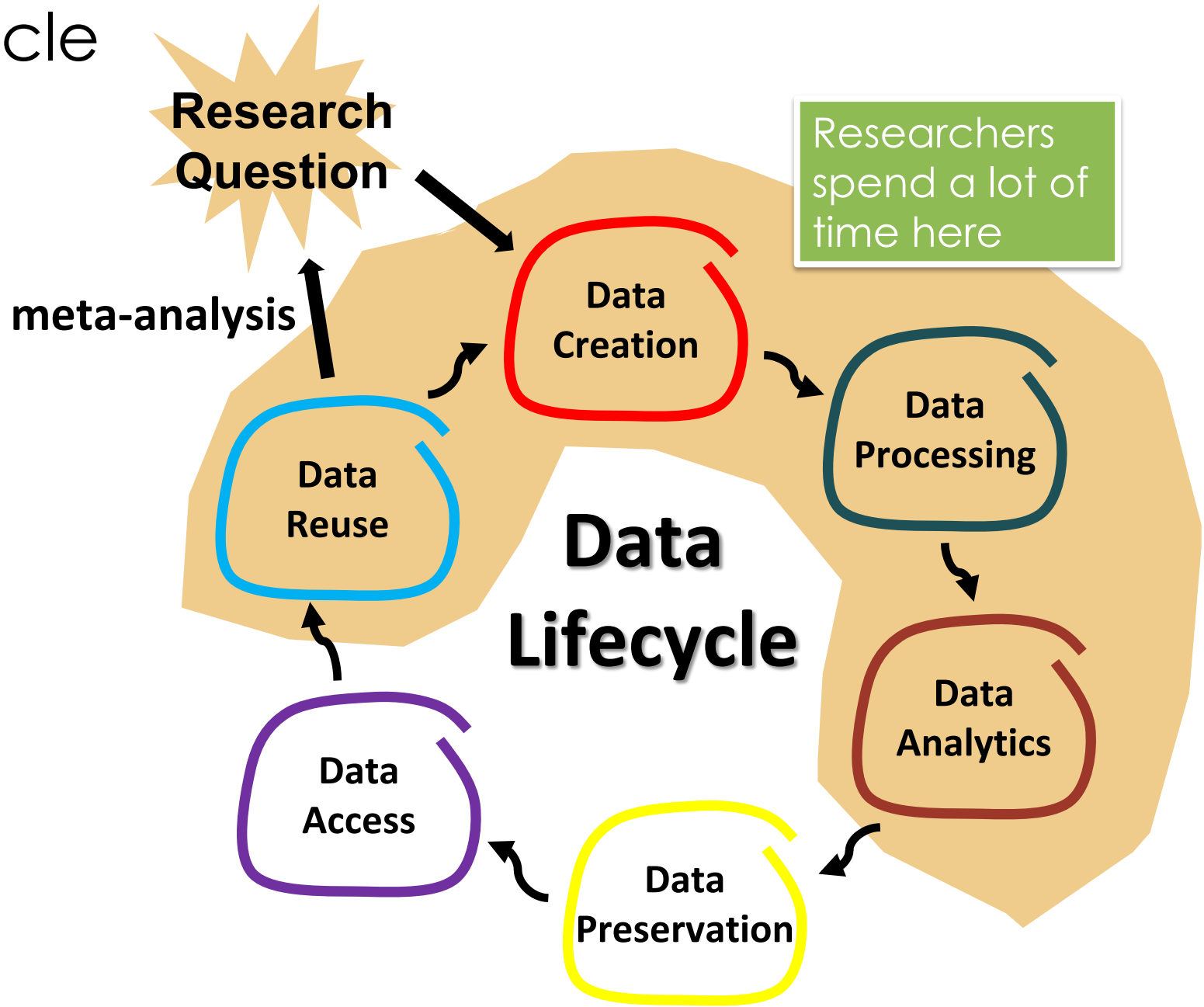
User Guide



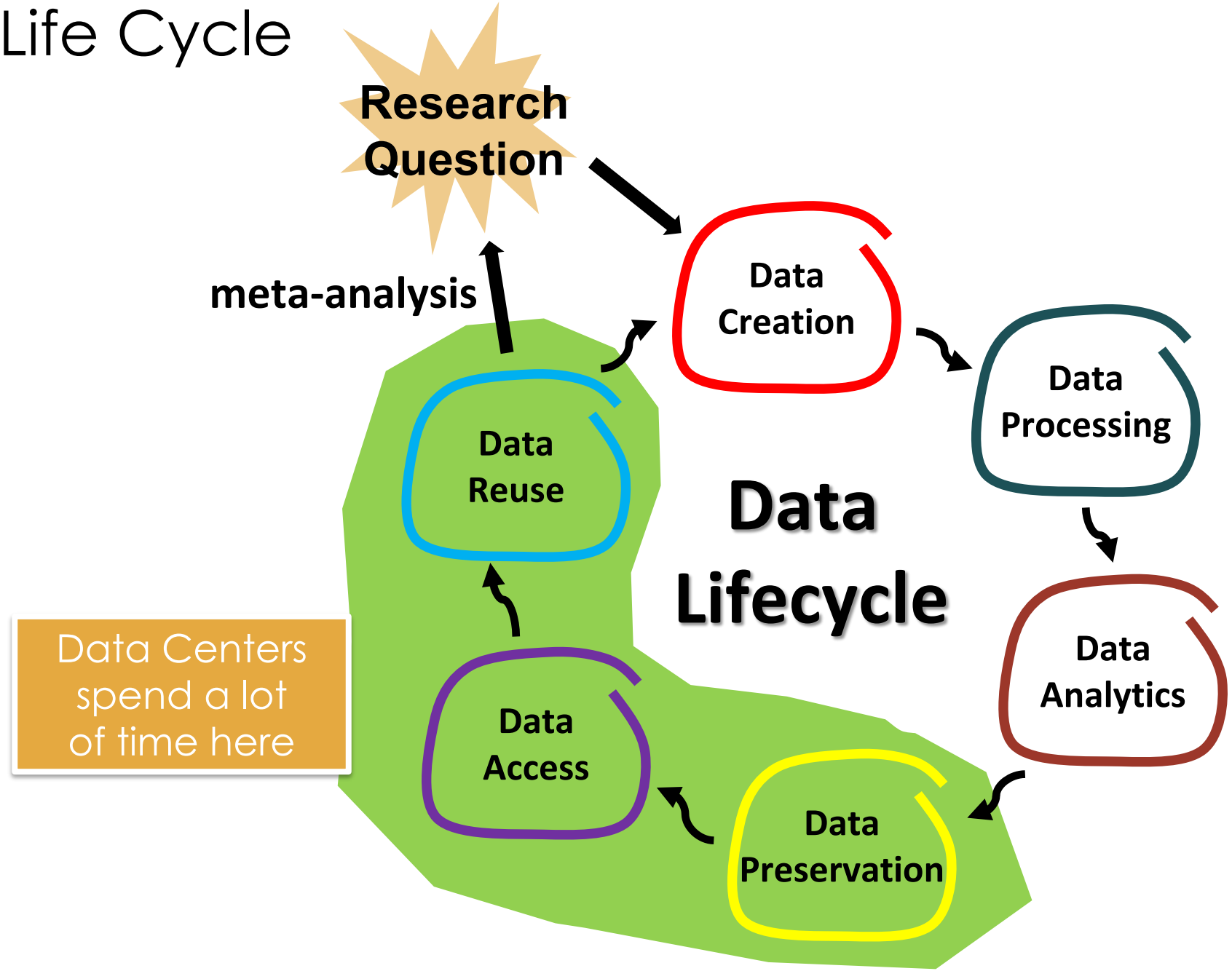
The Data Life Cycle



The Data Life Cycle

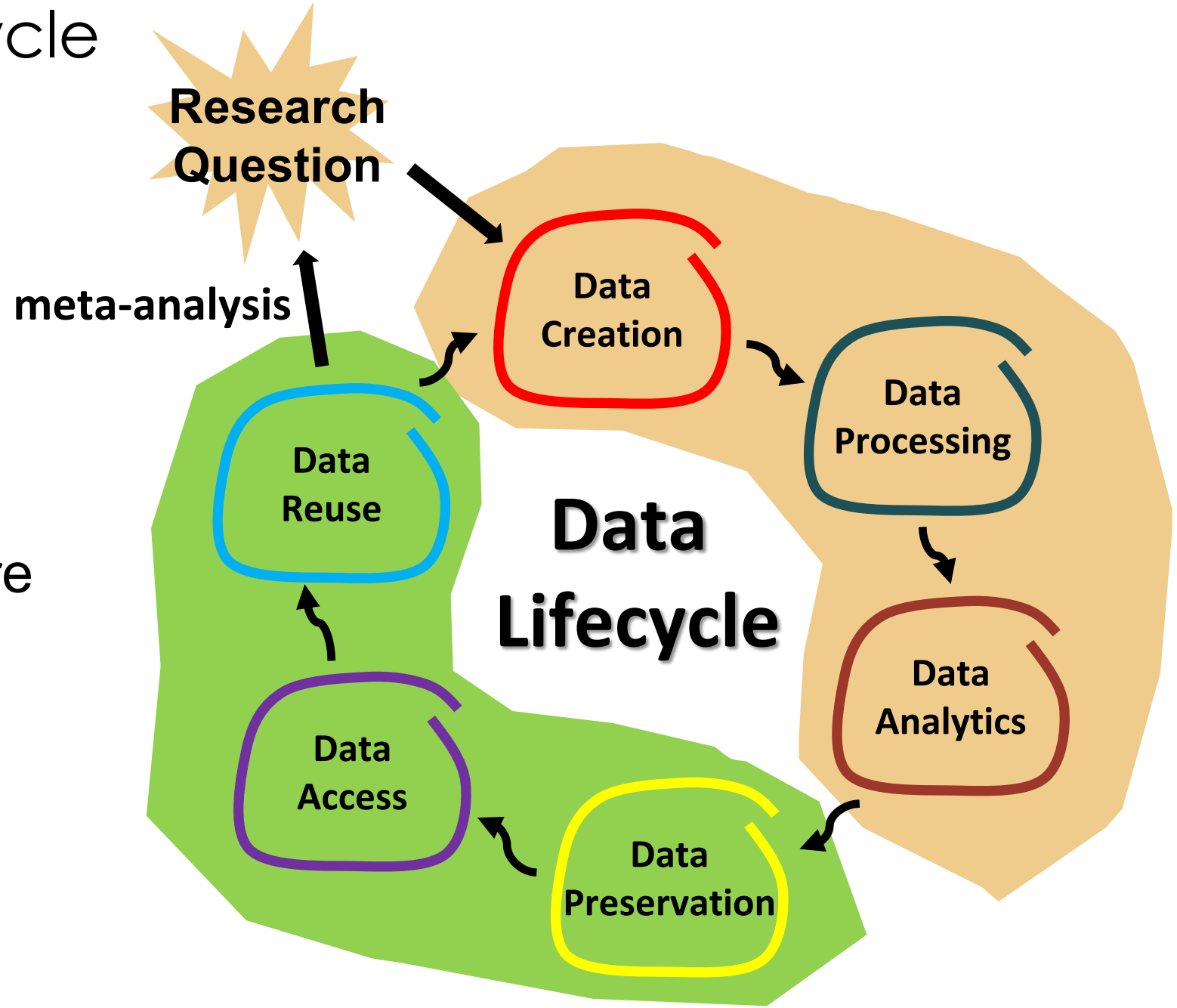


The Data Life Cycle



The Data Life Cycle

Good Data Management supports the entire Data Lifecycle



Good Data Management

- **START** … with a Data Management Plan (DMP)
- Follow protocols in standardized data and metadata
- Large and/or multi-investigator projects (e.g., ABoVE, multi-model comparisons) benefit from articulated coordination
- Start early, re-visit your DMP
- If possible, work with a data center

Good Data Management

- Data Management Plan (DMP)
  - DMP is a formal document that outlines how data are to be handled both **during** a research project, and **after** the project is completed
  - Proposals to NASA Terrestrial Ecology (TE), and many other agencies, require a DMP
  - Many helpful agency specific resources available

Good Data Management

- Data Management Plan (DMP)
  - NASA's Earth Science Division (ESD) Data Management Plan (DMP) Template for Data Producers (DPs) Includes
    - Mission/Investigation and Instrument/Sensor Overview
    - Data Product Summary
    - Data Acquisition
    - Data Processing
    - Data Analysis
    - Data Quality
    - Data Distribution and Archiving
    - Data Product Documentation

Good Data Management

- Follow protocols in standardized data and metadata
  - Define Parameters
    - Units, Date & Time, Coordinates & Geospatial Information
  - Coded Fields & Data Flags
  - Missing Values (“nodata”)
  - Use Consistent Data Organization
  - Use Archive Stable File Formats
  - Descriptive File Names
  - Use Preserved Versions of Programming to Manipulate your Data
  - Perform Basic Data Quality Assurance
  - Document your data

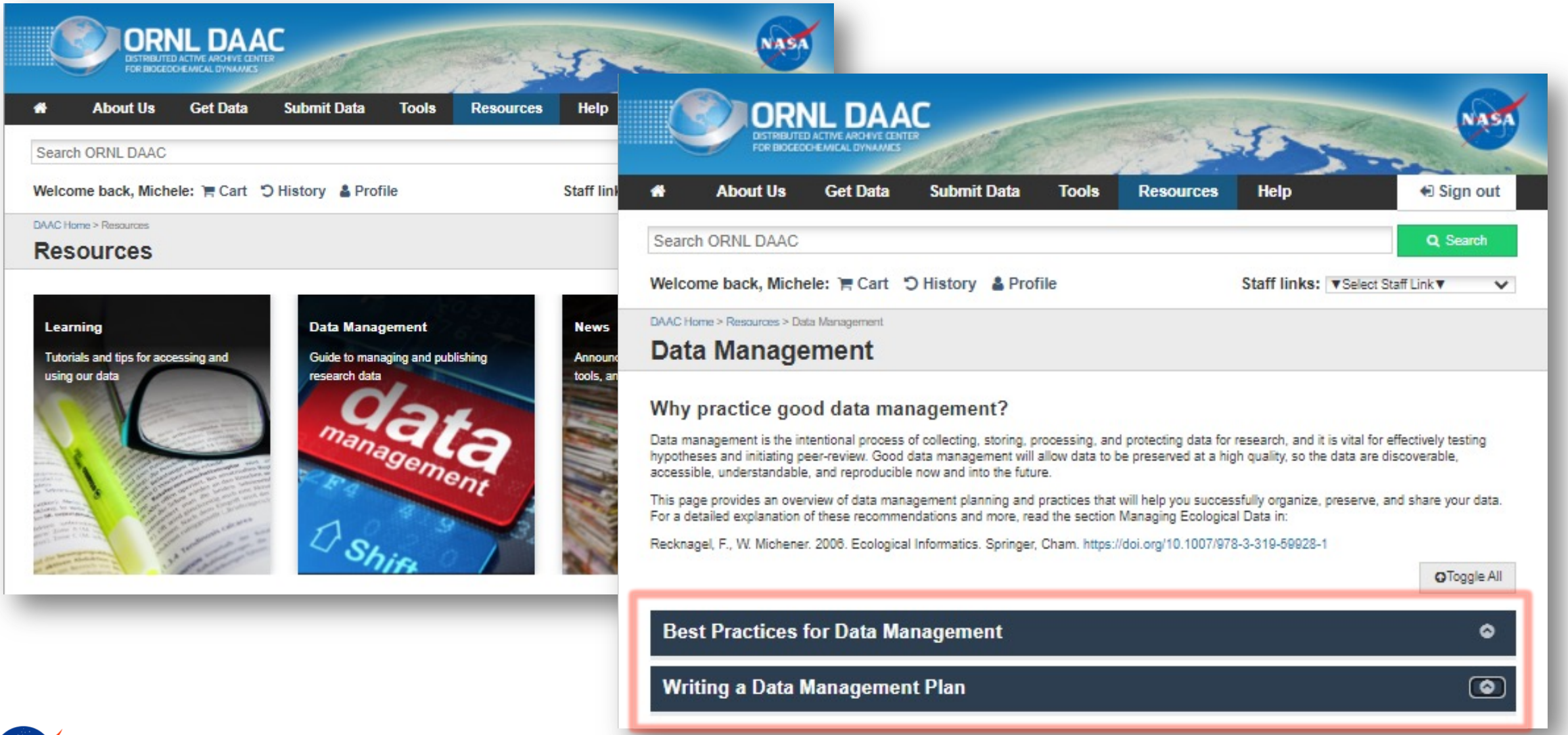
Good Data Management

- Follow protocols in standardized data and metadata
  - NASA created the Data Product Development Guide for Data Producers to guide the development of Earth science data products in ready-to-archive format
    - Ramapriyan, H. K., and P. J. T. Leonard. 2020. Data Product Development Guide (DPDG) for Data Producers version1. NASA Earth Science Data and Information System Standards Office, 9 July 2020. https://doi.org/10.5067/DOC/ESO/RFC-041VERSION1

Good Data Management = Publication Requirements

- US Government Open Data Policy (M13-13)
  - Executive Order of May 9, 2013, Making Open and Machine Readable the New Default for Government Information
  - Making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discover
- NASA Terrestrial Ecology (TE)-funded research data must be archived within NASA's Earth Observing System Data and Information System (EOSDIS)
- Many scientific journals require data be published in conjunction with the manuscript (Nature, Science, PLoS, and Ecology); a DMP can expedite the process

Data Management Resources at the ORNL DAAC



ORNL DAAC
DISTRIBUTED ACTIVE ARCHIVE CENTER
FOR BIOGEOCHEMICAL DYNAMICS

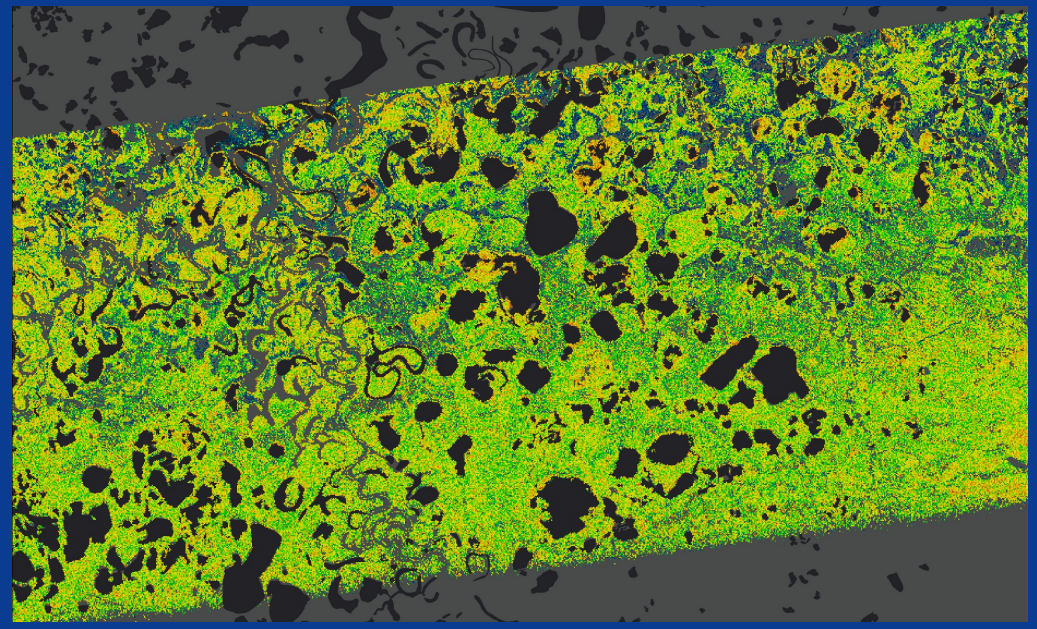
EOSDIS

Dataset Quality Assurance

Jessica Welch

NACP 7th OSM Data Management Workshop

March 24th, 2021



10.3334/ORNLDAAC/1657

The Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics operates under an interagency agreement between NASA and the U.S. Department of Energy

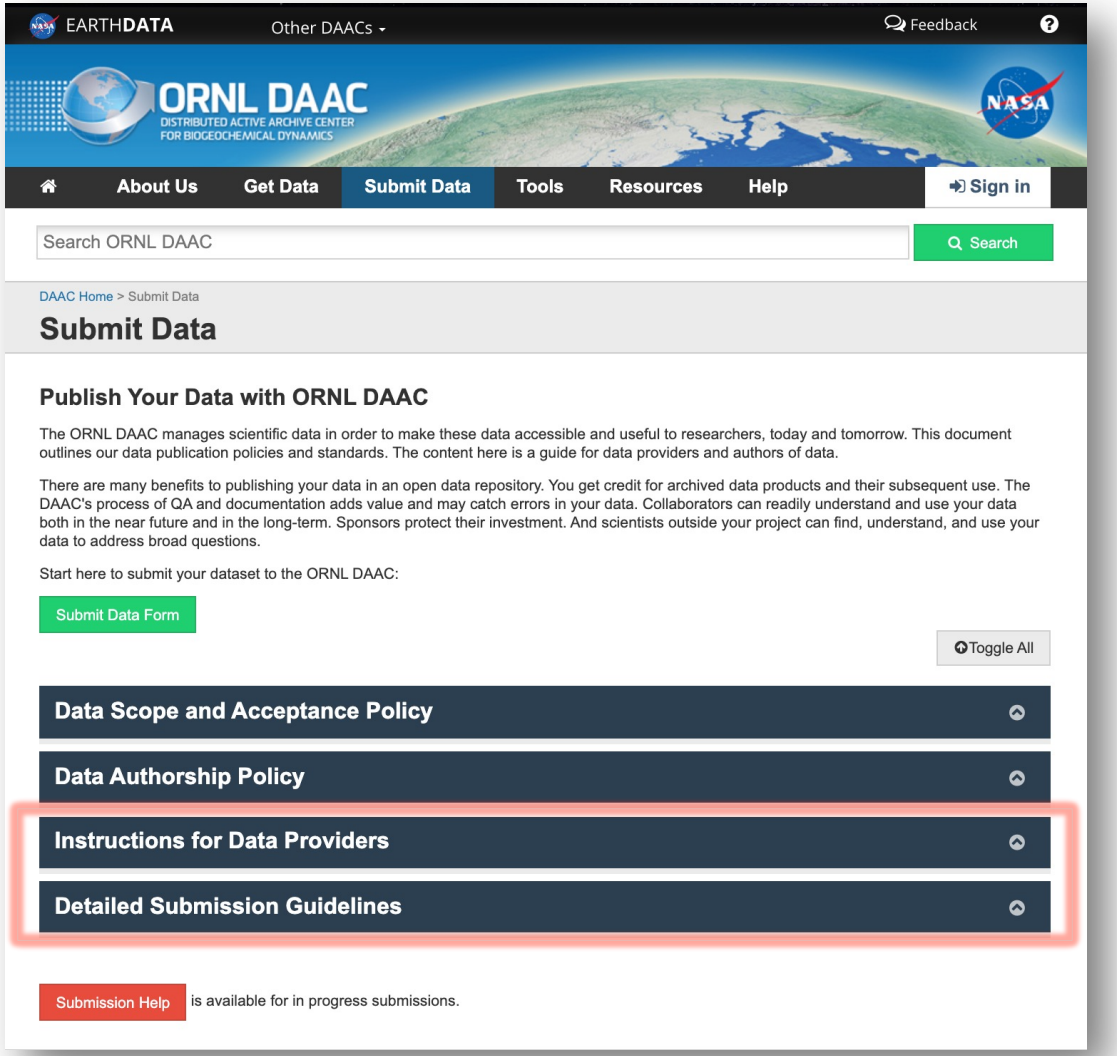
U.S. DEPARTMENT OF ENERGY

NASA

What is Dataset QA?

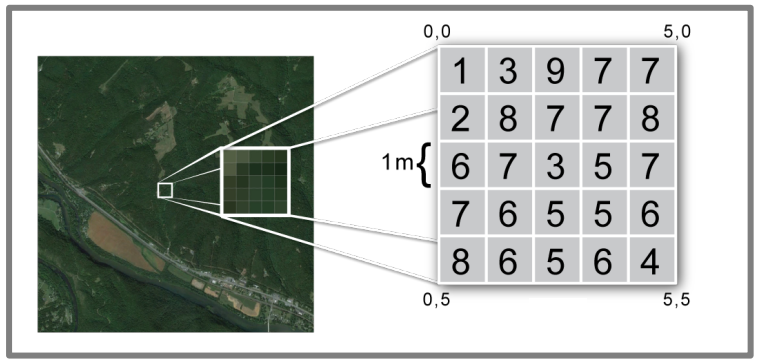
daac.ornl.gov/submit/

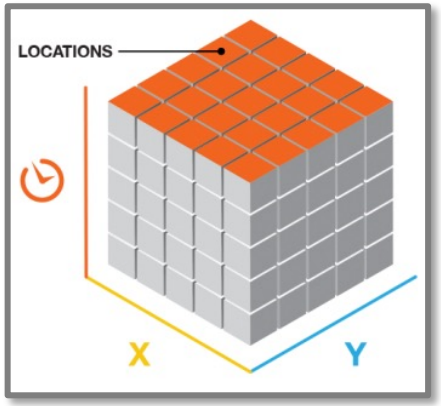
- Processes that ensure the dataset is ready for release
- Includes systematic checks, comparison with a standard, and error detection
- Important for proper collection of metadata that will make a dataset findable

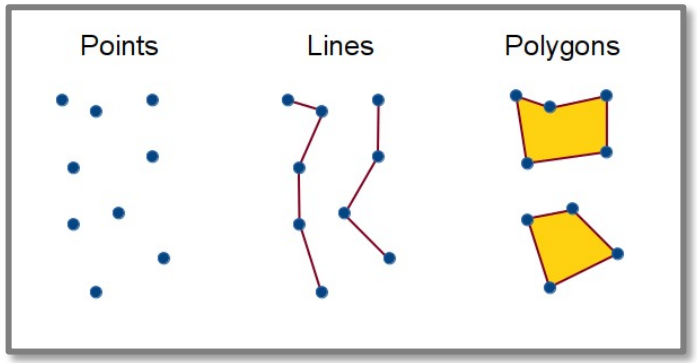


Archive Stable File Formats

- Tabular
  - Delimited text (CSV, TXT)
- Raster
  - GeoTIFF (TIF)
  - netCDF (NC, NC4)
  - HDF (H5, HDF5)
- Vector
  - Shapefiles (SHP)
  - Google (KML)







QA Exercise (5 mins)

daac.ornl.gov/resources/workshops/nacp-7th-osm-workshop/

- Examine each of the demonstration data files and look for QA issues.
  - soil_CO2_flux_final_OLD.csv
  - CAPECOD_final_Jones112756389.tif
  - snow depth 00500m 2006 qademo.nc

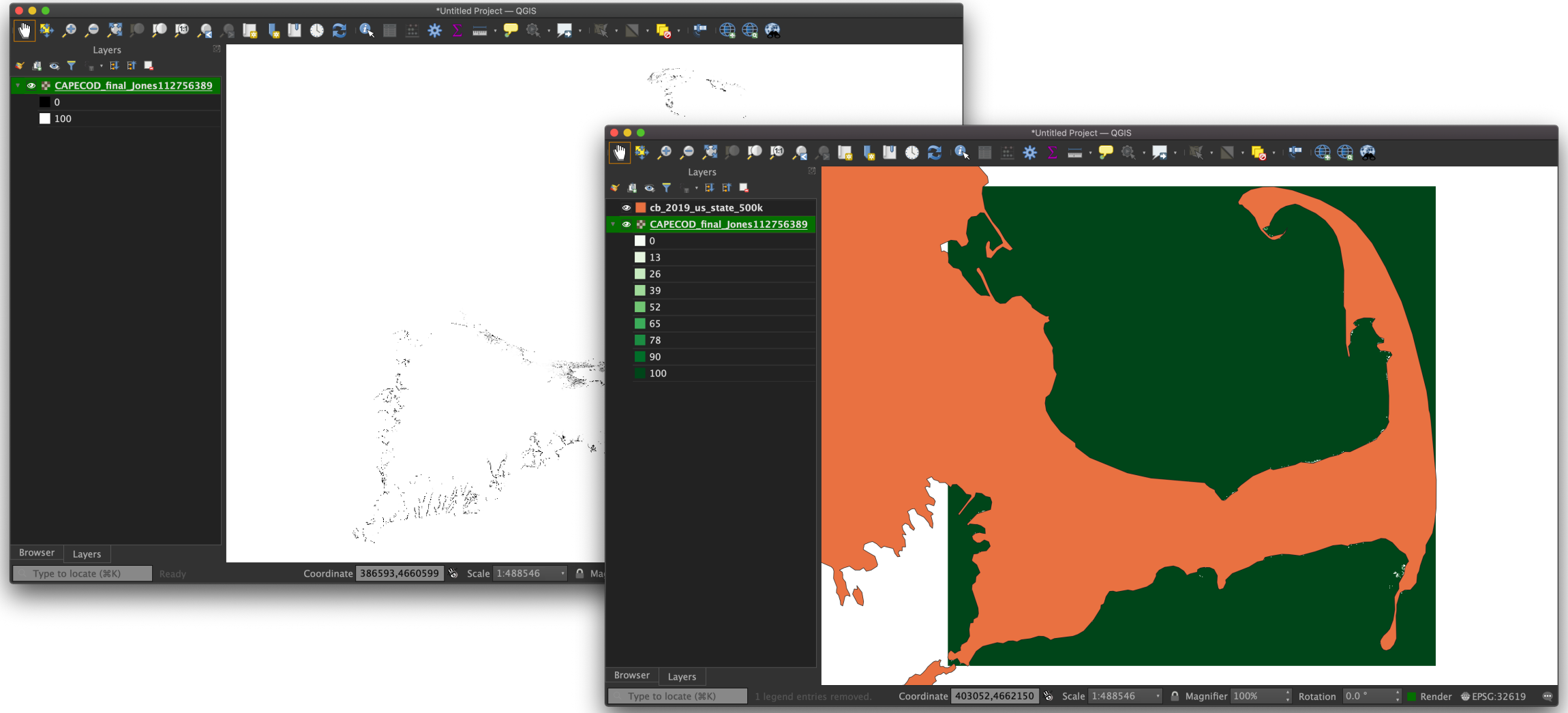
QA Exercise

soil_CO2_flux_final_OLD.csv

Soil Flux Compilation, v2.3								
Data collected from plots by JW and AB.								
Flux provided in units of daily average (g C m-2 day-1).								
loc	country	biome	latitude	longitude	soil temp	meas_yr	flux	modis_lai
Anchorage	USA	Boreal	61.2	-149.311	-6.9	2007	0.476	2.8
D1	Chile	Temperate	-45.4	-72	14.9	?	0.87	2
Eight Mile Lake	USA	Tundra	63.878	161.33	-0.82	2008	1.048	2.3
Itatinga	Brazil	Tropical	23.0333333	-48.633333	19.2	2012	0.379	3
Lompoloj_nkk_	Finland	Boreal	67.997	161.406	0	2008	0.255	2.1
Manitoba, BOREAS	Canada	Boreal	53.876	-149.76	0.86	1994	332	
McMurdo Station	Antarctica	Temperate	-77.8463	166.668	0	2009	0.036	
mean					1.836		0.382	2.1
North Slope	USA	Tundra	68.175	-149.253	0	prior to 2007	0.126	0.8
Northwest Territories	Canada	Boreal	61.31	-148.467	-1.1	2015	0.459	2.9
Svalbard	Norway	Tundra	78.224	-148.19	-8.3	2004	0	0.3
Wisconsin	USA	Temperate	46.17	-89.67	4.2	2015	0.476	2.7
Yakutsk	Russia	Boreal	62.255	24.209	0	2001	0.086	
Zotino, Central Siberia	Russia	Boreal	60.75	161.53	-0.41	1999	0.486	

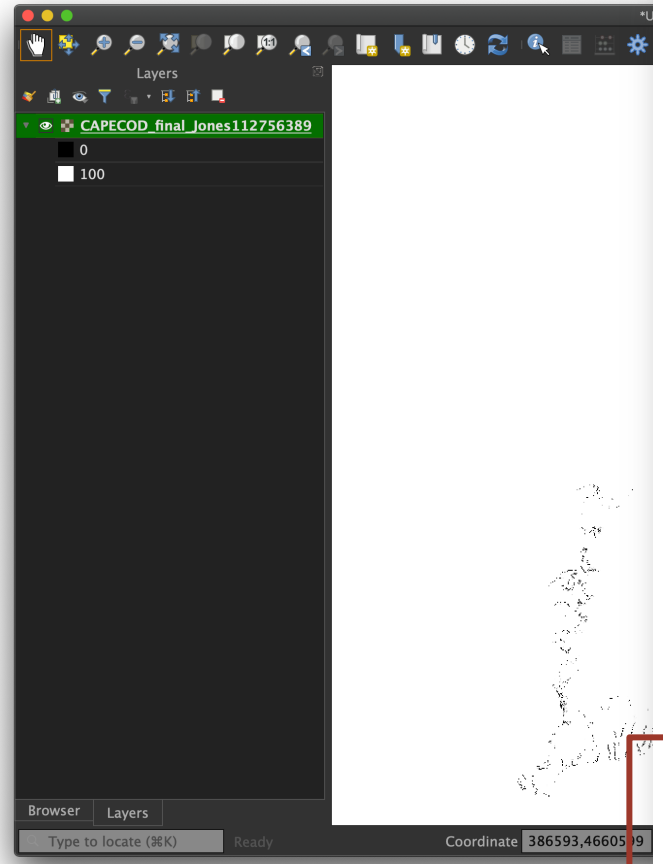
QA Exercise

CAPECOD_final_Jones112756389.tif



QA Exercise

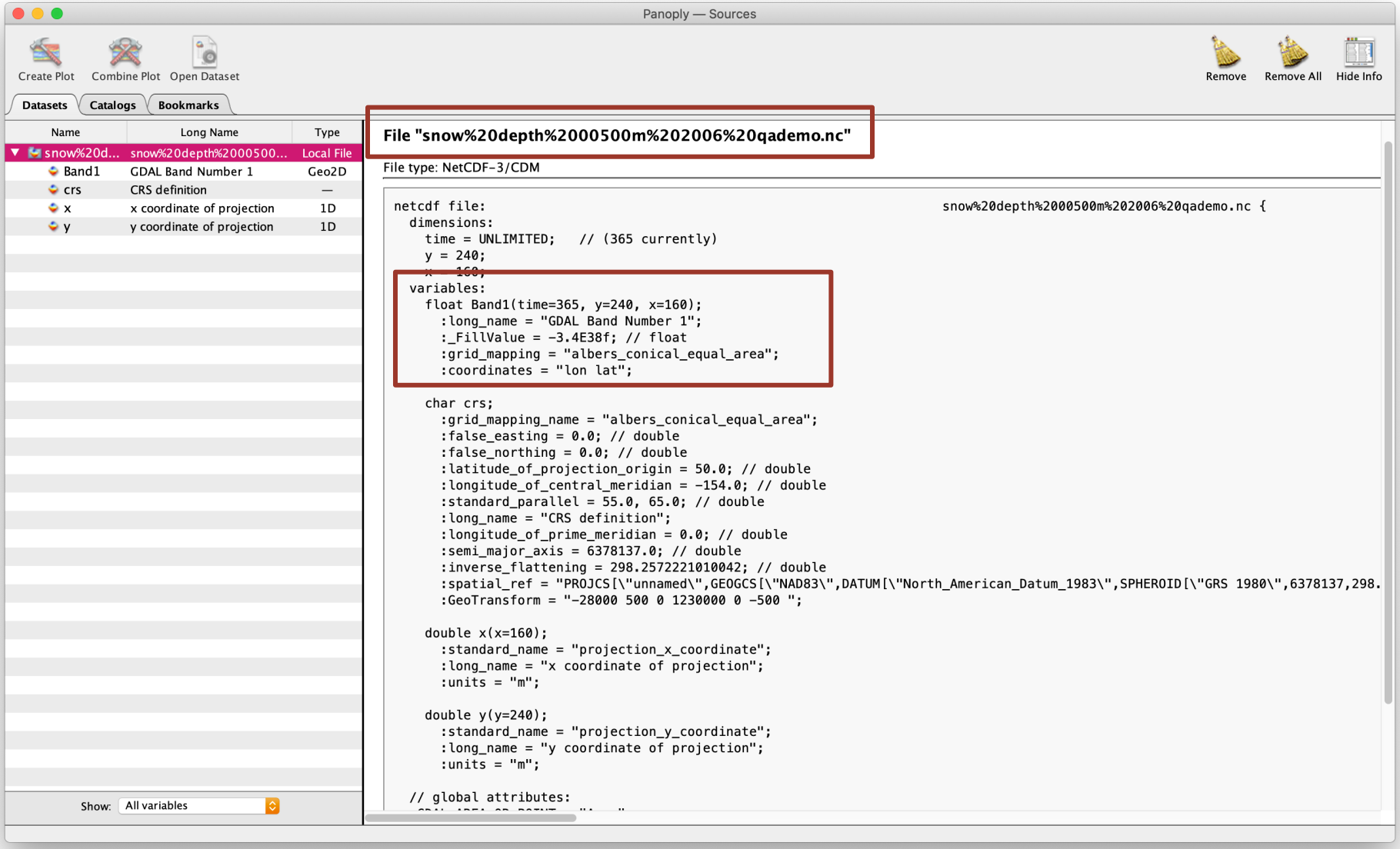
CAPECOD_final_Jones112756389.tif




```
Driver: GTiff/GeoTIFF
Files: CAPECOD_final_Jones112756389.tif
Size is 2116, 2076
Coordinate System is:
PROJCS["WGS_1984_UTM_Zone_19N",
    GEOGCS["WGS 84",
        DATUM["WGS_1984",
            SPHEROID["WGS 84",6378137,298.257223563,
                AUTHORITY["EPSG","7030"]],
            AUTHORITY["EPSG","6326"]],
        PRIMEM["Greenwich",0],
        UNIT["degree",0.0174532925199433],
        AUTHORITY["EPSG","4326"]],
    PROJECTION["Transverse_Mercator"],
    PARAMETER["latitude_of_origin",0],
    PARAMETER["central_meridian",-69],
    PARAMETER["scale_factor",0.9996],
    PARAMETER["false_easting",500000],
    PARAMETER["false_northing",0],
    UNIT["metre",1,
        AUTHORITY["EPSG","9001"]],
    AUTHORITY["EPSG","32619"]]
Origin = (359205.000000000000000,4659615.000000000000000)
Pixel Size = (30.000000000000000,-30.000000000000000)
Metadata:
  AREA_OR_POINT=Area
Image Structure Metadata:
  COMPRESSION=LZW
  INTERLEAVE=BAND
Corner Coordinates:
Upper Left  (  359205.000, 4659615.000) ( 70d42' 7.42"W, 42d 4'33.56"N)
Lower Left  (  359205.000, 4597335.000) ( 70d41'14.23"W, 41d30'55.01"N)
Upper Right (  422685.000, 4659615.000) ( 69d56' 5.26"W, 42d 5' 5.30"N)
Lower Right (  422685.000, 4597335.000) ( 69d55'36.04"W, 41d31'26.13"N)
Center      (  390945.000, 4628475.000) ( 70d18'45.74"W, 41d48' 2.51"N)
Band 1 Block=128x128 Type=Byte, ColorInterp=Gray
  Min=0.000 Max=100.000
  Minimum=0.000, Maximum=100.000, Mean=58.827, StdDev=39.726
  NoData Value=-128
  Metadata:
    STATISTICS_MAXIMUM=100
    STATISTICS_MEAN=58.827189421691
    STATISTICS_MINIMUM=0
    STATISTICS_SKIPFACTORX=1
    STATISTICS_SKIPFACTORY=1
    STATISTICS_STDDEV=39.72550564231
  Image Structure Metadata:
    PIXELTYPE=SIGNEDBYTE
```

QA Exercise

snow depth 00500m 2006 qademo.nc



CSV Format - Basic Guidelines

- A single header line with descriptive variable (column) names
- No spaces in variable names – use underscores
- No empty lines or rows, and the same number of columns across rows
- Consistent data types and precision within a column
- Prefer a no-data value based on data type: -9999 and NA
- Specify temporal and spatial information appropriately
  - Include UTC time in GMT following the ISO-8601 format
  - Include coordinates in decimal degrees using Geographic Lat/Lon (WGS84) CRS
- Provide auxiliary information (e.g., notes, descriptions) in documentation and/or companion files

CSV Format - Basic Guidelines

srdb-data-V5.csv

	A	B	C	D	F	H	I	J	K	L	M	N	O	T	
1	Record_number	Entry_date	Study_number	Author	Quality_flag	Country	Region	Site_name	Site_ID	Study_midyear	YearsOfData	Latitude	Longitude	Age_disturbance	Species
2	1	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-D1998	CA-BND-3YRWD	2001.5	1.5	56.63	-99.94	3	Picea mariana; P
3	2	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-D1995	CA-BND-6YRWD	2001.5	1.5	56.46	-99.97	6	Picea mariana; P
4	3	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-D1989	CA-BND-12YRWD	2001.5	1.5	55.91	-98.98	12	Picea mariana; P
5	4	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-D1981	CA-BND-20YRWD	2001.5	1.5	55.86	-98.48	20	Picea mariana; P
6	5	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-D1964	CA-BND-37YRWD	2001.5	1.5	55.92	-98.39	37	Picea mariana; P
7	6	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-D1930	CA-BND-71YRWD	2001.5	1.5	55.91	-98.52	71	Picea mariana
8	7	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-D1850; NOBS	CA-BND-151YRWD	2001.5	1.5	55.88	-98.48	151	Picea mariana
9	8	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-W1998	CA-BND-3YRPD	2001.5	1.5	56.63	-99.94	3	Picea mariana; P
10	9	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-W1995	CA-BND-6YRPD	2001.5	1.5	56.46	-99.98	6	Picea mariana; P
11	10	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-W1989	CA-BND-12YRPD	2001.5	1.5	55.9	-98.98	12	Picea mariana; P
12	11	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-W1981	CA-BND-20YRPD	2001.5	1.5	55.86	-98.48	20	Picea mariana; P
13	12	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-W1964	CA-BND-37YRPD	2001.5	1.5	55.92	-98.39	37	Picea mariana; P
14	13	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-W1930	CA-BND-71YRPD	2001.5	1.5	55.91	-98.51	71	Picea mariana
15	14	2008-09-26	2534	Bond-Lamberty	Q02	Canada	Manitoba	BOREAS NSA-W1850; NOBS	CA-BND-151YRPD	2001.5	1.5	55.88	-98.48	151	Picea mariana
16	15	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-D1998	CA-BND-3YRWD	1999.5	1.5	56.63	-99.94	3	Picea mariana; P
17	16	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-D1995	CA-BND-6YRWD	1999.5	1.5	56.46	-99.97	6	Picea mariana; P
18	17	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-D1989	CA-BND-12YRWD	1999.5	1.5	55.91	-98.98	12	Picea mariana; P
19	18	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-D1981	CA-BND-20YRWD	1999.5	1.5	55.86	-98.48	20	Picea mariana; P
20	19	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-D1964	CA-BND-37YRWD	1999.5	1.5	55.92	-98.39	37	Picea mariana; P
21	20	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-D1930	CA-BND-71YRWD	1999.5	1.5	55.91	-98.52	71	Picea mariana
22	21	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-D1850; NOBS	CA-BND-151YRWD	1999.5	1.5	55.88	-98.48	151	Picea mariana
23	22	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-W1998	CA-BND-3YRPD	1999.5	1.5	56.63	-99.94	3	Picea mariana; P
24	23	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-W1995	CA-BND-6YRPD	1999.5	1.5	56.46	-99.98	6	Picea mariana; P
25	24	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-W1989	CA-BND-12YRPD	1999.5	1.5	55.9	-98.98	12	Picea mariana; P
26	25	2008-09-26	2182	Wang	Q02	Canada	Manitoba	BOREAS NSA-W1981	CA-BND-20YRPD	1999.5	1.5	55.86	-98.48	20	Picea mariana; P

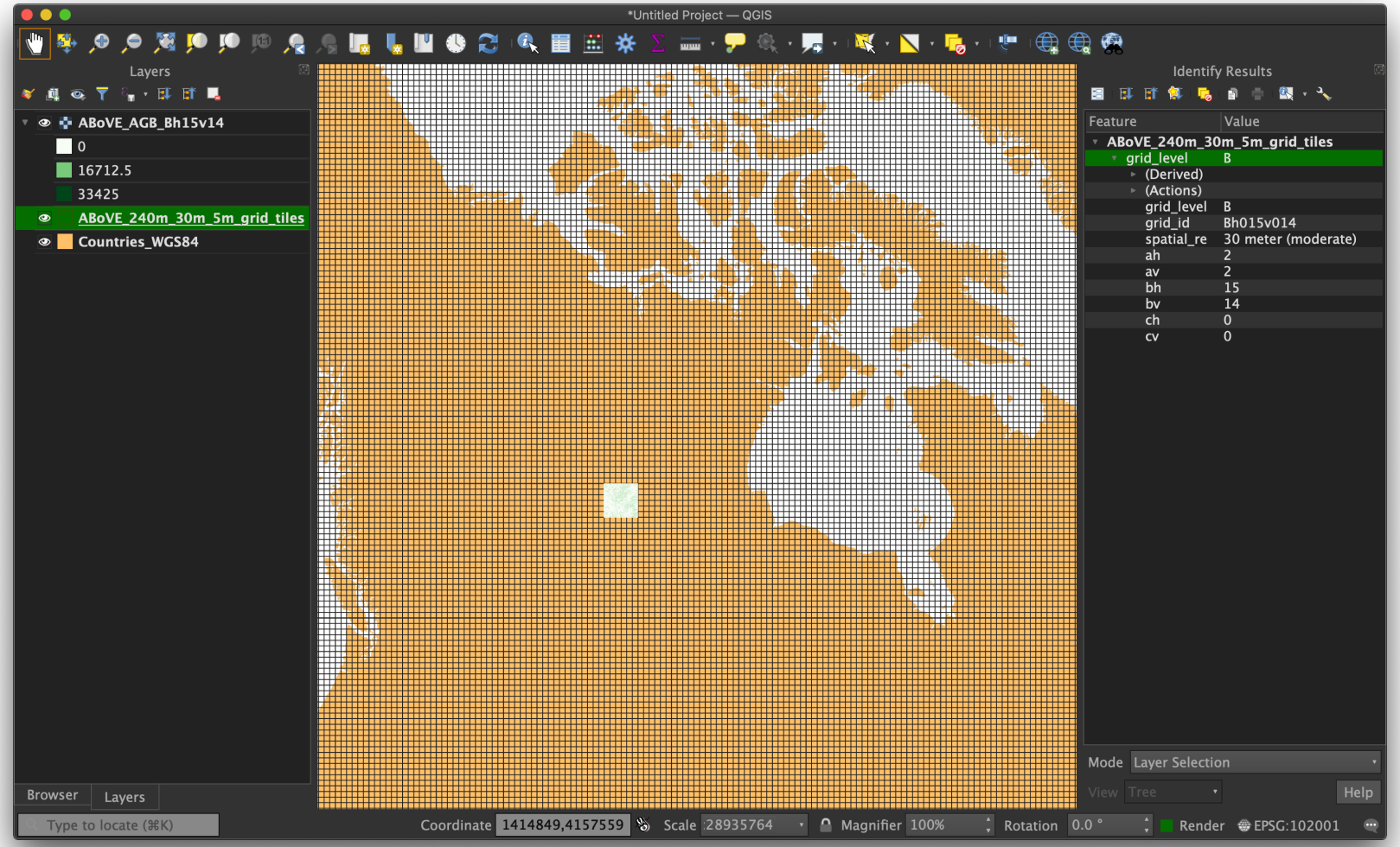
srdb-data-V5

A Global Database of Soil Respiration Data, Version 5.0.
https://doi.org/10.3334/ORNLDAAC/1827

GeoTIFF Format - Basic Guidelines

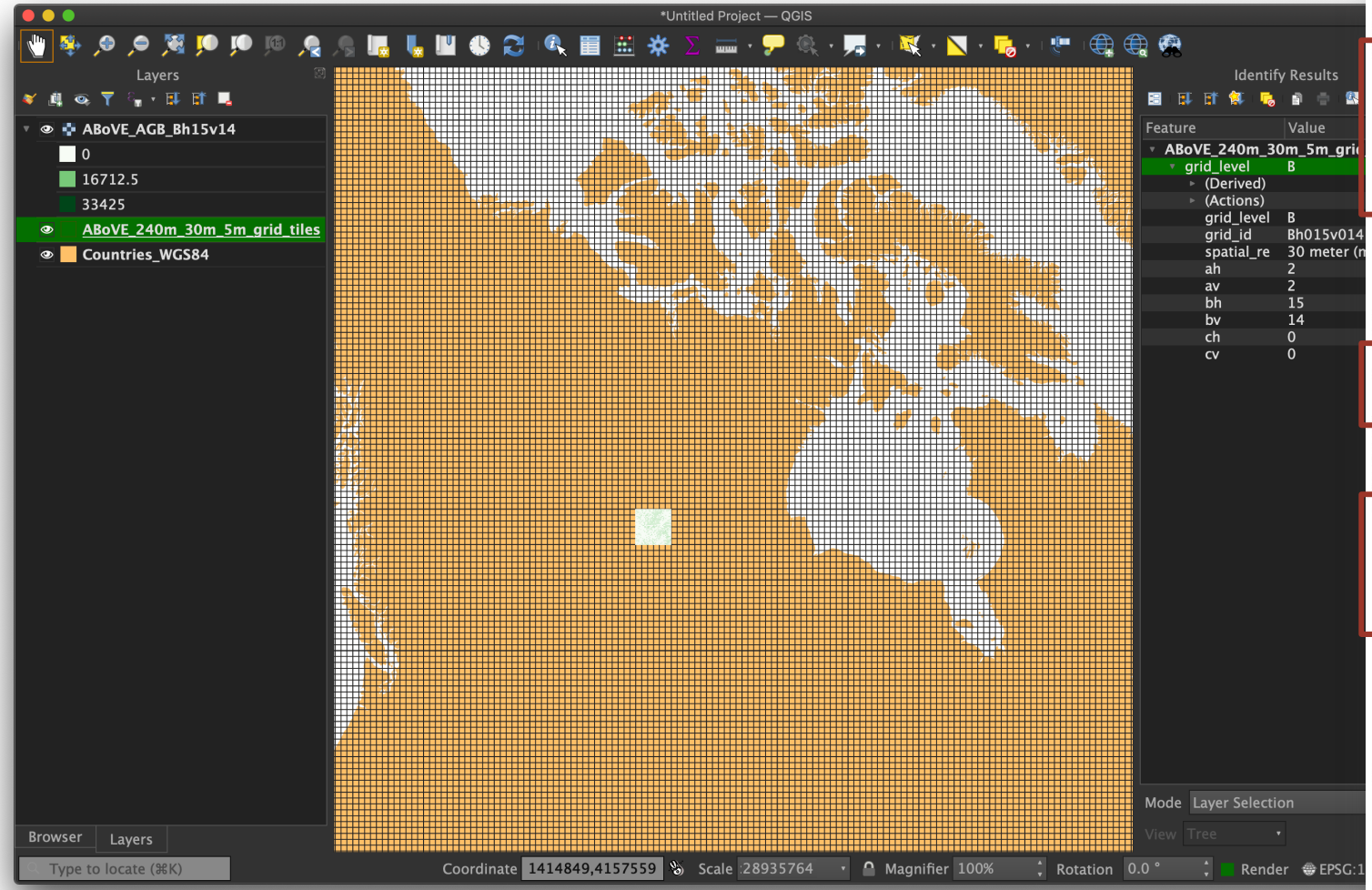
- Avoid mixing multiple variables into one file
- Use appropriate data type (e.g., byte v. integer v. float)
- Define spatial information
  - Spatial/Coordinate Reference System (S/CRS)
  - Spatial extent
- Set a proper no-data value
- Apply internal compression
  - Use only DEFLATE or LZW compression method
- Check basic statistics (e.g., min, max) of data values
- Use a base map to ensure coordinates are correct

GeoTIFF Format - Basic Guidelines



ABoVE: Annual Aboveground Biomass for Boreal Forests of ABoVE Core Domain, 1984-2014. https://doi.org/10.3334/ORNLDAAC/1808

GeoTIFF Format - Basic Guidelines




```
Driver: GTiff/GeoTIFF
Files: ABoVE_AGB_Bh15v14.tif
Size is 6000, 6000
Coordinate System is:
PROJCS["Canada_Albers_Equal_Area_Conic",
    GEOGCS["NAD83",
        DATUM["North_American_Datum_1983",
            SPHEROID["GRS 1980",6378137,298.2572221010042,
                AUTHORITY["EPSG","7019"]],
            AUTHORITY["EPSG","6269"]],
        PRIMEM["Greenwich",0],
        UNIT["degree",0.0174532925199433],
        AUTHORITY["EPSG","4269"]],
    PROJECTION["Albers_Conic_Equal_Area"],
    PARAMETER["standard_parallel_1",50],
    PARAMETER["standard_parallel_2",70],
    PARAMETER["latitude_of_center",40],
    PARAMETER["longitude_of_center",-96],
    PARAMETER["false_easting",0],
    PARAMETER["false_northing",0],
    UNIT["metre",1,
        AUTHORITY["EPSG","9001"]]]
Origin = (-700020.000000001629815,2120000.000000013969839)
Pixel Size = (30.000000000000025,-30.000000000000025)
Metadata:
  AREA_OR_POINT=Area
Image Structure Metadata:
  COMPRESSION=LZW
  INTERLEAVE=PIXEL
Corner Coordinates:
Upper Left  ( -700020.000, 2120000.000) (108d15'45.20"W, 58d33'18.89"N)
Lower Left  ( -700020.000, 1940000.000) (107d43' 5.97"W, 56d59' 8.66"N)
Upper Right ( -520020.000, 2120000.000) (105d 9'17.52"W, 58d48'29.64"N)
Lower Right ( -520020.000, 1940000.000) (104d44'41.22"W, 57d13'40.65"N)
Center      ( -610020.000, 2030000.000) (106d28'13.52"W, 57d54'13.61"N)
Band 1 Block=6000x1 Type=UInt16, ColorInterp=Gray
  Minimum=0.000, Maximum=33425.000, Mean=1776.310, StdDev=2036.444
  NoData Value=65535
  Metadata:
    STATISTICS_MAXIMUM=33425
    STATISTICS_MEAN=1776.3100794593
    STATISTICS_MINIMUM=0
    STATISTICS_STDDEV=2036.4438963114
    STATISTICS_VALID_PERCENT=99.999
Band 2 Block=6000x1 Type=UInt16, ColorInterp=Undefined
  Minimum=0.000, Maximum=29721.000, Mean=1790.964, StdDev=2050.796
  NoData Value=65535
  Metadata:
    STATISTICS_MAXIMUM=29721
    STATISTICS_MEAN=1790.9641968302
    STATISTICS_MINIMUM=0
    STATISTICS_STDDEV=2050.7960303926
    STATISTICS_VALID_PERCENT=99.999
Band ...
Band 31 Block=6000x1 Type=UInt16, ColorInterp=Undefined
  Minimum=0.000, Maximum=39249.000, Mean=1349.827, StdDev=1917.720
  NoData Value=65535
  Metadata:
    STATISTICS_MAXIMUM=39249
    STATISTICS_MEAN=1349.8268687803
    STATISTICS_MINIMUM=0
    STATISTICS_STDDEV=1917.7195001463
    STATISTICS_VALID_PERCENT=99.999
```

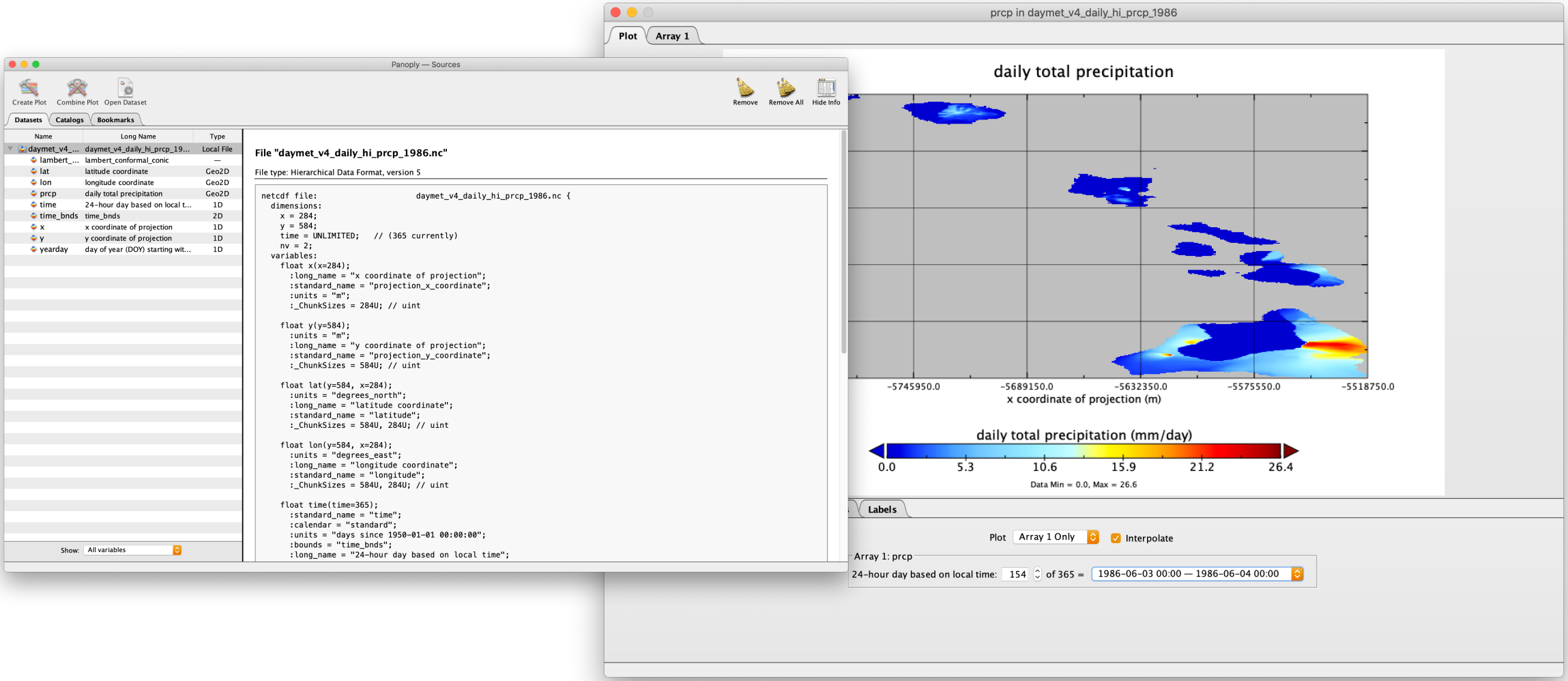
ABoVE: Annual Aboveground Biomass for Boreal Forests of ABoVE Core Domain, 1984-2014. https://doi.org/10.3334/ORNLDAAC/1808

NetCDF Format - Basic Guidelines

- Follow metadata standards available at cfconventions.org
- Variables made self-descriptive using attributes
- Provide the *standard_name* attribute when possible
- Specify *calendar* and *units* attributes for time variable
- Include *time_bnds* variable for timeseries data
- Provide cell center coordinates in WGS84 CRS
- Values of longitude variable go from west (-180) to east (180)
- Values of latitude variable go from south (-90) to north (90)

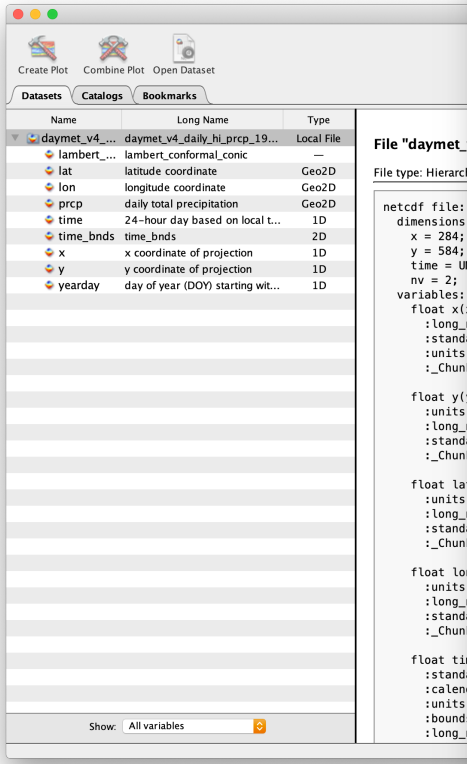
1. NetCD-what? An Ecologist's Guide to Working with Daymet and other NetCDF-formatted Data
2. NetCDF Why and How: Creating Publication Quality NetCDF Datasets

NetCDF Format - Basic Guidelines



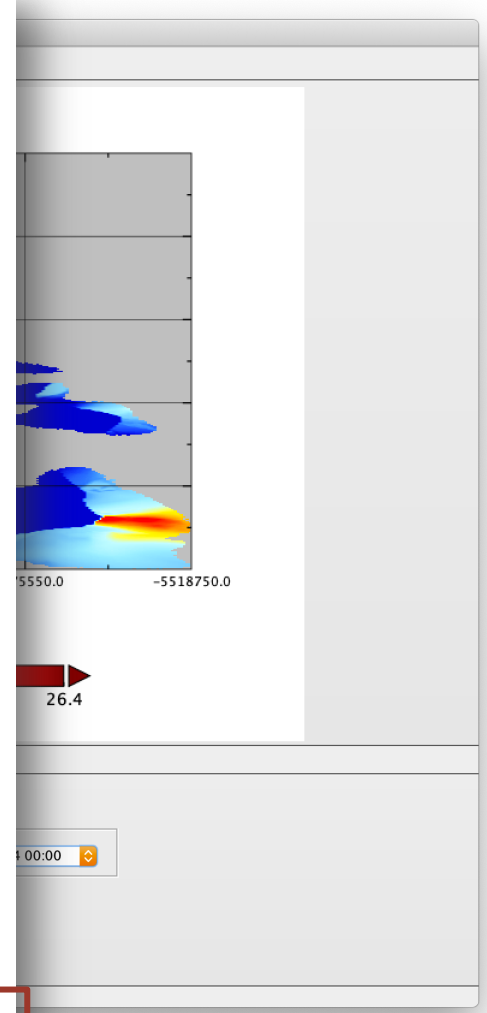
Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4. https://doi.org/10.3334/ORNLDAAC/1840

NetCDF Format - Basic Guidelines




```
netcdf daymet_v4_daily_hi_prcp_1986 {
dimensions:
        x = 284 ;
        y = 584 ;
        time = UNLIMITED ; // (365 currently)
        nv = 2 ;
variables:
        float x(x) ;
                x:units = "m" ;
                x:long_name = "x coordinate of projection" ;
                x:standard_name = "projection_x_coordinate" ;
        float y(y) ;
                y:units = "m" ;
                y:long_name = "y coordinate of projection" ;
                y:standard_name = "projection_y_coordinate" ;
        float lat(y, x) ;
                lat:units = "degrees_north" ;
                lat:long_name = "latitude coordinate" ;
                lat:standard_name = "latitude" ;
        float lon(y, x) ;
                lon:units = "degrees_east" ;
                lon:long_name = "longitude coordinate" ;
                lon:standard_name = "longitude" ;
        float time(time) ;
                time:standard_name = "time" ;
                time:calendar = "standard" ;
                time:units = "days since 1950-01-01 00:00:00" ;
                time:bounds = "time_bnds" ;
                time:long_name = "24-hour day based on local time" ;
        short yearday(time) ;
                yearday:long_name = "day of year (DOY) starting with day 1 on Januaray 1st" ;
        float time_bnds(time, nv) ;
        short lambert_conformal_conic ;
                lambert_conformal_conic:grid_mapping_name = "lambert_conformal_conic" ;
                lambert_conformal_conic:longitude_of_central_meridian = -100. ;
                lambert_conformal_conic:latitude_of_projection_origin = 42.5 ;
                lambert_conformal_conic:false_easting = 0. ;
                lambert_conformal_conic:false_northing = 0. ;
                lambert_conformal_conic:standard_parallel = 25., 60. ;
                lambert_conformal_conic:semi_major_axis = 6378137. ;
                lambert_conformal_conic:inverse_flattening = 298.257223563 ;
        float prcp(time, y, x) ;
                prcp:_FillValue = -9999.f ;
                prcp:long_name = "daily total precipitation" ;
                prcp:units = "mm/day" ;
                prcp:missing_value = -9999.f ;
                prcp:coordinates = "lat lon" ;
                prcp:grid_mapping = "lambert_conformal_conic" ;
                prcp:cell_methods = "area: mean time: sum" ;

// global attributes:
                :start_year = 1986s ;
                :source = "Daymet Software Version 4.0" ;
                :Version_software = "Daymet Software Version 4.0" ;
                :Version_data = "Daymet Data Version 4.0" ;
                :Conventions = "CF-1.6" ;
                :citation = "Please see http://daymet.ornl.gov/ for current Daymet data citation information" ;
                :references = "Please see http://daymet.ornl.gov/ for current information on Daymet references" ;
}
```



id for North America, Version 4.

Free & Open-Source Tools

- GDAL is a translator library for raster and vector geospatial data formats.

  gdal.org/
- NCO is a toolkit that manipulates and analyzes data stored in netCDF-accessible formats, including DAP, HDF4, and HDF5.

  nco.sourceforge.net/
- Panoply plots geo-referenced and other arrays from netCDF, HDF, GRIB, and other datasets.

  giss.nasa.gov/tools/panoply/
- QGIS is a Geographic Information System to create, edit, visualize, analyze, and publish geospatial information.

  qgis.org/

ORNL DAAC
DISTRIBUTED ACTIVE ARCHIVE CENTER
FOR BIOGEOCHEMICAL DYNAMICS

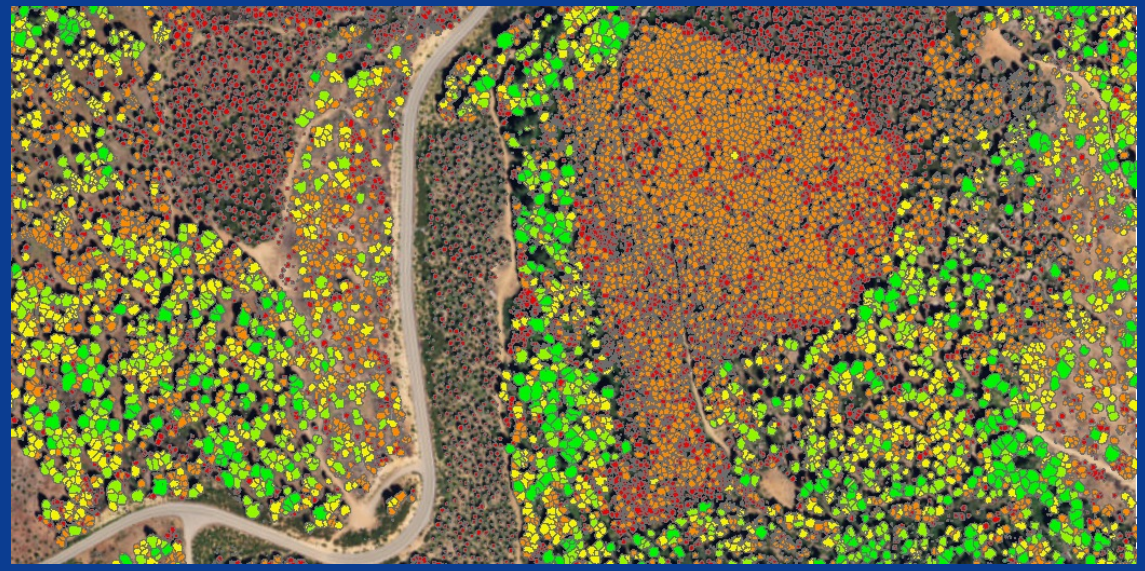
EOSDIS

Submitting Data to the ORNL DAAC

Yaxing Wei

NACP 7th OSM Data Management Workshop

March 24th, 2021



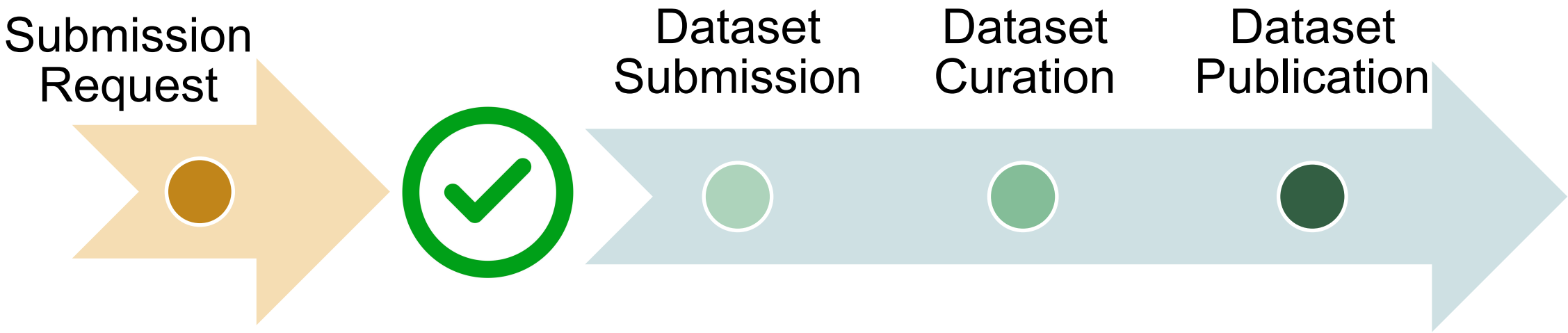
10.3334/ORNLDAAC/1537

The Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics operates under an interagency agreement between NASA and the U.S. Department of Energy

U.S. DEPARTMENT OF ENERGY

NASA

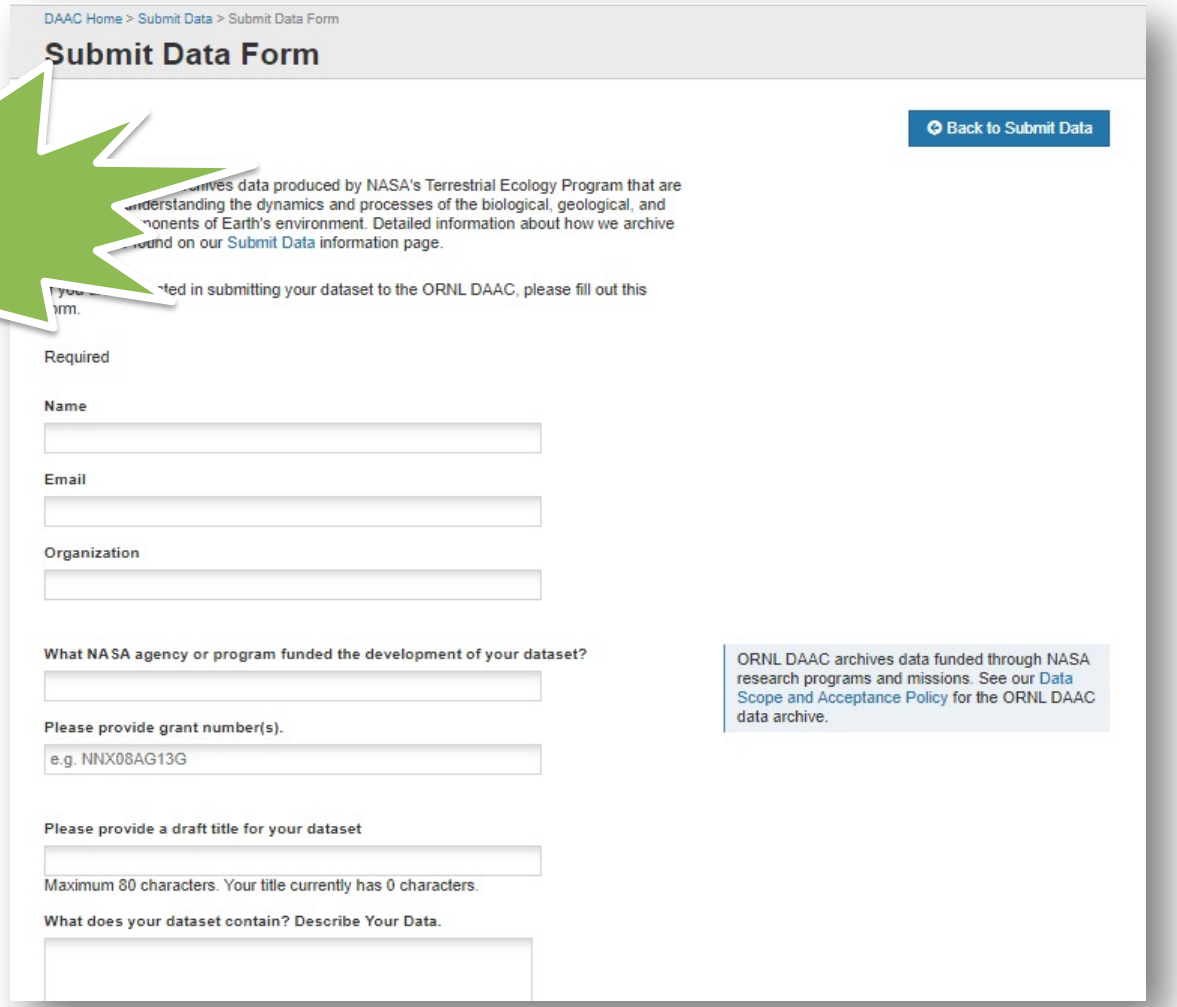
Process Overview



Submission Request

daac.ornl.gov/submit/

Do it early!



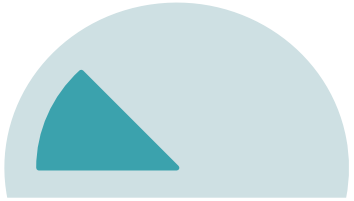
Data Scope and Acceptance Policy

- The ORNL DAAC archives data and model products that were generated with funding from the NASA Terrestrial Ecology program and other programs within the NASA Carbon Cycle and Ecosystems Focus area.
- It's unusual for data generated through other funding sources to be published at the ORNL DAAC. Such data submission requests require review and approval from NASA.
- Find out more at https://daac.ornl.gov/submit/

Dataset Submission

Once approved, you will receive an email describing the submission steps. Your submission will be assigned a unique identifier and upload area.

Create XCAMS Account

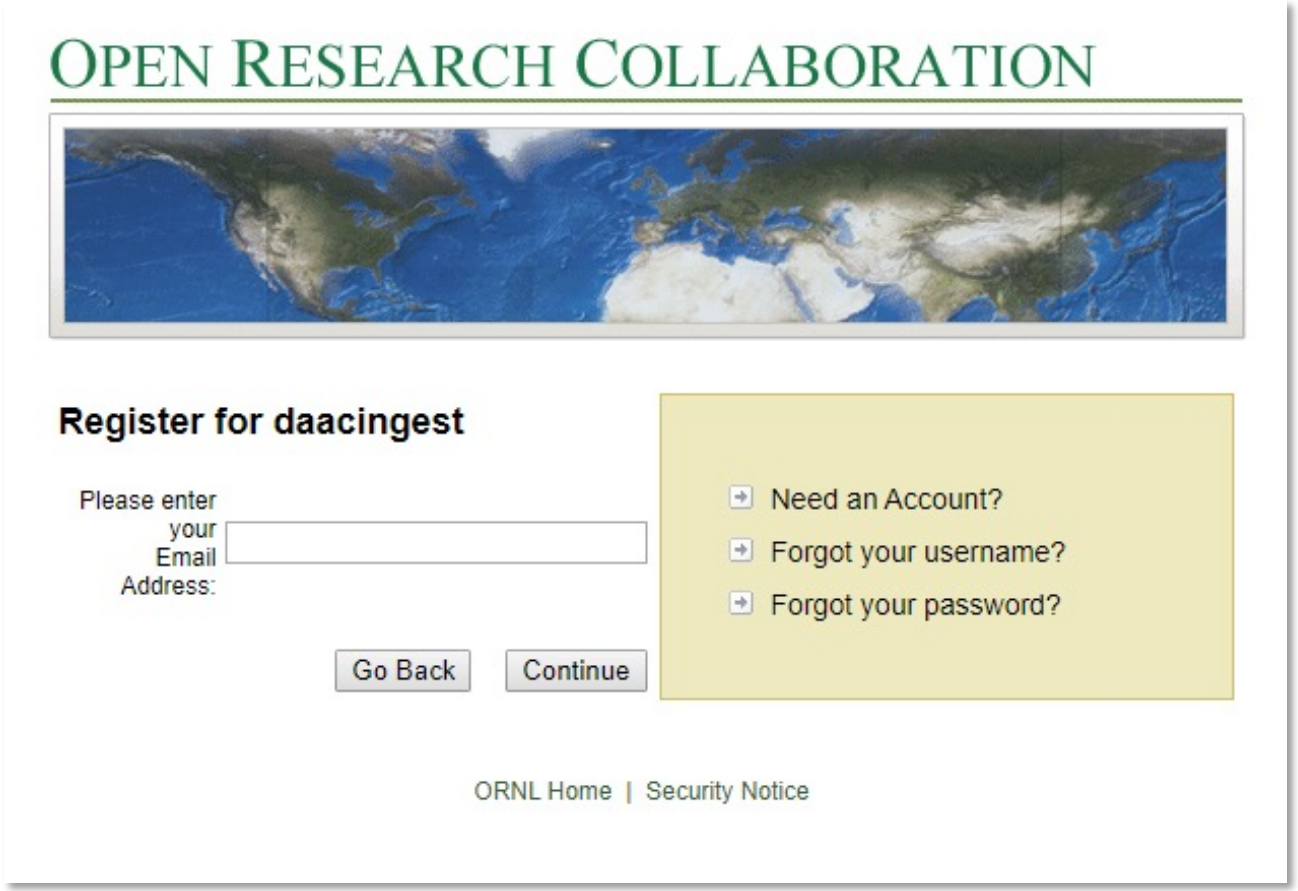


Upload the Files

Answer Data Provider Questions

Create XCAMS Account

The email contains a link to create an ORNL XCAMS account.



Upload the Files

The email provides you a link and guidance on how to upload files.

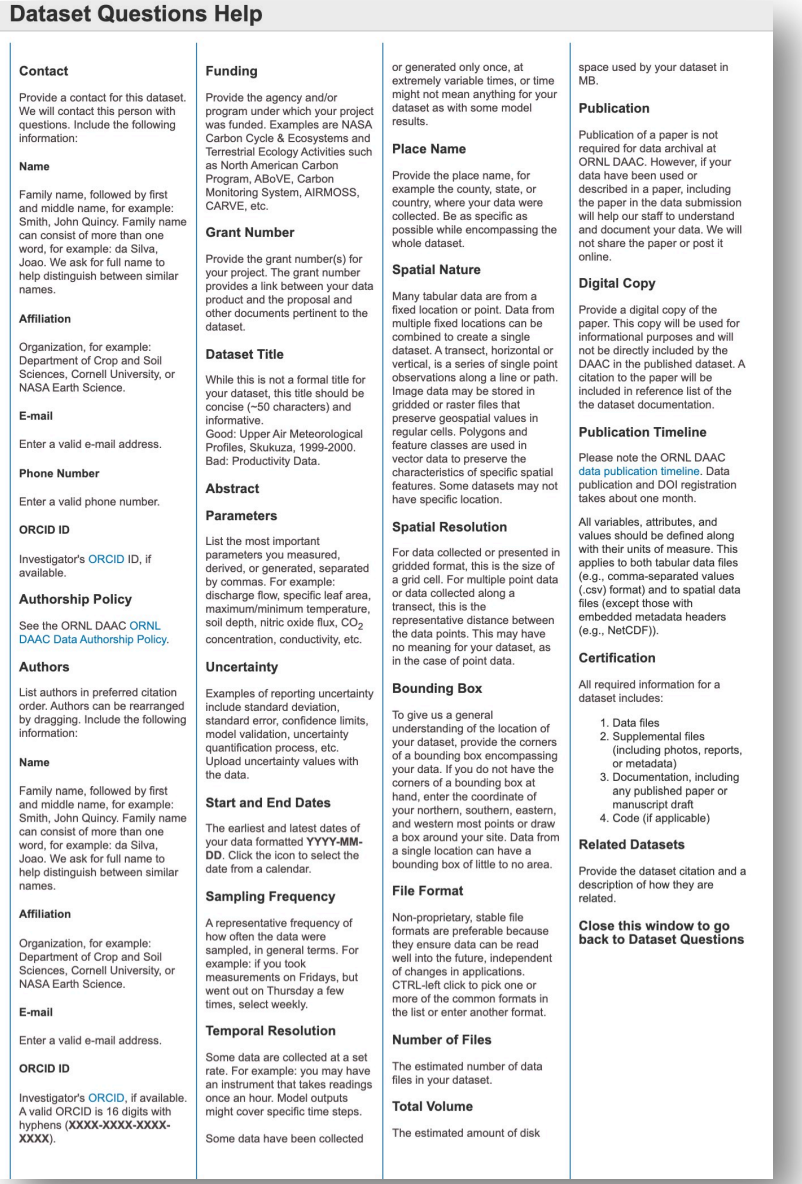
- Primary data files
  - High-level, derived products
  - Field and input data
  - Uncertainty estimates
- Dataset documentation
- Published or draft manuscript
- Supplemental files
  - Photos, reports, metadata, etc.
  - Associated code

Answer Data Provider Questions

daacdpq.ornl.gov/questions/help/

The email provides you a link to describe your dataset.

- Data contacts and authors
- Dataset description
  - Title, abstract, parameters, uncertainty
- Temporal and spatial characteristics
- Data formats and documentation
- Related archived datasets



Dataset Questions Help

Contact

Provide a contact for this dataset. We will contact this person with questions. Include the following information:

Name

Family name, followed by first and middle name, for example: Smith, John Quincy. Family name can consist of more than one word, for example: da Silva, Joao. We ask for full name to help distinguish between similar names.

Affiliation

Organization, for example: Department of Crop and Soil Sciences, Cornell University, or NASA Earth Science.

E-mail

Enter a valid e-mail address.

Phone Number

Enter a valid phone number.

ORCID ID

Investigator's ORCID ID, if available.

Authorship Policy

See the ORNL DAAC ORNL DAAC Data Authorship Policy.

Authors

List authors in preferred citation order. Authors can be rearranged by dragging. Include the following information:

Name

Family name, followed by first and middle name, for example: Smith, John Quincy. Family name can consist of more than one word, for example: da Silva, Joao. We ask for full name to help distinguish between similar names.

Affiliation

Organization, for example: Department of Crop and Soil Sciences, Cornell University, or NASA Earth Science.

E-mail

Enter a valid e-mail address.

ORCID ID

Investigator's ORCID, if available. A valid ORCID is 16 digits with hyphens (**XXXX-XXXX-XXXX-XXXX**).

Funding

Provide the agency and/or program under which your project was funded. Examples are NASA Carbon Cycle & Ecosystems and Terrestrial Ecology Activities such as North American Carbon Program, ABoVE, Carbon Monitoring System, AIRMOSS, CARVE, etc.

Grant Number

Provide the grant number(s) for your project. The grant number provides a link between your data product and the proposal and other documents pertinent to the dataset.

Dataset Title

While this is not a formal title for your dataset, this title should be concise (~50 characters) and informative.
Good: Upper Air Meteorological Profiles, Skukuza, 1999-2000.
Bad: Productivity Data.

Abstract

Parameters

List the most important parameters you measured, derived, or generated, separated by commas. For example: discharge flow, specific leaf area, maximum/minimum temperature, soil depth, nitric oxide flux, CO_2 concentration, conductivity, etc.

Uncertainty

Examples of reporting uncertainty include standard deviation, standard error, confidence limits, model validation, uncertainty quantification process, etc. Upload uncertainty values with the data.

Start and End Dates

The earliest and latest dates of your data formatted **YYYY-MM-DD**. Click the icon to select the date from a calendar.

Sampling Frequency

A representative frequency of how often the data were sampled, in general terms. For example: if you took measurements on Fridays, but went out on Thursday a few times, select weekly.

Temporal Resolution

Some data are collected at a set rate. For example: you may have an instrument that takes readings once an hour. Model outputs might cover specific time steps.

Some data have been collected or generated only once, at extremely variable times, or time might not mean anything for your dataset as with some model results.

Place Name

Provide the place name, for example the county, state, or country, where your data were collected. Be as specific as possible while encompassing the whole dataset.

Spatial Nature

Many tabular data are from a fixed location or point. Data from multiple fixed locations can be combined to create a single dataset. A transect, horizontal or vertical, is a series of single point observations along a line or path. Image data may be stored in gridded or raster files that preserve geospatial values in regular cells. Polygons and feature classes are used in vector data to preserve the characteristics of specific spatial features. Some datasets may not have specific location.

Spatial Resolution

For data collected or presented in gridded format, this is the size of a grid cell. For multiple point data or data collected along a transect, this is the representative distance between the data points. This may have no meaning for your dataset, as in the case of point data.

Bounding Box

To give us a general understanding of the location of your dataset, provide the corners of a bounding box encompassing your data. If you do not have the corners of a bounding box at hand, enter the coordinate of your northern, southern, eastern, and western most points or draw a box around your site. Data from a single location can have a bounding box of little to no area.

File Format

Non-proprietary, stable file formats are preferable because they ensure data can be read well into the future, independent of changes in applications. CTRL-left click to pick one or more of the common formats in the list or enter another format.

Number of Files

The estimated number of data files in your dataset.

Total Volume

The estimated amount of disk space used by your dataset in MB.

Publication

Publication of a paper is not required for data archival at ORNL DAAC. However, if your data have been used or described in a paper, including the paper in the data submission will help our staff to understand and document your data. We will not share the paper or post it online.

Digital Copy

Provide a digital copy of the paper. This copy will be used for informational purposes and will not be directly included by the DAAC in the published dataset. A citation to the paper will be included in reference list of the the dataset documentation.

Publication Timeline

Please note the ORNL DAAC data publication timeline. Data publication and DOI registration takes about one month.

All variables, attributes, and values should be defined along with their units of measure. This applies to both tabular data files (e.g., comma-separated values (.csv) format) and to spatial data files (except those with embedded metadata headers (e.g., NetCDF)).

Certification

All required information for a dataset includes:

1. Data files
2. Supplemental files (including photos, reports, or metadata)
3. Documentation, including any published paper or manuscript draft
4. Code (if applicable)

Related Datasets

Provide the dataset citation and a description of how they are related.

Close this window to go back to Dataset Questions

What is Data Documentation?

daac.ornl.gov/submit/#guidelines/

- Conveys information for an average user to know how to use your data.
  - For example, a "README" file for users who are unfamiliar with the project, methods, data, etc.
- We offer a documentation template to help you get started.

```
--------------------------------------------------------------------------
DATASET OVERVIEW

Dataset Working Title:

Investigators: (full name as formatted for manuscripts, email address, ORCID ID)

Point of Contact: (name, email, institution)

Summary: (overview describing the variables in the dataset and how they were derived)

Keywords:

Acknowledgments: (include funding sources)

Related Publication: (include DOIs)

Related Datasets: (include DOIs)
--------------------------------------------------------------------------
DATASET CHARACTERISTICS

Spatial Resolution: (e.g., 1000 m, 2 degrees, point locations)

Projection Information:

Temporal Resolution: (e.g., daily, annual, one-time estimate)

Temporal Coverage: (e.g., YYYY-MM-DD to YYYY-MM-DD)

File Format and Number:

File Naming Convention:

File Descriptions: (file names, units, and descriptions)

Data Dictionary: (variables, units, and descriptions that define jargon, acronyms, abbreviations, etc.)
--------------------------------------------------------------------------
APPLICATION & DERIVATION

(What is the significance and application of the data?)
--------------------------------------------------------------------------
QUALITY ASSESSMENT
```

Track Your Dataset's Progress

The email provides you a link to track progress.

ORNL DAAC Dataset Publication Progress Tracker

Data Set ID - 38786036ad

Status

Submission Opened	Data Upload	Questions	Submission Closed	QA	Documentation	Publication
Completed 2021-02-24 1 days	Completed 2021-02-26 2 days	Completed 2021-03-02 6 days	Completed 2021-03-02 6 days	In Progress	Open	Open

The Submit Data Page

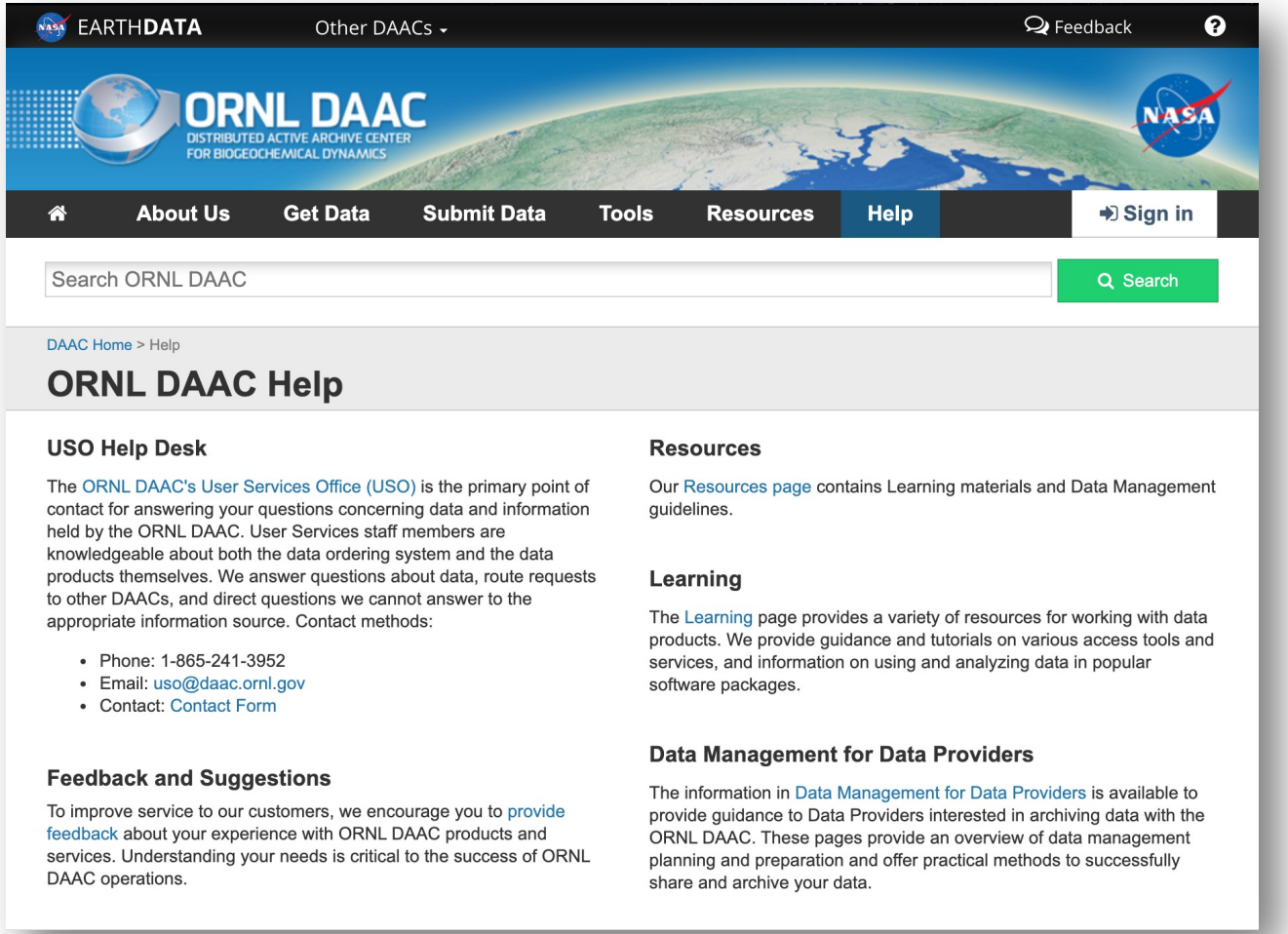
daac.ornl.gov/submit/

Workshop Summary

daac.ornl.gov/help/

You should now understand

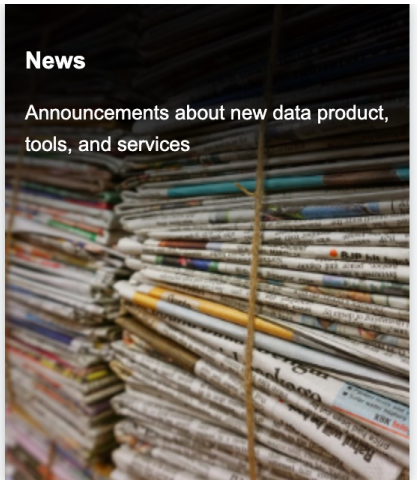
- Benefits of Publishing your Data
- Best Practices of Data Management
- Dataset Quality Assurance
- Submitting Data to the ORNL DAAC



Get more from the ORNL DAAC

daac.ornl.gov/resources/

- Learning – webinars, tutorials, etc.
  - NetCDF Why and How: Creating Publication Quality NetCDF Datasets
  - Introduction to Geospatial Analysis in R
  - Understanding AVIRIS-NG data in ENVI format with rotated grid
  - Accessing Data through ORNL DAAC Web Services
  - MODIS and VIIRS Data Tools and Services at your Fingertips
- Data Management
- News



Go FAIR! Produce Open Science!

- Findable, Accessible, Interoperable, Reusable (F.A.I.R.)
  doi.org/10.1038/sdata.2016.18
  go-fair.org

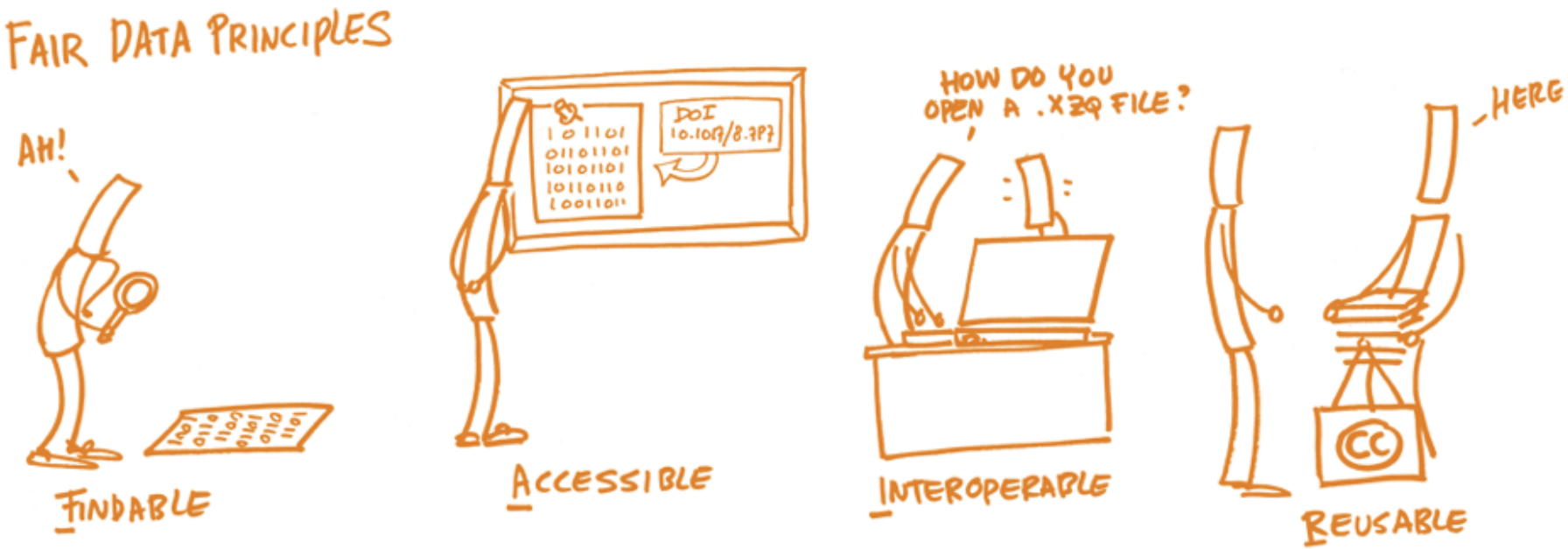
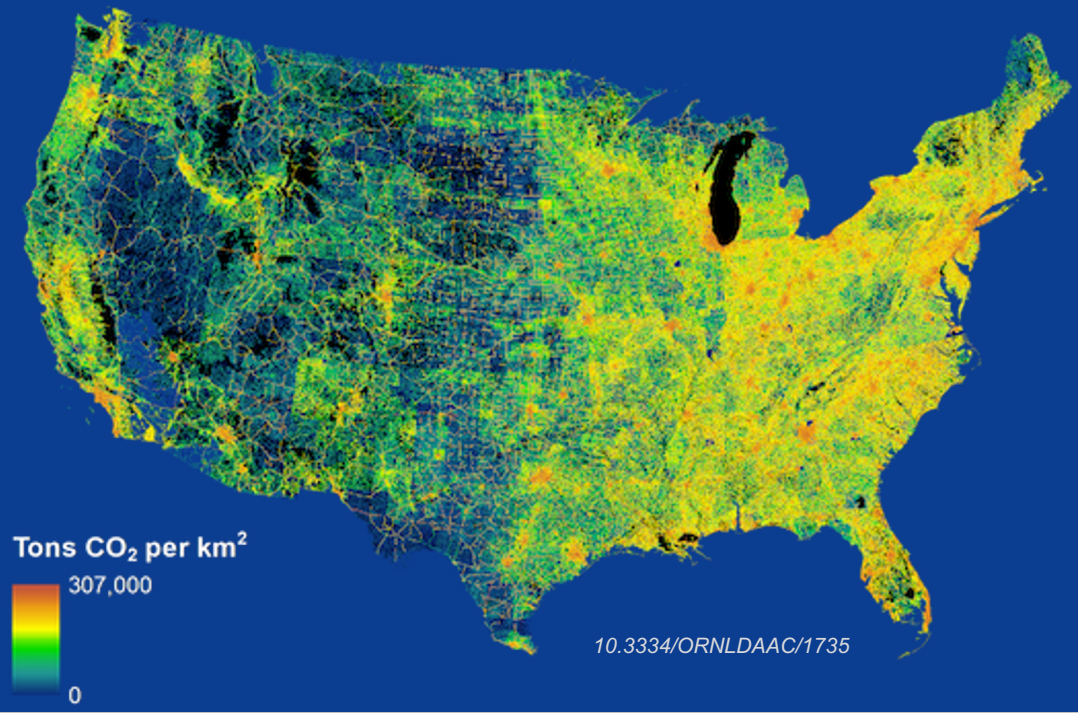


Image Source: book.fosteropenscience.eu/

ORNL DAAC
DISTRIBUTED ACTIVE ARCHIVE CENTER FOR BIOGEOCHEMICAL DYNAMICS

EOSDIS

Question & Answer Session



Kay Shope, Tammy Walker

Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)

March 24th, 2021

The Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics operates under an interagency agreement between NASA and the U.S. Department of Energy

U.S. DEPARTMENT OF ENERGY

NASA