



# Best Data Management Practices



Robert Cook, Yaxing Wei, Les Hook, Alison Boyer, Suresh Santhana Vannan, Viv Hutchison\*, Michele Thornton, and Tammy Beaty

Climate Change Science Institute and Environmental Sciences Division,

Oak Ridge National Laboratory

\*US Geological Survey, Denver



## Introduction

Scientists spend considerable time conducting field studies and experiments, running computer models, and analyzing the data compiled, but an often-overlooked activity is effectively managing the data for their own use and also for sharing and archiving.

Here we provide guidance on fundamental data management practices that investigators should perform during the course of data collection to improve the usability of their data sets.

## Benefits of good data management practices

### Short-term

- Spend less time doing data management and more time doing research
- Easier to prepare and use data for yourself and collaborators

### Long-term

- Scientists outside your project can find, understand, and use your data to address broad questions
- You get credit for archived data products and sponsors see good return on their investment

### Open data for the US Government



- US Agencies have data sharing policies and requirements
- Scientific journals (Nature, Science, and PLoS) have data sharing requirements

## An effective Data Management Plan for a proposal

A Data Management Plan describes what you will do with your data during and after your research. An effective Plan contains:

1. Information about the data
  - Description of data to be produced
  - How will it be managed in short-term?
2. Description of Data
  - Structure, number of files, volume, format
3. Metadata Content & Format
  - Characteristics of the data
4. Policies for Access, Sharing, & Reuse
5. Long-term Storage & Data Management
  - Where will the data be archived?



Detailed Template: <http://daac.ornl.gov/PI/plan.shtml>

## Metadata

Information to find, understand, and use the data

### Why

Why were the data collected?

### Who

Who collected the data?  
Who to contact for questions?  
Who to contact to order?  
Who owns the data?

### Where

Where were the data collected?  
Where were the data processed?  
Where are the data located?

### What

What are the data about?  
What project were they collected under?  
What are the constraints on their use?  
What is the quality?  
What parameters were measured?  
What instruments were used?  
What format are the data in?

### When

When were the data collected?

### How

How were the data collected?  
How do I access the data?  
How do I order the data?  
How much do the data cost?  
How was the quality assessed?

## Manage your data

### Define the contents of your data files

For others to use your data, they must fully understand:

- Parameter names, units, and coded values.
- Dates and time
- Spatial coordinates and elevation
- Formats



Unit database and conversion between units  
UDUNITS



CF Standard Name  
Climate Forecast (CF) standards promote sharing

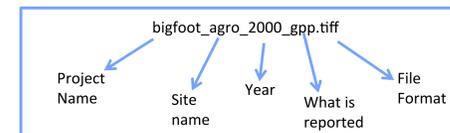
### Assign descriptive data set titles

Data set titles should be as descriptive as possible

- *NACP Site: Terrestrial Biosphere Model and Aggregated Flux Data in Standard Format*
- *CMS: Forest Biomass and Productivity, 1-degree and 5-km, Conterminous US, 2005*

### Assign descriptive file names

File names should reflect the contents of the file and uniquely identify the data file



### Use consistent data organization

Use consistent data organization, parameter formats, and common site names across the data set.

| Station | Date     | Temp | Precip |
|---------|----------|------|--------|
| Units   | YYYYMMDD | C    | mm     |
| HOGI    | 19961001 | 12   | 0      |
| HOGI    | 19961002 | 14   | 3      |
| HOGI    | 19961003 | 19   | -9999  |

Spreadsheet: Columns contain all the parameters that make up the record.

| Station | Date     | Parameter | Value | Unit |
|---------|----------|-----------|-------|------|
| HOGI    | 19961001 | Temp      | 12    | C    |
| HOGI    | 19961002 | Temp      | 14    | C    |
| HOGI    | 19961001 | Precip    | 0     | mm   |
| HOGI    | 19961002 | Precip    | 3     | mm   |

Database: Each row represents a record. This generates a long and skinny file.

### Use stable file formats

- Select a stable, well-documented, and non-proprietary format that can be read well into the future
  - Tabular: .csv, .txt
  - Spatial: geotiff, netCDF, shapefile

| COUNTRY | LAT          | LONG         | DATE       | DISTURBANCE | TAXONOMY        | PLANT PAR |
|---------|--------------|--------------|------------|-------------|-----------------|-----------|
| none    | decimal degr | decimal degr | year-month | none        | none            | none      |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Baphia massi L  |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Baphinia pet L  |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Rubiaceae L     |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Brachystegia L  |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Brachystegia L  |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Burkea africe L |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Fabaceae L      |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Constratum L    |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Copaifera ba L  |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Diospyrus ba L  |           |
| Zambia  | -15.44       | 23.52        | 2000-02    | CO          | Hannoa chio L   |           |

### Protect your data

- Ensure error-free file transfers by comparing checksums before and after transfers
- Maintain back-up copies of data and ensure that you can read the backup

## Best practices for geospatial data

### Critical Information for Geospatial Data

- What: Data structure (e.g., raster or vector) and resolution / scale
- Where: Geospatial information
  - Spatial Reference System (SRS): datum and projection
  - Extent and location

Example 1: FLUXNET Sites  
Use European Petroleum Survey Group (EPSG) codes to represent common Spatial Reference Systems

| STED   | STNAME   | STATUS | TOWER_BEGIN (YYY | TOWER_END (YYY | LAT (dec. deg) | LONG (dec. deg) |
|--------|--|--------|------------------|----------------|----------------|-----------------|
| PA-Bar | Panama - Barro Colorado Island                   | Active | 2012             |                | 9.1540         | -79.8480        |
| US-Akn | USA - SC - Aiken                                 | Active | 2011             |                | 33.3833        | -81.5656        |
| US-ORv | USA - OH - Oientangy River Wetland Research Park | Active | 2011             |                | 40.0201        | -83.0183        |
| US-Dia | USA - CA - Diablo                                | Active | 2010             |                | 37.6773        | -121.5296       |

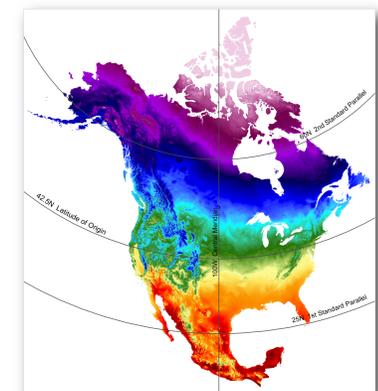
Used by Global Positioning System

### Example 2: Daymet Data

Use standard ways to define customized Spatial Reference System

```
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:longitude_of_prime_meridian = 0. ;
SRS Definition following CF Convention
```

```
*proj=lcc +lat_1=25 +lat_2=60 +lat_0=42.5 +lon_0=-100 +x_0=0
+y_0=0 +datum=WGS84 +units=m +no_defs *
SRS Definition in PROJ.4 format (widely supported by Open Source GIS software)
```



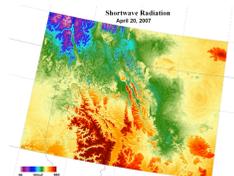
Average Annual Maximum Temperature for 1998  
(<http://daymet.ornl.gov>)

### Good Formats

- Open and non-proprietary
- Self descriptive; file contains interpretive metadata

- Feature Data Formats
  - Shapefile
  - ASCII

- Coverage Data Formats
  - GeoTIFF
  - netCDF v3/v4
  - HDF/HDF-EOS



## Archive your data

When the results of your project have been published, it is time to share your data products with the scientific community by archiving them at the ORNL DAAC.

A data set contains:

- Data files
- Documentation
- Answers to data provider questions: why, what, where, when, how, and who  
<http://daac.ornl.gov/PI/questions.shtml>

## Additional resources

[http://daac.ornl.gov/PI/pi\\_info.shtml](http://daac.ornl.gov/PI/pi_info.shtml)

Contact: Tammy Beaty ([beatytw@ornl.gov](mailto:beatytw@ornl.gov))

Sponsored by the National Aeronautics and Space Administration under Interagency Agreement NNG14HH39L. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S Dept. of Energy under contract DE-AC05-00OR22725