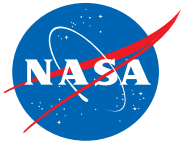


Environmental Data Management Best Practices Webinar

Part 2 – Geospatial data

NASA EarthData Webinar

Introduction to the ORNL DAAC

Suresh K.S. Vannan

Environmental Sciences Division

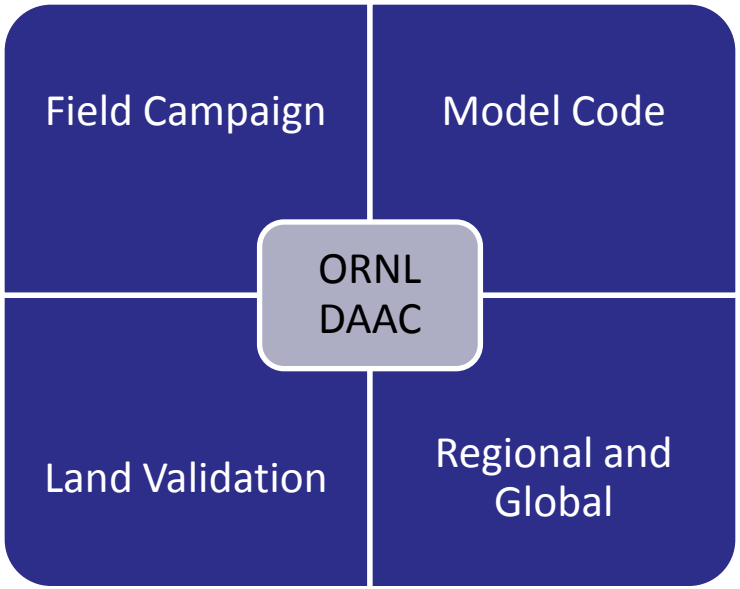
Oak Ridge National Laboratory

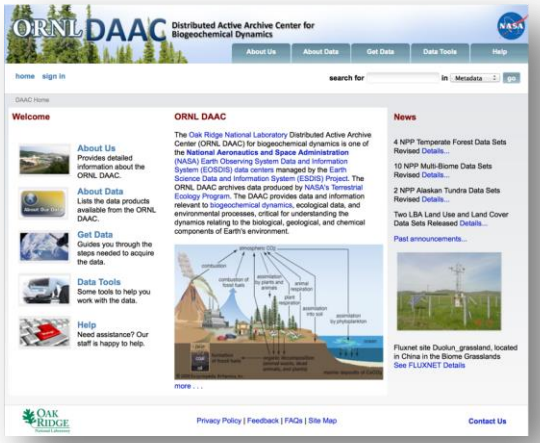
September 12, 2013

NASA EarthData Webinar

About ORNL DAAC

The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) archives data produced by NASA's Terrestrial Ecology Program in support of NASA's Carbon Cycle and Ecosystems Focus Area.





http://daac.ornl.gov

What's in it for you?

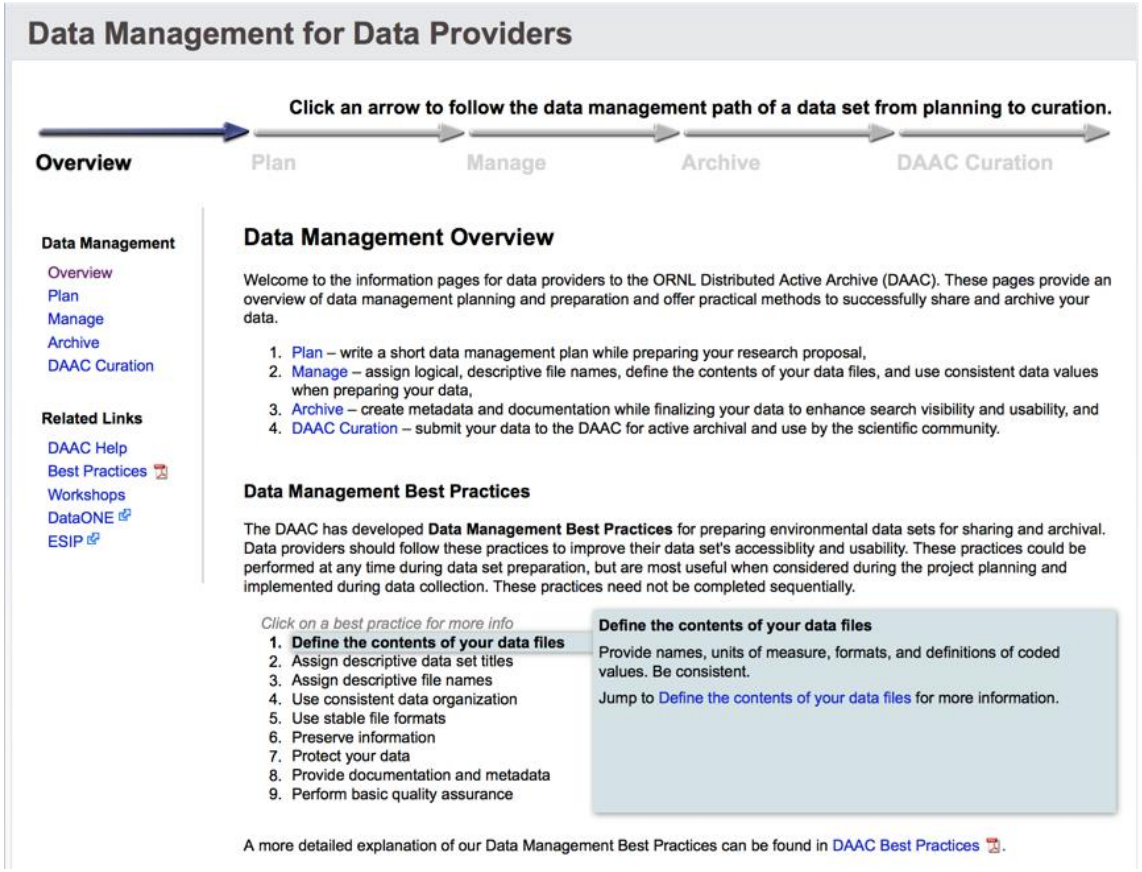
- Protect research investments
- Data in general cannot be collected again (especially observational data)
- Credit for data publication and further science

ELSEVIER

THOMSON REUTERS

Resources

http://daac.ornl.gov/PI/pi_info.shtml



Data Management for Data Providers

Click an arrow to follow the data management path of a data set from planning to curation.

Overview → Plan → Manage → Archive → DAAC Curation

Data Management

- Overview
- Plan
- Manage
- Archive
- DAAC Curation

Related Links

- DAAC Help
- Best Practices
- Workshops
- DataONE
- ESIP

Data Management Overview

Welcome to the information pages for data providers to the ORNL Distributed Active Archive (DAAC). These pages provide an overview of data management planning and preparation and offer practical methods to successfully share and archive your data.

1. Plan – write a short data management plan while preparing your research proposal,
2. Manage – assign logical, descriptive file names, define the contents of your data files, and use consistent data values when preparing your data,
3. Archive – create metadata and documentation while finalizing your data to enhance search visibility and usability, and
4. DAAC Curation – submit your data to the DAAC for active archival and use by the scientific community.

Data Management Best Practices

The DAAC has developed **Data Management Best Practices** for preparing environmental data sets for sharing and archival. Data providers should follow these practices to improve their data set's accessiblity and usability. These practices could be performed at any time during data set preparation, but are most useful when considered during the project planning and implemented during data collection. These practices need not be completed sequentially.

Click on a best practice for more info

1. **Define the contents of your data files**
2. Assign descriptive data set titles
3. Assign descriptive file names
4. Use consistent data organization
5. Use stable file formats
6. Preserve information
7. Protect your data
8. Provide documentation and metadata
9. Perform basic quality assurance

Define the contents of your data files

Provide names, units of measure, formats, and definitions of coded values. Be consistent.

Jump to Define the contents of your data files for more information.

A more detailed explanation of our Data Management Best Practices can be found in DAAC Best Practices.

Best Practices for Preparing Geospatial Data for Sharing and Archiving

Yaxing Wei

Environmental Sciences Division

Oak Ridge National Laboratory

September 12, 2013

NASA EarthData Webinar

Presenter: Yaxing Wei

- Geospatial Information Scientist, NASA's ORNL Distributed Active Archive Center for Biogeochemical Dynamics
- Data Management Support
  - North American Carbon Program
  - National Hydropower Asset Assessment Project
- Oak Ridge National Laboratory, Oak Ridge, TN
- weiy@ornl.gov
- Phone: +1 865 241-3403

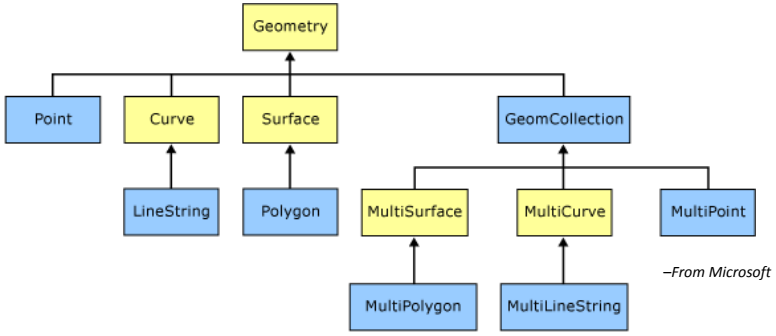
ORNL, Oak Ridge, TN

Agenda

- Geospatial Data
- Critical Information for Geospatial Data
- Best Practices
  - Provide Geospatial Information
  - Provide Temporal Information
  - Provide Data Content
  - Choose Data Format
  - Selected Geospatial Data Tools
- Benefits of Following Best Practices

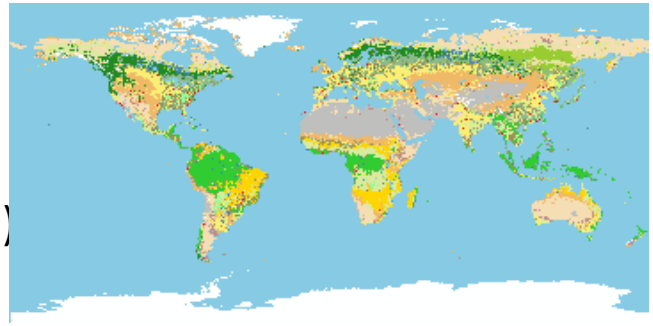
Geospatial Data

- Data with location information
  - Feature data: "object" with location and other properties
    - AmeriFlux sites, National Hydrography Dataset, ecoregion boundaries



–From Microsoft

  - Coverage data: "phenomenon" spanning spatial extent / temporal period / …
    - AmeriFlux site GPP time series (1-D)
    - one year of MODIS land cover (2-D)
    - global 1° monthly model output NEE (3-D)
    - ….



MODIS IGBP Land Cover (2007)

Critical Information for Geospatial Data

- Where: *geospatial information*
  - Spatial Reference System: datum and projection
  - Spatial extent/resolution
- When: *temporal information*
  - Calendar
  - Time units & extent/resolution
- What: *data content*
  - Variable name, units, missing value, …
- Who, Why, and How

Bottom Line

The critical information has to be provided and correct!

Best Practices

- **Best Practices for Providing Geospatial Information**
- Best Practices for Providing Temporal Information
- Best Practices for Providing Data Content
- Best Practices for Choosing Data Formats
- Best Practices for Geospatial Data Tools

Geospatial Example (1)

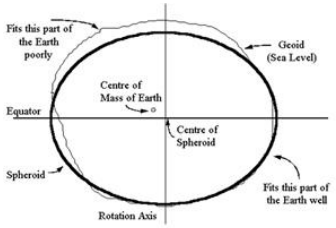
- AmeriFlux sites as Point data

SITEID	SITENAME	STATUS	TOWER_BEGAN (YYYY)	TOWER_END (YYYY)	LAT (dec. deg)	LONG (dec. deg)	ELEV (m)	IGBP	CLIMATE_KOPPEN	MAT (deg C)	MAP (mm)
PA-Bar	Panama - Barro Colorado Island	Active	2012		9.1540	-79.8480	150	EBF	Af	26.00	2800.00
US-Akn	USA - SC - Aiken	Active	2011		33.3833	-81.5656	92	MF	TBD		
US-ORv	USA - OH - Olentangy River Wetland Research Park	Active	2011		40.0201	-83.0183	221	WET	TBD		
US-Dia	USA - CA - Diablo	Active	2010		37.6773	-121.5296	323	GRA	TBD		

Table from http://ameriflux.lbl.gov

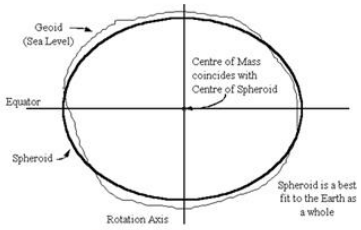
Spatial Reference System (SRS)

- Datum: a system which allows the location of latitudes and longitudes (and heights) to be identified onto the surface of the Earth
  - Sphere / Spheroid



Local-Referencing

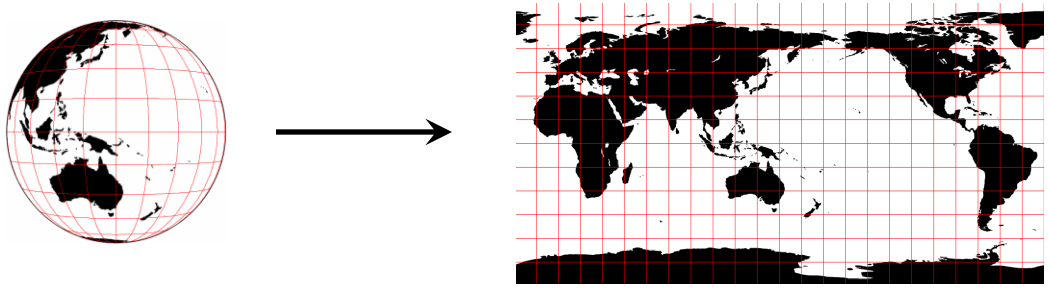
Examples: North American Datum of 1927 (NAD27); North American Datum of 1983 (NAD83)



Global-Referencing

Example: World Geodetic System of 1984 (WGS84)

- Projection: define a way to flatten the Earth surface



- SRID: code representing pre-defined popular SRS, e.g. EPSG:4326
  - http://spatialreference.org

Geospatial Example (1) Con't

- AmeriFlux site as Point data

SRS: WGS 84 (EPSG:4326) ← Used by GPS

SITEID	SITENAME	STATUS	TOWER_BEGAN (YYYY)	TOWER_END (YYYY)	LAT (dec. deg)	LONG (dec. deg)	ELEV (m)	IGBP	CLIMATE_KOPPEN	MAT (deg C)	MAP (mm)
PA-Bar	Panama - Barro Colorado Island	Active	2012		9.1540	-79.8480	150	EBF	Af	26.00	2800.00
US-Akn	USA - SC - Aiken	Active	2011		33.3833	-81.5656	92	MF	TBD		
US-ORv	USA - OH - Olentangy River Wetland Research Park	Active	2011		40.0201	-83.0183	221	WET	TBD		
US-Dia	USA - CA - Diablo	Active	2010		37.6773	-121.5296	323	GRA	TBD		

Table from http://ameriflux.lbl.gov

Precision:

Geospatial Example (2)

- Define Geospatial Information (Regular Grid)
  - Grid cells are rectangular (e.g. NACP regional terrestrial biosphere model outputs)
    - Define your SRS
      - Sphere-based GCS (radius of the Earth: 6,370,997m)
    - Provide X/Y spatial resolution: size of a grid cell
      - X: 1-degree, Y: 1-degree
    - Provide spatial extent: outer boundary of all cells
      - West: -170, South: 10, East: -50, North: 84

Option 1

    - Provide coordinates of each grid cell center
    - Provide coordinates of 4 borders of each grid cell

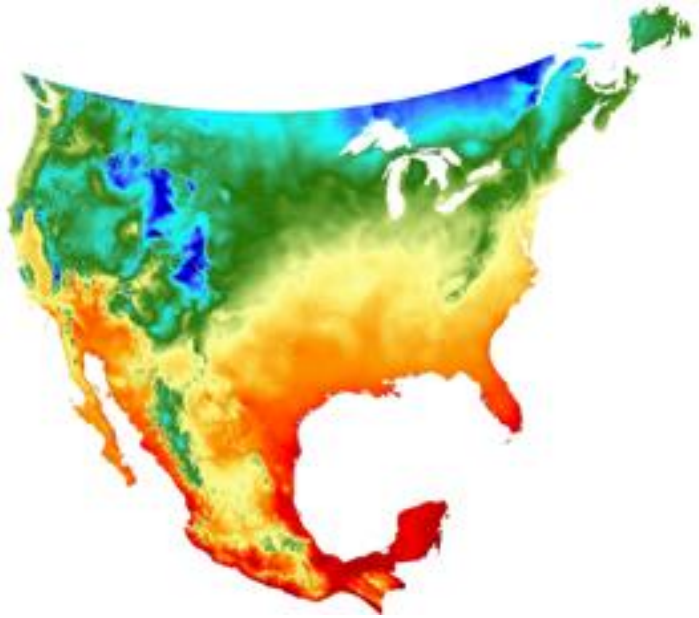
Option 2

Geospatial Example (2) Con't

- Define Geospatial Information (Irregular Grid)
  - Cells are not rectangular (e.g. 10242 Spherical Geodesic Grid)
    - Define your SRS
    - Provide coordinates for all vertices of each polygon in counter-clockwise order
    - Provide coordinates for the mass center of each polygon (if possible)

Geospatial Example (3)

- SRS for Daymet data
  - Datum: North_American_Datum_1983
  - Projection: Lambert Conformal Conic
    - units: meters
    - 1st standard parallel: 25 deg N
    - 2nd standard parallel: 60 deg N
    - Central meridian: 100 deg W
    - Latitude of origin: 42.5 deg N
    - false easting: 0
    - false northing: 0



Daymet Minimum Temperature

(http://daymet.ornl.gov)

Choose Proper Projection

- Preserve Direction
  - Projection: Lambert Conformal Conic
  - Research: navigation, weather, …
- Preserve Area
  - Projection: Albers Equal Area
  - Research: land use, density of bird population, …
- Preserve Distance
  - Projection: Equidistant Conic
  - Research: earthquake, …

Best Practices

- Best Practices for Providing Geospatial Information
- **Best Practices for Providing Temporal Information**
- Best Practices for Providing Data Content
- Best Practices for Choosing Data Formats
- Best Practices for Geospatial Data Tools

Temporal Example (1)

- Specify Calendar
  - **julian**: one leap year in every 4 years
  - **gregorian**: leap year if either (1) it is divisible by 4 but not by 100 or (2) it is divisible by 400
  - **proleptic_gregorian**: gregorian calendar extended to dates before 1582-10-15
  - **365_day**: no leap year, Feb. always has 28 days
  - **360_day**: 30 days for each month
  - **366_day**: all leap years

gregorian is the internationally used civil calendar

Temporal Example (2)

- Specify Time
  - "the measurement was made at 6 in the afternoon **on March 22, 2010 and it took 1 hour 20 minutes and 30 seconds" - BAD**
- ISO 8601: date, time, and duration
  - Date/Time point: YYYY-MM-DDThh:mm:ss.sTZD

    2010-03-22T18:00:00.00-06:00
  - Duration: P[n]Y[n]M[n]DT[n]H[n]M[n]S

    PT1H20M30S

Best Practices

- Best Practices for Providing Geospatial Information
- Best Practices for Providing Temporal Information
- **Best Practices for Providing Data Content**
- Best Practices for Choosing Data Formats
- Best Practices for Geospatial Data Tools

Define Data Content

- Variable Name
  - Brief and descriptive, short name and long name
  - Try to follow your communities' rules

    Climate & Forecast (CF) Standard Names
- Use Keywords to Tag Your Data

  GCMD Science Keywords

  Semantic Web for Earth and Environmental Terminology (SWEET) Ontologies
- Description and Abstract

Data Units

- Separate content from units

  kg_C_per_square_meter_per_second

- Follow standards

  UDUNITS-2

  kg/m2/s

  kg m-2 s-1

  celsius, degC, Kelvin, degK, degree (*angular degree*)

  60 second, (5 meter)/(30 second)

Missing Values

- Avoid NaN
- Use a Missing Value Code

  -9999.0
- Define a Range of Valid Values

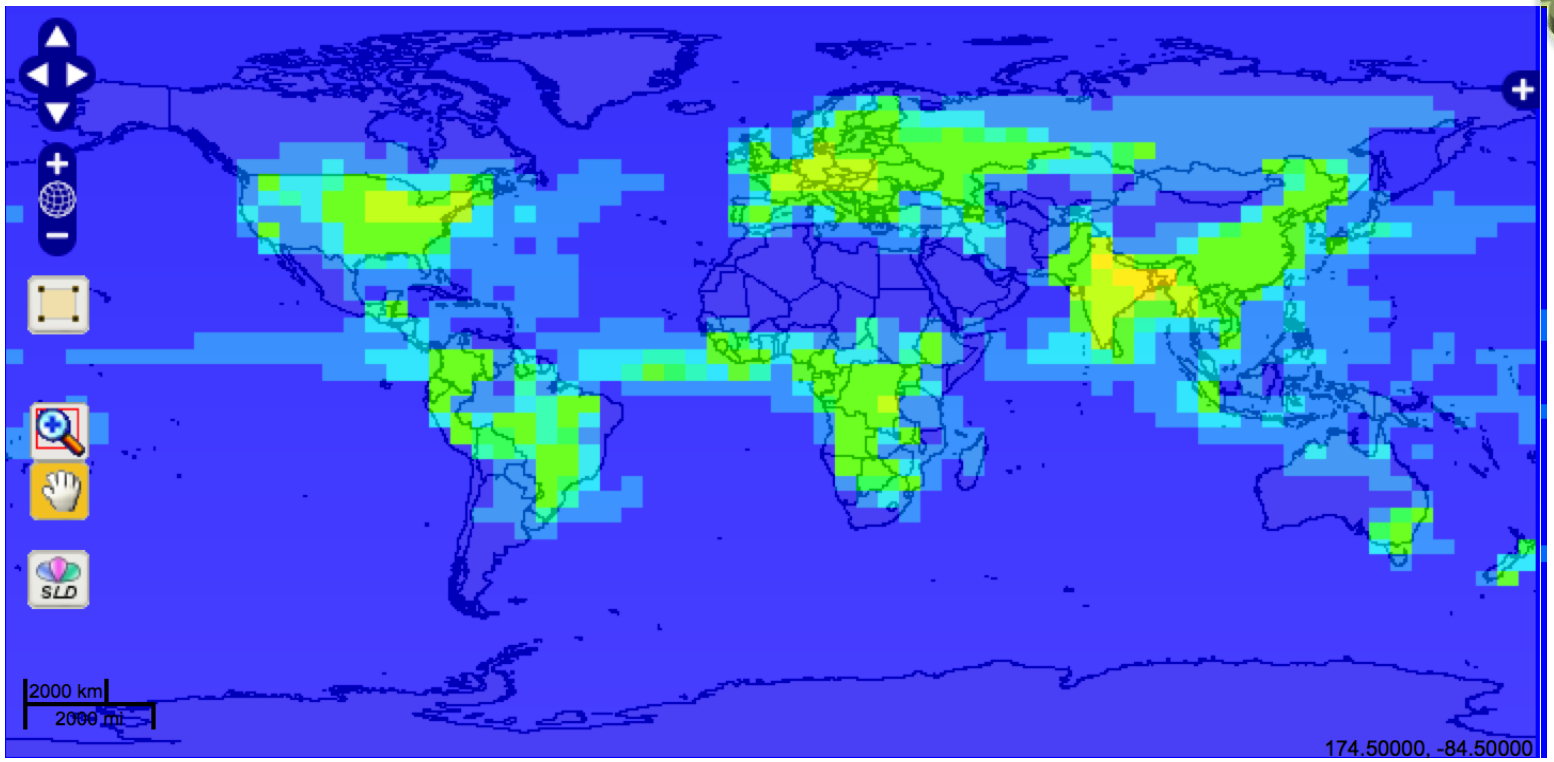
  valid_min=-50.0, valid_max=100.0

  or

  valid_range=-50.0,100.0

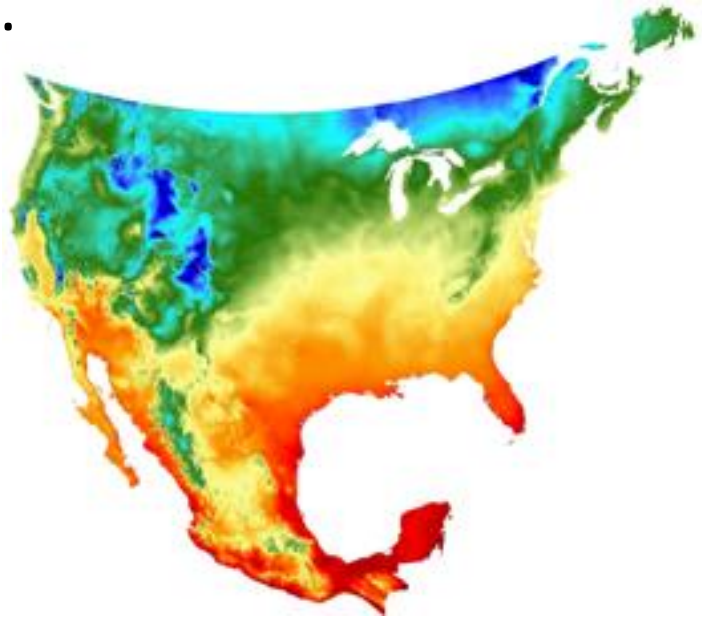
Bad Practice Example (1)

- Global Maps Of Atmospheric Nitrogen Deposition, 1860, 1993, and 2050



Bad Practice Example (2)

- Time in Daymet
  - Daymet has data for 365 days in each year
  - Calendar: 365_day
  - No! It has leap years. It removed December 31st instead of Feb 29th in leap years.
  - Calendar: gregorian



A Not-so-Good Practice Example

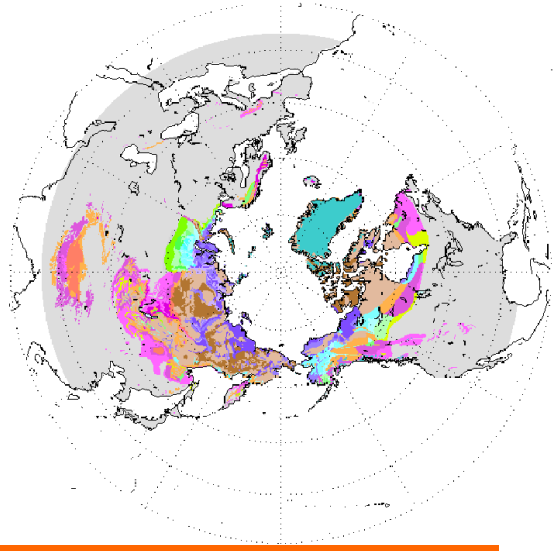
- Circum-Arctic Map of Permafrost and Ground Ice Conditions
  - It provides a 25km by 25km gridded map in BINARY format along with a header file and SRS definition in readme

Header:

nrows 721
ncols 721
nbits 8
byteorder I
ulxmap -9024309
ulymap 9024309
xdim 25067.525
ydim 25067.525

SRS Definition:

Projection: Lambert Azimuthal
Units: meters
Spheroid: defined
Major Axis: 6371228.00000
Minor Axis: 6371228.000
longitude of center of projection: 0
latitude of center of projection: 90
false easting (meters): 0.00000
false northing (meters): 0.00000



Bad Data Format

Best Practices

- Best Practices for Providing Geospatial Information
- Best Practices for Providing Temporal Information
- Best Practices for Providing Data Content
- **Best Practices for Choosing Data Formats**
- Best Practices for Geospatial Data Tools

“Good” Formats

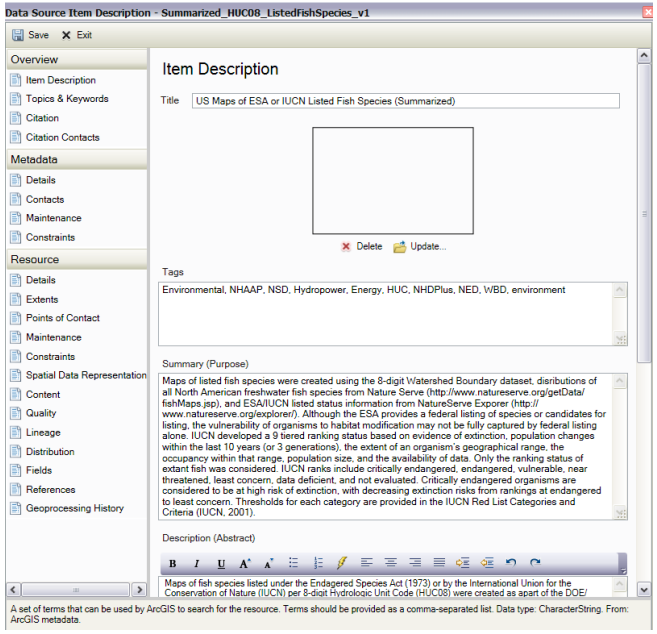
- Open and non-proprietary
- Simple and commonly used
- More importantly, self-descriptive
  - Metadata is included inside data

- Feature Data Formats
  - Shapefile
  - KML
  - GML
  - ESRI Geodatabase

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

Shapefile

- Ideal for feature data
  - point, line, and polygon
- SRS can be embedded inside files (*.prj)
- Metadata can be embedded inside files (*.xml)



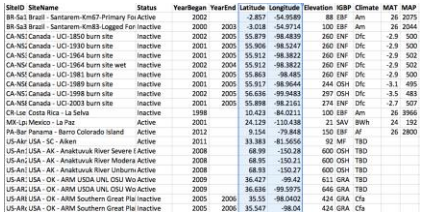
NetCDF

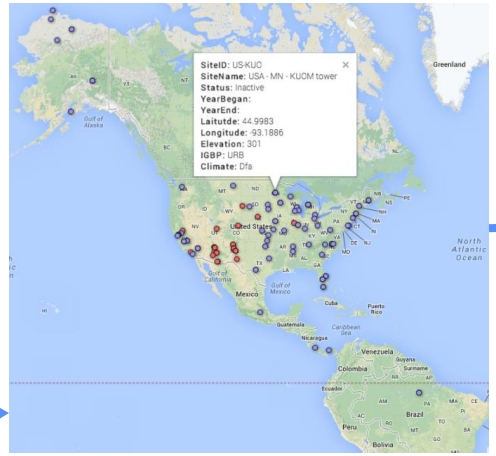
- Ideal for multi-dimensional data
- CF metadata convention
  - Standard variable names
  - Spatial/temporal coordinates
  - Cell boundaries/shape/methods
  - Missing data
  - Data units
  - …

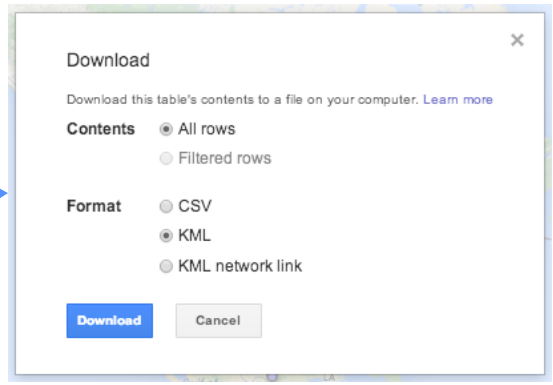
KML – Keyhole Markup Language

- Ideal for feature data: point, line, and polygon
- Ideal for visualize and quality check data
  - Google Map, Google Earth, …

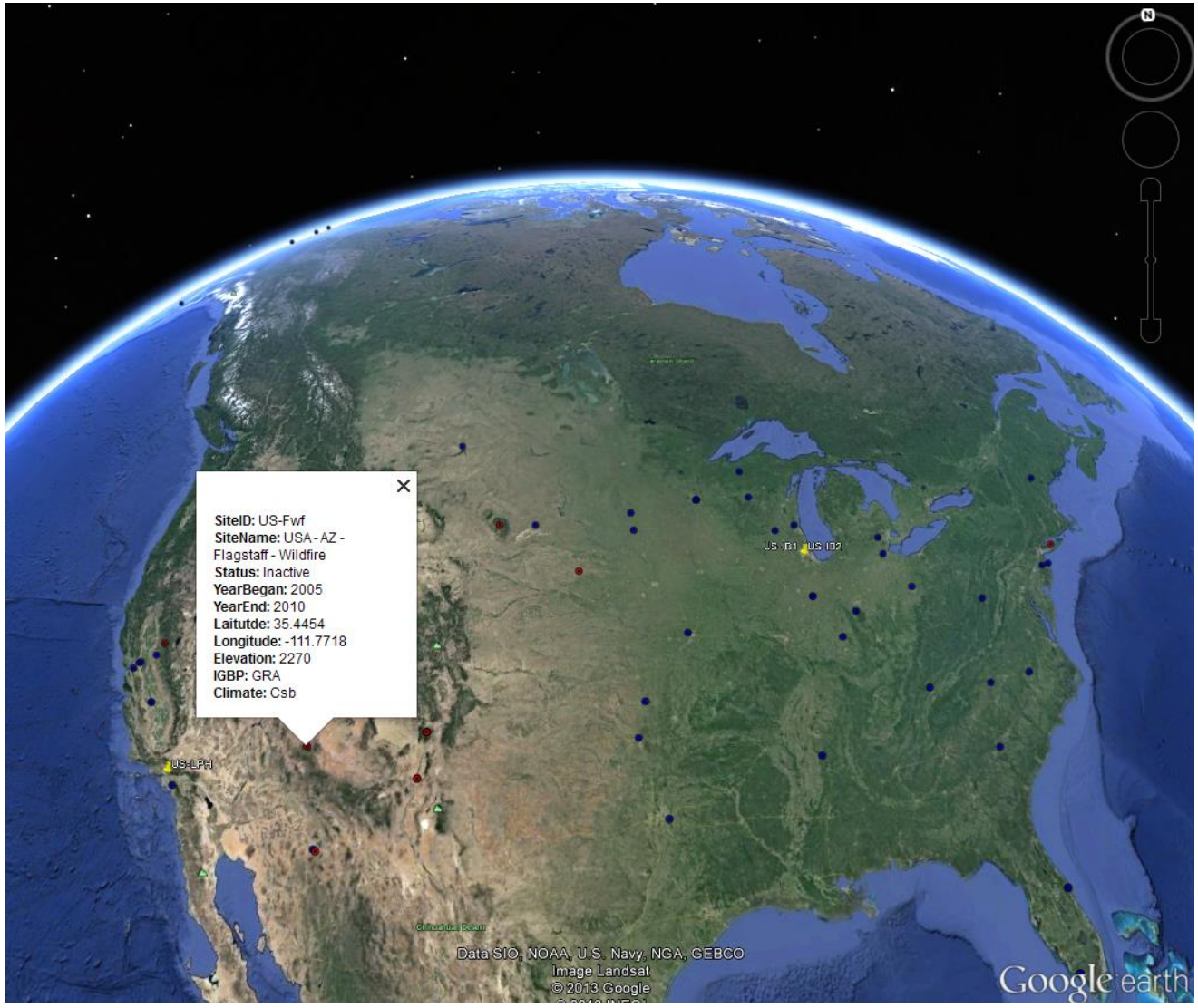
Google Fusion Table







KML in Google Earth



Best Practices

- Best Practices for Providing Geospatial Information
- Best Practices for Providing Temporal Information
- Best Practices for Providing Data Content
- Best Practices for Choosing Data Formats
- **Best Practices for Geospatial Data Tools**

GDAL/OGR

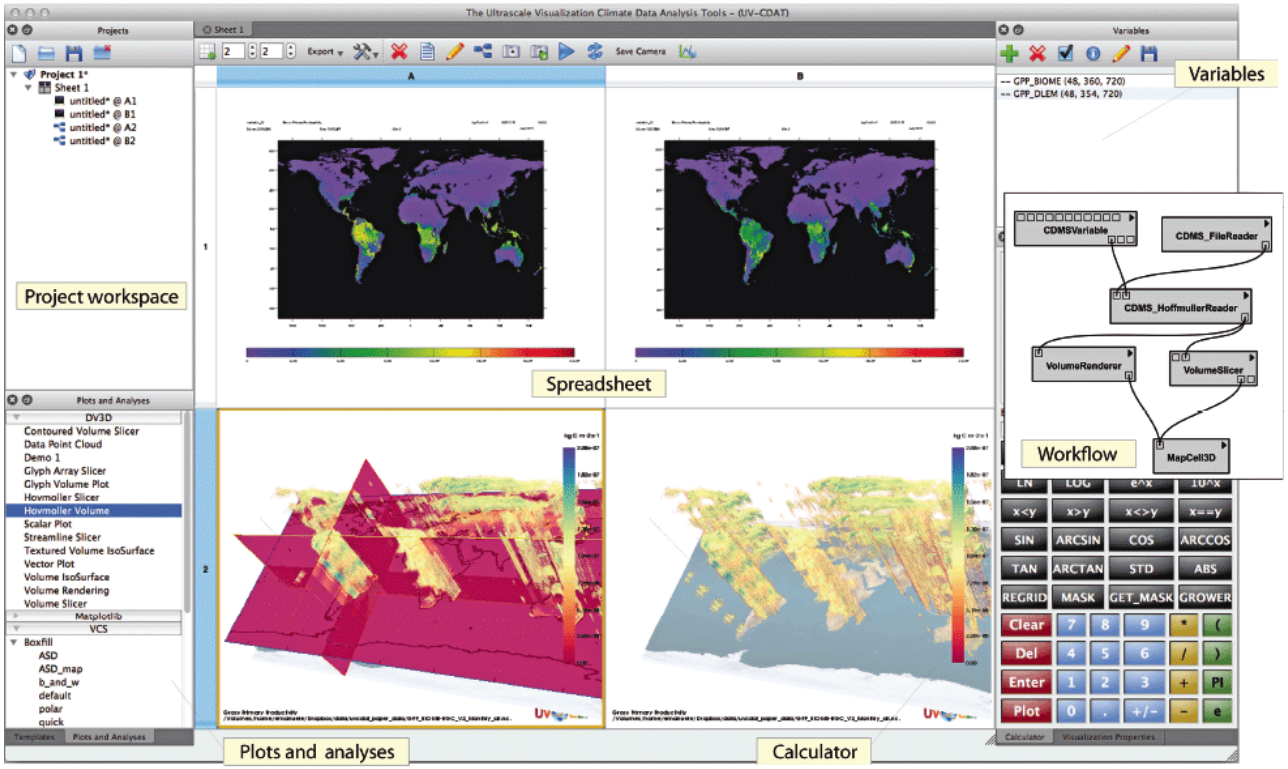
- GDAL: Geospatial Data Abstraction Library
  - Raster Data
- OGR: Simple Feature Library
  - Feature Data
- Available on many OS
  - Linux, Unix, Mac OS X, Windows, …
- Ideal for data conversion
  - gdal_translate: 130+ raster data formats
  - ogr2ogr: 70+ feature data formats

NCO: NetCDF Operator

- Ideal for quick manipulation of netCDF data
- Command line utilities
  - **ncatted**: edit attributes
  - **ncrename**: rename attr, dim, and var
  - **ncbo**: binary operator (+, -, *, /)
  - **ncks**: extract vars, copy var from another file, …
  - **ncrcat**: merge multiple time steps together
  - **ncap2**: write simple scripts
  - ….

UV-CDAT

- UV-CDAT: Ultrascale Visualization Climate Data Analysis Tools
  - Support netCDF, HDF, …



Santos et al., 2013

Benefits from Following Best Practices

- Make your data easily understood by others
  - promote sharing and research
- Make your data ready to be used by tools
  - ArcGIS, Matlab, R, NCO, CDO, NCL, VisTrails, UV-CDAT, …
- Bring science researchers (you) and data management people (us) closer.
  - Benefit from the information infrastructures we provide
  - Your data can be ingested into many existing Web services to provide on-demand data distribution to users
- Value of your data can be preserved into the future

Summary

- Provide geospatial, temporal, other information completely and accurately
- Choose good formats to organize the data content and make them self-descriptive
- Provide metadata in standard ways
- There are many benefits