Better than Free:
Data Explorations with Public Data and Software Tools

Introduction to Data Formats

Mary Haley

SEA 2016 Tutorial

April 6, 2016

Definition: data format (sometimes called *file format*)

The structure of a computer file. There are hundreds of different formats for data (databases, text, images, video, etc).

Each format defines how the sequence of bits and bytes are laid out, with the ASCII text file being the simplest.

Quick survey (1/2)

What types of data do you work with?

- ASCII, CSV (comma separated values)
- Binary (Fortran, C, GRIB)
- Excel type spreadsheets
- Self-describing (NetCDF, HDF)
- Not sure

IBTrACS WMO: International Best Tracks Archive for Climate Stewardship -- WMO DATA ONLY -- Version: v03r04

Serial_Num,Season,Num,Basin,Sub_basin,Name,ISO_time,Nature,Latitude,Longitude,Wind(WMO),Pres(WMO),Center,Wind(WMO) Percentile,Pres(WMO) Percentile,Track_type

N/A,Year,#,BB,BB,N/A,YYYY-MM-DD HH:MM:SS,N/A,deg_north,deg_east,kt,mb,N/A,%,%,N/A

1848011S09080,1848,02, SI, MM,XXXX848003,1848-01-11 06:00:00, NR, -8.60, 79.80, 0.0, 0.0,reunion,-100.000,-100.000,main

1848011S09080,1848,02, SI, MM,XXXX848003,1848-01-12 06:00:00, NR, -9.00, 78.90, 0.0, 0.0,reunion,-100.000,-100.000,main

1848011S09080,1848,02, SI, MM,XXXX848003,1848-01-13 06:00:00, NR,-10.40, 73.20, 0.0, 0.0,reunion,-100.000,-100.000,main

1848011S09080,1848,02, SI, MM,XXXX848003,1848-01-14 06:00:00, NR,-12.80, 69.90, 0.0, 0.0,reunion,-100.000,-100.000,main

1848011S09080,1848,02, SI, MM,XXXX848003,1848-01-15 06:00:00, NR,-13.90, 68.90, 0.0, 0.0,reunion,-100.000,-100.000,main

1848011S09080,1848,02, SI, MM,XXXX848003,1848-01-16 06:00:00, NR,-15.30, 67.70, 0.0, 0.0,reunion,-100.000,-100.000,main

1848011S09080,1848,02, SI, MM,XXXX848003,1848-01-17 06:00:00, NR,-16.50, 67.00, 0.0, 0.0,reunion,-100.000,-100.000,main

1848011S09080,1848,02, SI, MM,XXXX848003,1848-01-18 06:00:00, NR,-18.00, 67.40, 0.0, 0.0,reunion,-100.000,-100.000,main

Sample CSV file

```
USC0004029189403TMAX-9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999
-9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999      6  6   28  6   39  6   56  6
117  6  150  6  172  6  156  6  167  6  178  6  117  6  128  6  111  6
USC0004029189403TMIN-9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999
-9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999    -83  6  -67  6  -44  6  -78  6
-50  6  -39  6  -11  6   11  6   39  6   50  6   56  6  -11  6    0  6
USC0004029189403PRCP   76  6    0P 6    0P 6    0P 6   51  6    0P 6    0P 6    0P 6   25  6    0P 6
0P 6    0P 6    0P 6    0P 6   20  6    0P 6    0P 6    0P 6   18  6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6
0P 6    0P 6    0P 6   81  6   53  6    0P 6
USC0004029189403SNOW   76  6    0  6    0  6    0  6   51  6    0  6    0  6    0  6   25  6    0  6    0  6
0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0  6    0
6    0  6    0  6    0  6
USC0004029189404TMAX  144  6  106  6  128  6  167  6  172  6  183  6  139  6  161  6  217  6
194  6  117  6  161  6  178  6  139  6   56  6   94  6  178  6  194  6  233  6  217  6  211  6  144  6  117
6  189  6  167  6   56  6   39  6   94  6  167  6  183  6-9999
USC0004029189404TMIN   28  6  -11  6   -6  6   -6  6   -6  6    6  6   61  6   -6  6   28  6   50  6    6  6
-22  6   11  6   22  6  -11  6  -39  6  -33  6    0  6   22  6   67  6   61  6   61  6   22  6   -6  6   61  6  -17
6  -50  6  -11  6   33  6   17  6-9999
USC0004029189404PRCP    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6
0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6    0P 6
211  6-9999      0P 6    0P 6    0P 6-9999
```

Sample ASCII file

Data formats: self-describing

Self-describing data formats are files that contain data, plus descriptive information about the data ("metadata").

Metadata can be information about the file itself and about the variables on the file

Data formats: self-describing

Metadata generally includes:

- Attributes
- Dimension names and sizes
- Coordinate information

Metadata: attributes

Sample file attributes

- *institution*
- *creation_date*
- *history*
- *resolution*
- *SatelliteName*
- *Conventions*

Sample variable attributes

- *_FillValue* / *missing_value*
- *long_name* / *description* / *standard_name*
- *scale_factor* / *add_offset*
- *units*
- *coordinates*

Metadata: dimension names and sizes

Defines dimension sizes of variables on the file

- time
- level / lv_ISBL9
- lat / latitude / lat_3
- lon / longitude / lon_3
- x / y / z
- Cell_Across_Swath_mod07

Metadata: coordinate information

Special variables on the file that contain coordinate values

- Time steps
- Latitude and longitude coordinates
- Pressure levels / height

Pros of self-describing formats

- Well-written (*) files describe themselves
- Can query file for information
- Essential for subsetting and aggregation

* There are conventions; not everybody follows them!

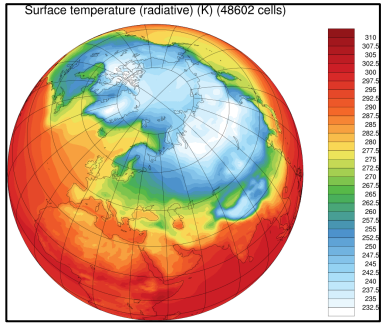
Cons of self-describing formats

- Complex – requires special software to read and write
- Some formats still evolving
- Can be large
- Conventions not always adopted or available

Self-describing data formats

NetCDF (**Net**work **C**ommon **D**ata **F**orm)

- Very common in climate sciences
- Developed and supported by Unidata
- Two versions: NetCDF-3 and NetCDF-4
  http://www.unidata.ucar.edu/software/netcdf/
- Conventions
  http://www.unidata.ucar.edu/software/netcdf/conventions.html



NetCDF tip

Determining the type of your NetCDF file

```
ncdump -k   my_netcdf_file
```

classic 64-bit offset netCDF-4

netCDF-4 classic model

NetCDF Conventions

- CF (Climate and Forecast) Conventions
  http://cfconventions.org/

- More conventions
  http://www.unidata.ucar.edu/software/netcdf/conventions.html

Look for "Conventions" attribute on file

Another NetCDF tip

CF-convention compliance checker

http://puma.nerc.ac.uk/cgi-bin/cf-checker.pl

Final NetCDF tip

Doing quick operations across multiple NetCDF files

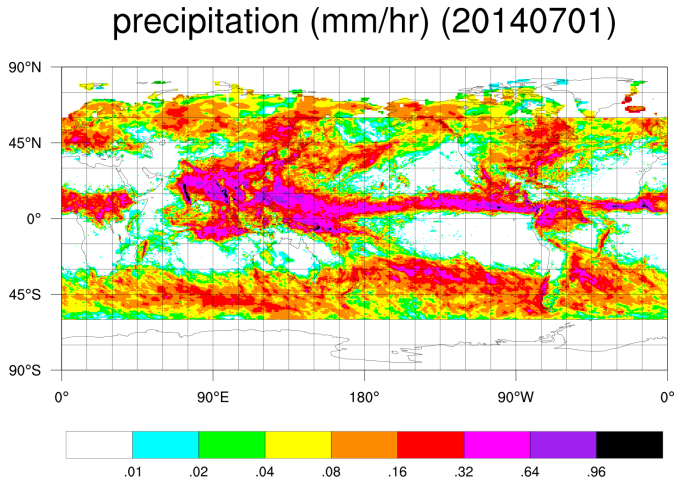
NetCDF Climate Operators (NCO)
http://nco.sourceforge.net/

Climate Data Operators (CDO)
https://code.zmaw.de/projects/cdo

Self-describing data formats

HDF (**H**ierarchical **D**ata **F**ormat)

- Tailored for large and complex datasets
- Used by a wide variety of scientific disciplines
- Two versions HDF4 / HDF5
  http://www.hdfgroup.org

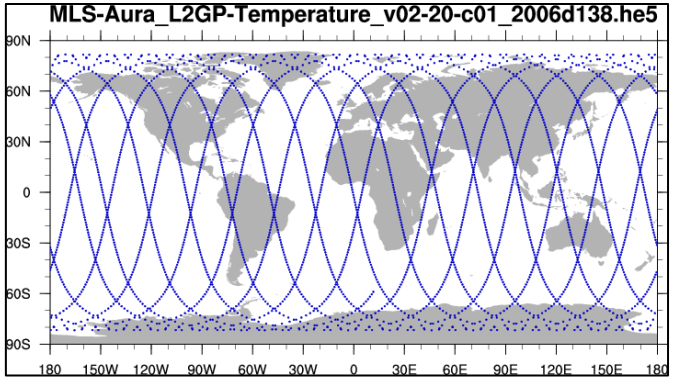


Self-describing data formats

HDF-EOS (**HDF E**arth **O**bserving **S**ystem)

- HDF4 and HDF5 subset with conventions, data types, and metadata
- Used for NASA EOS missions (mostly satellite)
- Geo-located data

  http://hdfeos.net

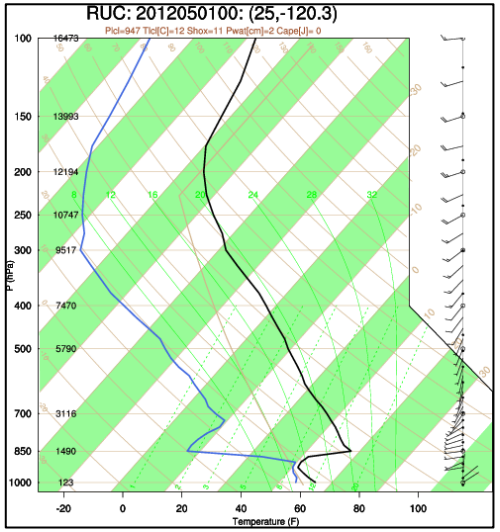


~~Self~~-describing data formats

GRIB (**Gri**dded **B**inary)

General **R**egularly-distributed **I**nformation in **B**inary form

- World Meteorological Organization standard
- Historical / forecast weather data
- Actually a “record” format
- Requires look-up tables for the metadata (GRIB code tables)



Why is this all important?

- "Know your data"
- Match tools with data
- What's important to you?
  - File portability
  - File size
  - Handling complex data
  - Readability
  - Descriptive information
  - Easy to use

Help available from NCAR

Experts in RDA

RDA blog
http://ncarrda.blogspot.com/

Climate Data Guide
https://climatedataguide.ucar.edu/