Best Practices for Preparing Data for Sharing and Archiving

Bob Cook

Environmental Sciences Division

Oak Ridge National Laboratory

September 12, 2013

NASA EarthData Webinar

Presenter: Bob Cook

- Biogeochemist
- Chief Scientist, NASA's ORNL Distributed Active Archive Center for Biogeochemical Dynamics
- Associate Editor, *Biogeochemistry*
- Oak Ridge National Laboratory, Oak Ridge, TN
- cookrb@ornl.gov
- Phone: +1 865 574-7319

ORNL, Oak Ridge, TN

Best Practices for Preserving Data

Agenda:

- Introduction
- Metadata
- Fundamental Best Practices

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
- Collaborators can readily understand and use data files

Long-term (data publication)

- Scientists outside your project can find, understand, and use your data to address broad questions
- You get credit for archived data products and their use in other papers
- Sponsors protect their investment

Metadata

Information to find, understand, and use the data

- *descriptors*
- *documentation*

What is Metadata?

Who

Who collected the data?
Who processed the data?
Who wrote the metadata?
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?

When

When were the data collected?
When were the data processed?

What

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

How

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

Why

Why were the data collected?

This is the metadata for this.

aurora with comet

What's Missing?

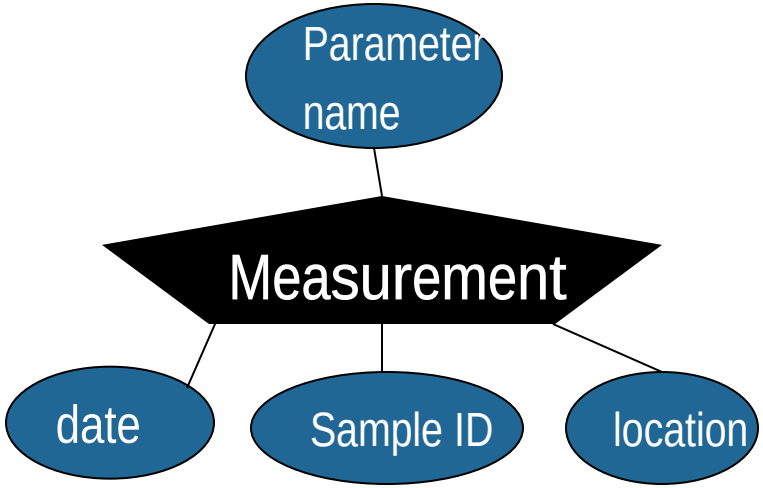
The 20-Year Rule

- The metadata accompanying a data set should be written for a user 20 years into the future
  - What does that investigator need to know to use the data?
- Prepare the data and documentation for a user who is unfamiliar with your project, methods, and observations

NRC (1991)

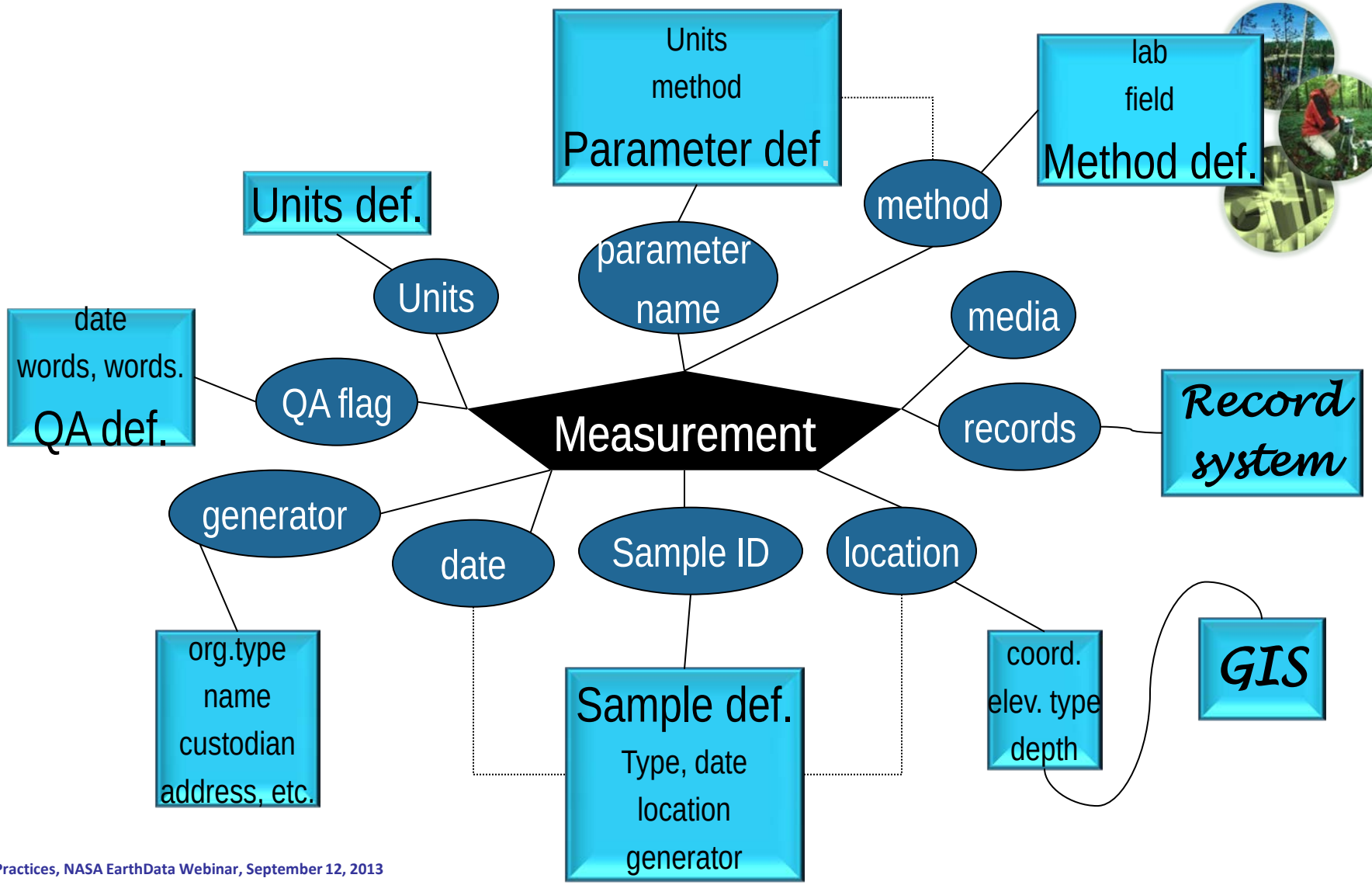
Metadata needed to Understand Data

The details of the data ….



Courtesy of Raymond McCord, ORNL

Metadata Needed to Understand Data



Fundamental Data Practices

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

1. Define the contents of your data files

- Content flows from science plan (hypotheses) and is informed from requirements of final archive.

- Keep a set of similar measurements together in one file
  - same investigator,
  - methods,
  - time basis, and
  - instrument
  - No hard and fast rules about contents of each file.

2. Define the parameters

Parameter Table

Column	Description	Units/Format
SITE	k=Kataba forest, p=Pandamatenga, m=Near Maun, e=HOORC/MPG Maun tower, o=Okwa river crossing, t=Tshane, skukuza=Skukuza Flux Tower	text
SPECIES	Scientific name up to 25 characters	text
DATE	Date of measurement	yyyymmdd
BA	Woody plant basal area	m2/ha
SEBA	Standard error of BA	m2/ha
DENSITY	Woody plant density (number of trees per hectare)	number/ha
SEDEN	Standard error of DENSITY (n=42 for KT, n=49 for Skukuza)	number/ha
STEMS	Number of stems per hectare (/ha)	number/ha
HEIGHT	Basal area-weighted average height	m2/ha
WOOD	Aboveground woody plant wood dry biomass	kg/ha
LEAF	Aboveground woody plant leaf dry biomass	kg/ha
LAI	Leaf Area Index calculated by allometry	m2/m2

Scholes (2005)

- Use your community's commonly accepted parameter names and units
- Be consistent
- Explicitly state units

2. Define the parameters (cont)

- Choose a format for each parameter, explain the format in the metadata, and use that format throughout the file
  - e.g., use yyyymmdd; January 2, 1999 is 19990102
  - Report in both local time and Coordinated Universal Time (UTC) and 24-hour notation (13:30 instead of 1:30 p.m.)
    - Avoid Daylight standard time
  - Use a code (e.g., -9999) for missing values

- See Hook et al. (2010) for additional examples of parameter formats
    - http://daac.ornl.gov/PI/bestprac.html#prac3

3. Use consistent data organization (one good approach)

Each row in a file represents a complete record, and the columns represent all the parameters that make up the record.

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

Note: -9999 is a missing value code for the data set

3. Use consistent data organization (a 2nd good approach)

Parameter name, value, and units are placed in individual rows.

This approach is used in relational databases.

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

3. Use consistent data organization (cont)

- Be consistent in file organization and formatting
  - don’t change or re-arrange columns
  - Include header rows (first row should contain file name, data set title, author, date, and companion file names)
  - column headings should describe content of each column, including one row for parameter names and one for parameter units

Example of Poor Data Practices for Collaboration and Data Sharing

C:\Documents and Settings\hampton\My Documents\NCEAS Distributed Graduate Seminars\[Wash Cres Lake Dec 15 Dont_Use.xls]Sheet1

Stable Isotope Data Sheet

Sampling Site / Identifier:	Wash Cresc Lake	Peter's lab	Don't use - old data
Sample Type:	Algal	Washed Rocks	
Date:	Dec. 16		
Tray ID and Sequence:	Tray 004		

Reference statistics: SD for delta ^{13}C = 0.07 SD for delta ^{15}N = 0.15

Position	SampleID		Weight (mg)	%C	delta 13C	delta 13C_ca	%N	delta 15N	delta 15N_ca	Spec. No.			
A1		ref	0.98	38.27	-25.05	-24.59	1.96	4.12	3.47	25354			
A2		ref	0.98	39.78	-25.00	-24.54	2.03	4.01	3.36	25356			
A3		ref	0.98	40.37	-24.99	-24.53	2.04	4.09	3.44	25358			
A4		ref	1.01	42.23	-25.06	-24.60	2.17	4.20	3.55	25360		Shore	Avg Con
A5	ALG01		3.05	1.88	-24.34	-23.88	0.17	-1.65	-2.30	25362	c	-1.26	-27.22
A6	Lk Outlet Alg		3.06	31.55	-30.17	-29.71	0.92	0.87	0.22	25364		1.26	0.32
A7	ALG03		2.91	6.85	-21.11	-20.65	0.48	-0.97	-1.62	25366	c		
A8	ALG05		2.91	35.56	-28.05	-27.59	2.30	0.59	-0.06	25368			
A9	ALG07		3.04	33.49	-29.56	-29.10	1.68	0.79	0.14	25370			
A10	ALG06		2.95	41.17	-27.32	-26.86	1.97	2.71	2.06	25372			
B1	ALG04		3.01	43.74	-27.50	-27.04	1.36	0.99	0.34	25374	c		
B2	ALG02		3	4.51	-22.68	-22.22	0.34	4.31	3.66	25376			
B3	ALG01		2.99	1.59	-24.58	-24.12	0.15	-1.69	-2.34	25378	c		
B4	ALG03		2.92	4.37	-21.06	-20.60	0.34	-1.52	-2.17	25380	c		
B5	ALG07		2.9	33.58	-29.44	-28.98	1.74	0.62	-0.03	25382			
B6		ref	1.01	44.94	-25.00	-24.54	2.59	3.96	3.31	25384			
B7		ref	0.99	42.28	-24.87	-24.41	2.37	4.33	3.68	25386			
B8	Lk Outlet Alg		3.04	31.43	-29.69	-29.23	1.07	0.95	0.30	25388			
B9	ALG06		3.09	35.57	-27.26	-26.80	1.96	2.79	2.14	25390			
B10	ALG02		3.05	5.52	-22.31	-21.85	0.45	4.72	4.07	25392			
C1	ALG04		2.98	37.90	-27.42	-26.96	1.36	1.21	0.56	25394	c		
C2	ALG05		3.04	31.74	-27.93	-27.47	2.40	0.73	0.08	25396			
C3		ref	0.99	38.46	-25.09	-24.63	2.40	4.37	3.72	25398			
				23.78			1.17						

Courtesy of Stefanie Hampton, NCEAS

Stable Isotope Data at ORNL: tabular csv format

Aranabar and Macko. 2005. doi:10.3334/ORNLDAAC/783

SITE	COUNTRY	LAT	LONG	DATE	DISTURBANC	TAXONOMY	PLANT PART	NOTES	C	N	d13C	d15N
units	none	decimal degr	decimal degr	year-month	none	none	none	none	%	%	per mil	per mil
Mongu	Zambia	-15.44	23.52	2000-02	C0	Baphia mass	L	none	51.6	3	-27	1.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Bauhinia pet	L	none	47.6	2.31	-27	4.7
Mongu	Zambia	-15.44	23.52	2000-02	C0	Rubiaceae	L	none	51.8	1.9	-29	2.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Brachystegia	L	none	53.2	3.21	-25.7	4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Brachystegia	L	none	53.2	3.08	-24.5	4.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Burkea africa	L	none	49.5	1.84	-27	-1.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Fabaceae	L	shrub	46.6	2.69	-28.1	3
Mongu	Zambia	-15.44	23.52	2000-02	C0	Combretum	L	none	48.7	-9999	-28	-9999
Mongu	Zambia	-15.44	23.52	2000-02	C0	Copaifera ba	L	none	57.5	1.63	-28	3
Mongu	Zambia	-15.44	23.52	2000-02	C0	Diospyrus ba	L	none	53.1	1.47	-27	4.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Hannoa chlo	L	none	49.2	1.92	-27.8	3.8
Mongu	Zambia	-15.44	23.52	2000-02	C0	Guibourtia c	L	none	53.3	2.73	-27.4	1.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Hannoa chlo	L	none	48.1	1.41	-27.4	3.8
Mongu	Zambia	-15.44	23.52	2000-02	C0	Indigofera sp	L	none	49.4	3.21	-28.6	0.7
Mongu	Zambia	-15.44	23.52	2000-02	C0	Indigofera sp	L	none	49.8	3.26	-27.5	0.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Ochna pulch	L	mature leaf	51.4	1.68	-26.4	3.1
Mongu	Zambia	-15.44	23.52	2000-02	C0	Parinari cura	L	none	51.3	1.55	-30.5	2.5
Mongu	Zambia	-15.44	23.52	2000-02	C0	Paropsia bra	L	none	52.8	2.55	-28.6	5.8
Mongu	Zambia	-15.44	23.52	2000-02	C0	Pseudolachn	L	none	47.2	1.74	-25.9	2.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Tephrosia sp	L	none	46.9	3.37	-28.7	0.5
Mongu	Zambia	-15.44	23.52	2000-02	C0	unidentified	LI	litter	47.1	2.13	-25.8	4.4
Mongu	Zambia	-15.44	23.52	2000-02	C0	Basidiomyco	S	saprophytic f	41.2	2.39	-20.1	3.5
Mongu	Zambia	-15.44	23.52	2000-02	C0	Basidiomyco	S	ectomycorrh	24.7	2.15	-20.8	3.8
Mongu	Zambia	-15.44	23.52	2000-02	C0	Lichen	W	none	46.3	2.27	-23.8	-0.7
Mongu	Zambia	-15.44	23.52	2000-02	C0	Lichen	W	none	41.9	1.59	-19.6	-2.6
Mongu	Zambia	-15.44	23.52	2000-02	C0	Lichen	W	none	41	1.38	-20.5	-2.5
Mongu	Zambia	-15.44	23.52	2000-02	C0	Basidiomyco	S	ectomycorrh	39.9	3.77	-21.3	5
Mongu	Zambia	-15.44	23.52	2000-02	C0	Basidiomyco	S	ectomycorrh	28.7	4.04	-21.4	5.9

4. Use stable file formats

Lose years of critical knowledge because modern PCs could not always open old file formats.

http://news.bbc.co.uk/2/hi/6265976.stm

4. Use stable file formats

Lose years of critical knowledge because modern PCs could not always open old file formats.

Lesson: Avoid proprietary formats
They may not be readable in the future

http://news.bbc.co.uk/2/hi/6265976.stm

4. Use stable file formats (cont)

- Use text file formats for tabular data
  - (e.g., .csv (**c**omma-**s**eparated **v**alues))

```
SAFARI 2000 Plant and Soil C and N Isotopes, Southern Africa, 1995-2000
SITE,COUNTRY,LAT,LONG,DATE,START_DEPTH,END_DEPTH,CHARACTERISTICS,C,N,d13C,d15N
units,none,decimal degrees,decimal
degrees,yyyy/mm/dd,cm,cm,none,percent,percent,per mil,per mil
USGS-1,Botswana,-21.62,27.37,1999/07/12,5,20,Hardveld,0.67,0.052,-17,8.9
USGS-2,Botswana,-21.07,27.42,1999/07/12,5,20,Hardveld,0.68,0.063,-18.3,8
USGS-3,Botswana,-20.72,26.83,1999/07/12,5,20,Hardveld,0.94,0.087,-17,6.8
USGS-4,Botswana,-20.52,26.41,1999/07/12,5,20,Hardveld,0.53,0.04,-19.9,5.5
USGS-5,Botswana,-20.55,26.15,1999/07/12,5,20,Lacustrine,2.11,0.162,-15.2,5.9
...
USGS-30,Botswana,-19.81,23.63,1999/07/18,5,20,Alluvium,0.67,0.063,-19.2,11.8
USGS-31,Botswana,-20.62,22.74,1999/07/18,5,20,Hardveld,0.23,0.014,-16.8,16.2
```

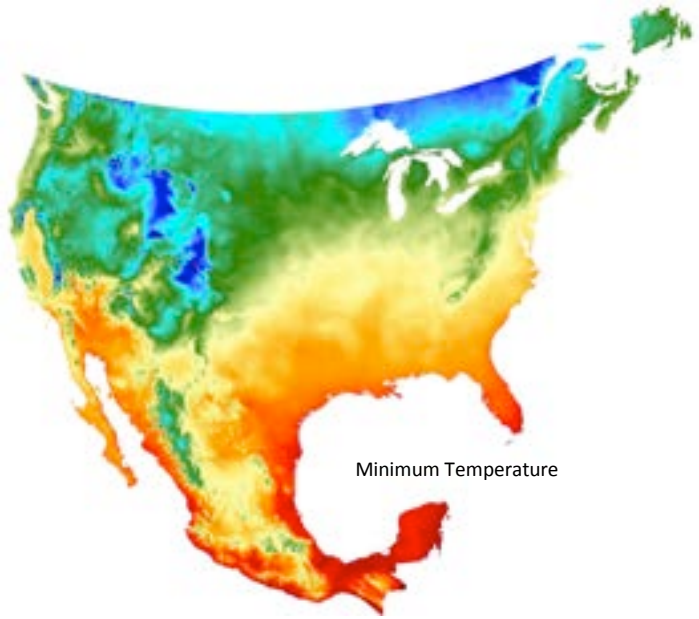
Aranibar, J. N. and S. A. Macko. 2005. SAFARI 2000 Plant and Soil C and N Isotopes, Southern Africa, 1995-2000. Data set. Available on-line [http://daac.ornl.gov/] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/783

4. Use stable file formats (cont)

Suggested Geospatial File Formats

Raster formats

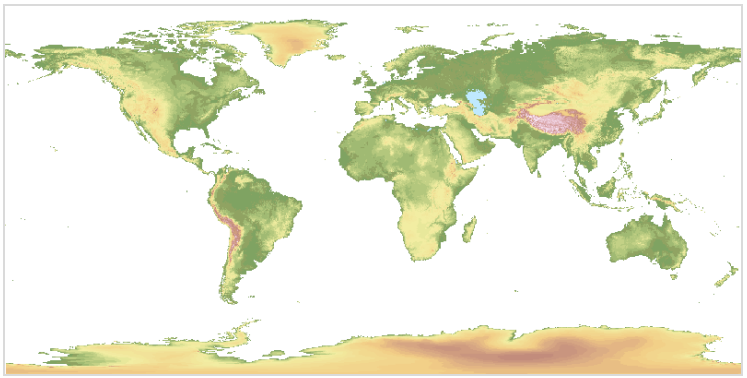
- Geotiff
- netCDF
  - with CF convention preferred
- HDF
- ASCII
  - plain text file gridded format with external projection information



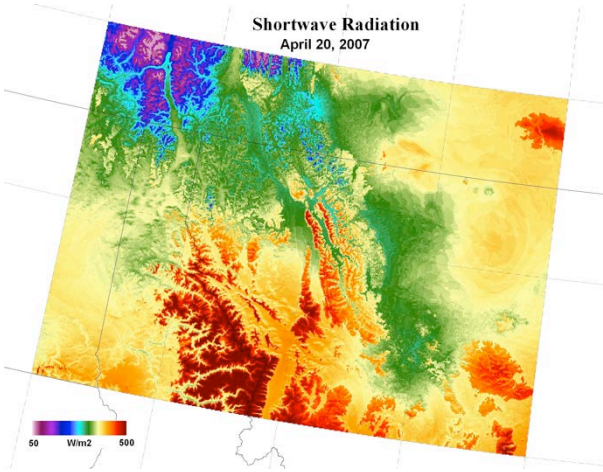
Minimum Temperature

Vector

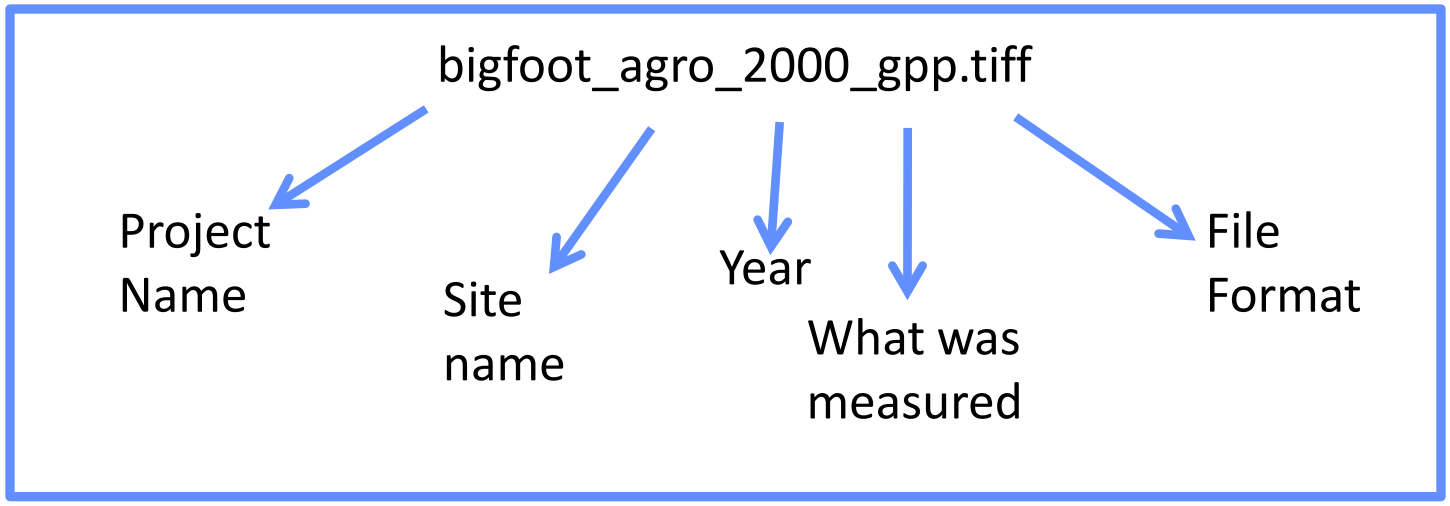
- Shapefile
- KML/GML



GTOPO30 Elevation



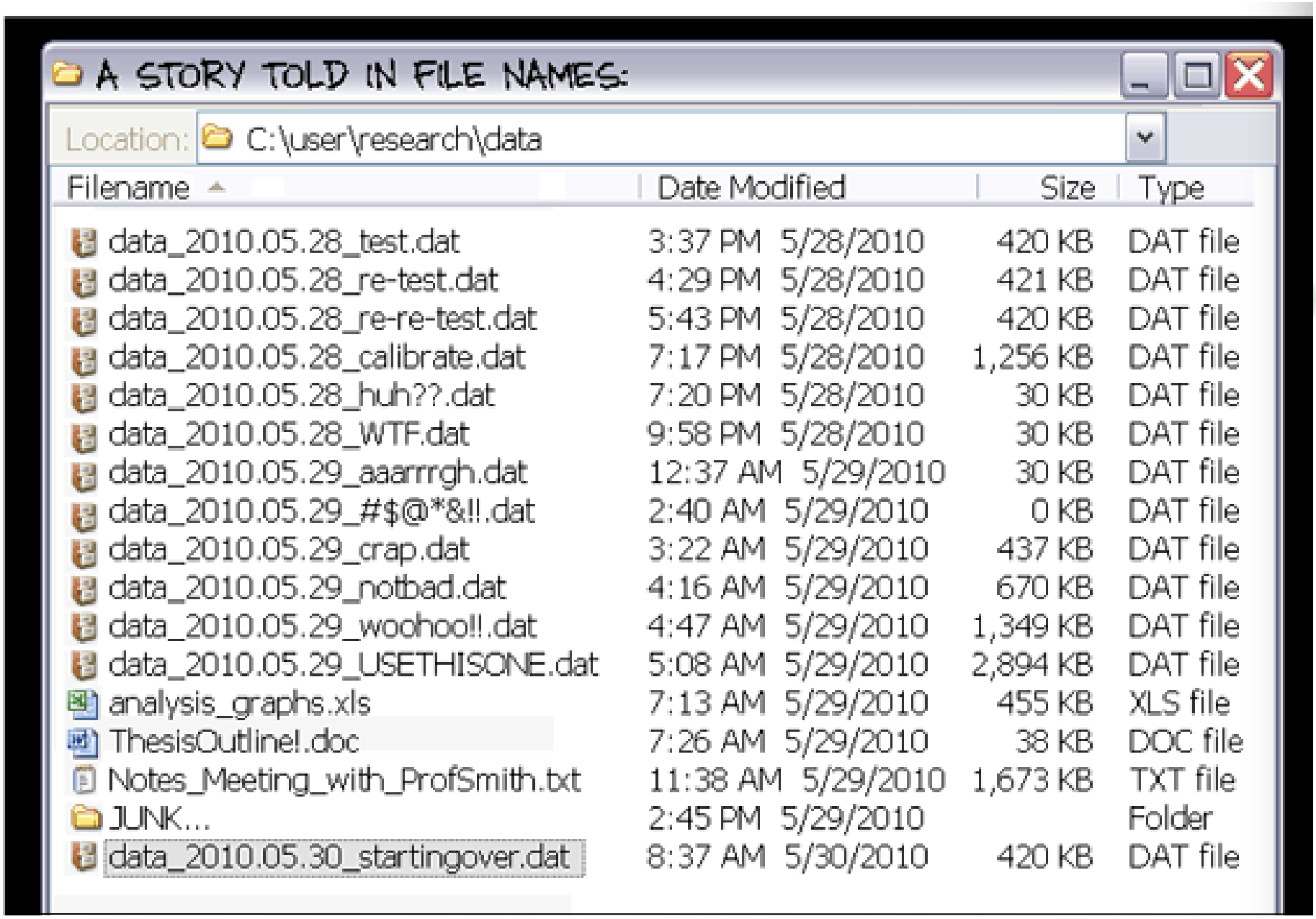
5. Assign descriptive file names



File Names should

- Be Unique
- Reflect contents
- Use ASCII characters only
- Avoid spaces and special characters

Bad:
My data.xls
2001 data.csv
best version.txt
Final data.csv

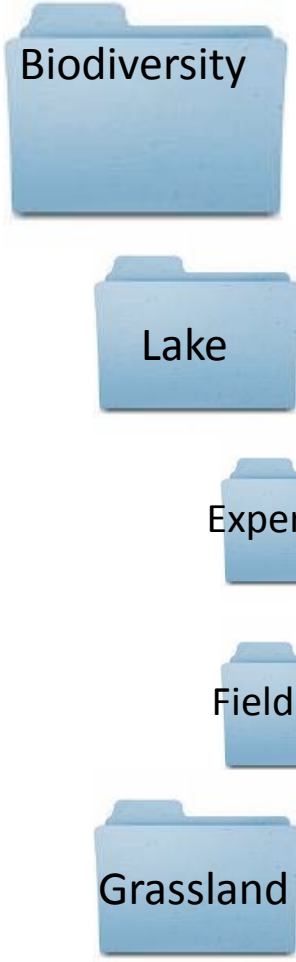


Courtesy of PhD Comics

5. Assign descriptive file names

Organize files logically

- Make sure your file system is logical and efficient



Biodiv_H20_heatExp_2005_2008.csv

Biodiv_H20_predatorExp_2001_2003.csv

…

Biodiv_H20_planktonCount_start2001_active.csv

Biodiv_H20_chla_profiles_2003.csv

…

From S. Hampton

6. Preserve processing information

Raw Data File

Giles_zoopCount_Diel_2001_2003.csv

TAX	COUNT	TEMPC
C	3.97887358	12.3
F	0.97261354	12.7
M	0.53051648	12.1
F	0	11.9
C	10.8823893	12.8
F	43.5295571	13.1
M	21.7647785	14.2
N	61.6668725	12.9
	...	

```
–### Giles_zoop_temp_regress_4jun08.r
–### Load data
–Giles<-
read.csv("Giles_zoopCount_Diel_2001_2003.csv")
–### Look at the data
–Giles
–plot(COUNT~ TEMPC, data=Giles)
–### Log Transform the independent variable (x+1)
–Giles$Lcount<-log(Giles$COUNT+1)
–### Plot the log-transformed y against x
–plot(Lcount ~ TEMPC, data=Giles)
```

Keep raw data raw:

•Do not Include transformations, interpolations, etc in raw file

•Make your raw data “read only” to ensure no changes

When processing data:

- Use a programming language (e.g., R, SAS, MATLAB)
  - Code is a record of the processing done
  - Codes can be revised, rerun

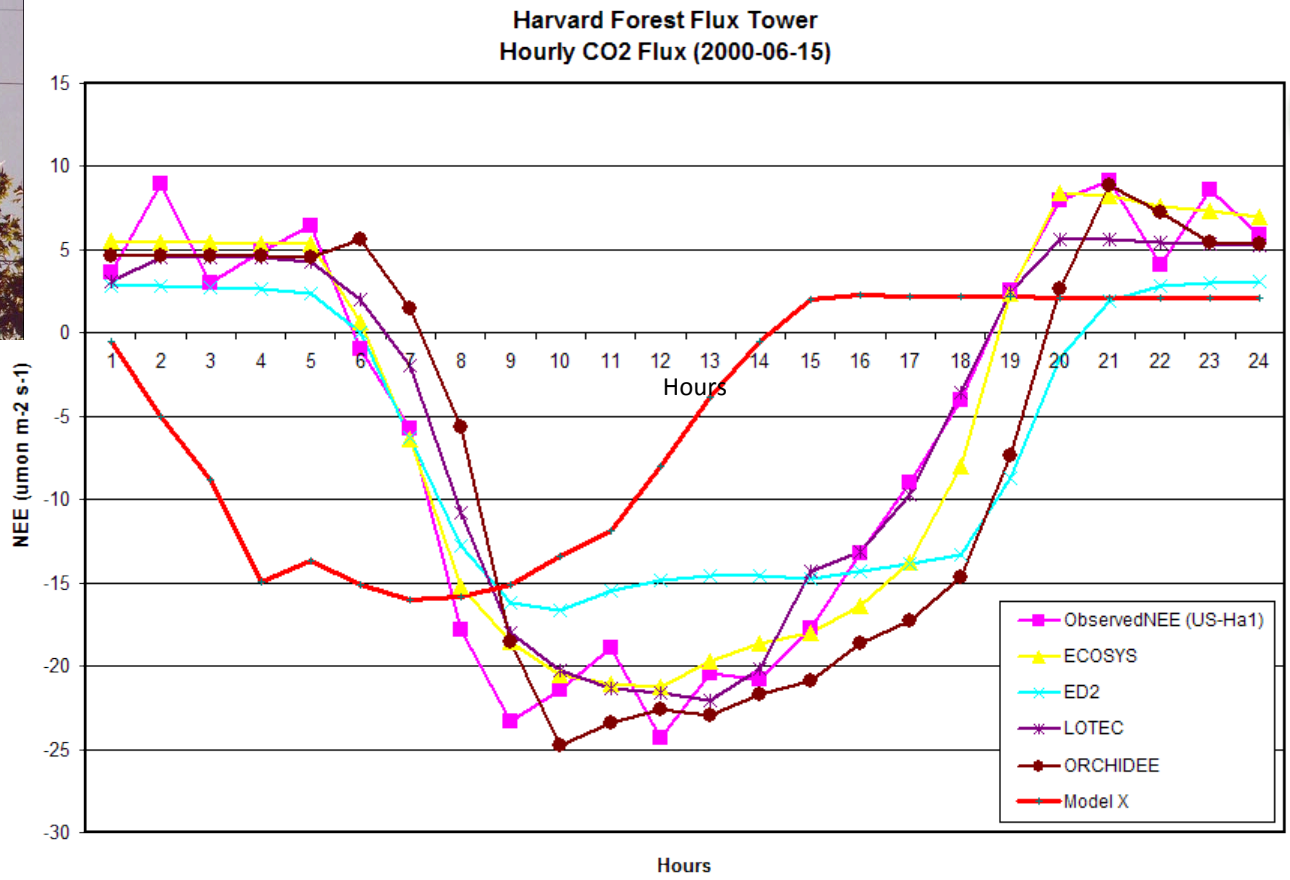
7. Perform basic quality assurance

- Assure that data are delimited and line up in proper columns
- Check that there no missing values (blank cells) for key parameters
- Scan for impossible and anomalous values
- Perform and review statistical summaries
- Map location data (lat/long) and assess errors

No better QA than to analyze data

Plot information to examine outliers

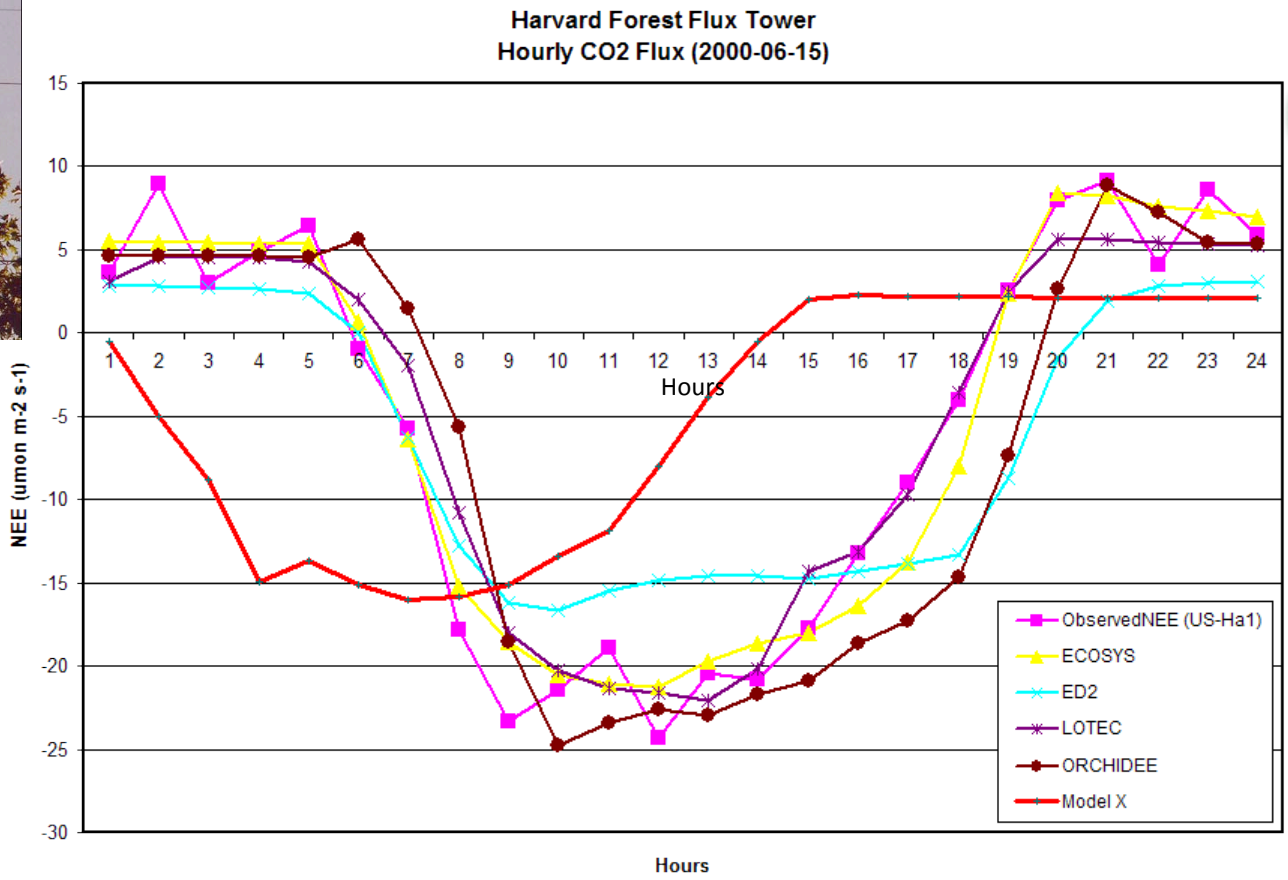
Model-Observation Intercomparison



Data from the North American Carbon Program Interim Synthesis (Courtesy of Dan Ricciuto and Yaxing Wei, ORNL)

Plot information to examine outliers

Model-Observation Intercomparison

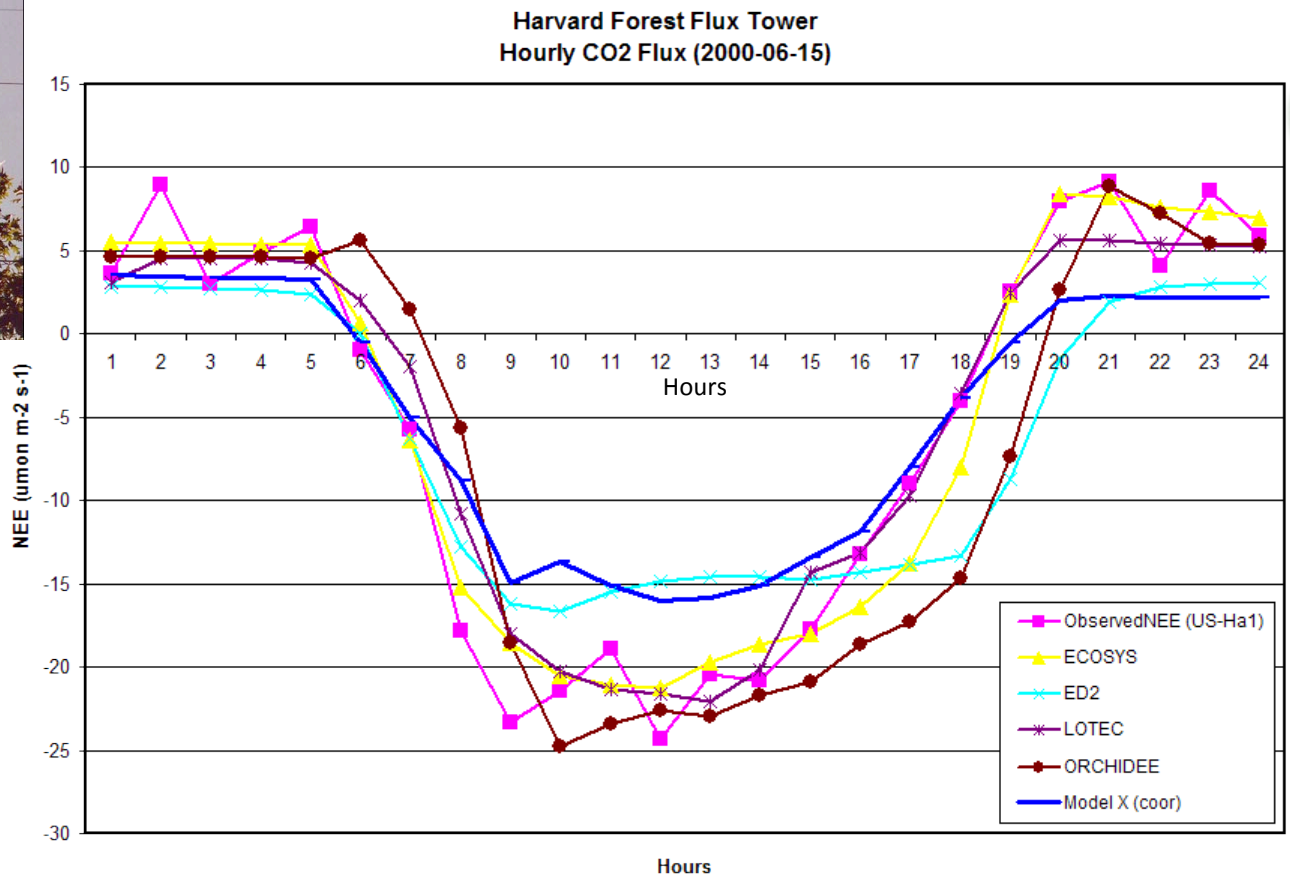


Model X uses UTC time, all others use Eastern Time

Data from the North American Carbon Program Interim Synthesis (Courtesy of Dan Ricciuto and Yaxing Wei, ORNL)

Plot information to examine outliers

Model-Observation Intercomparison



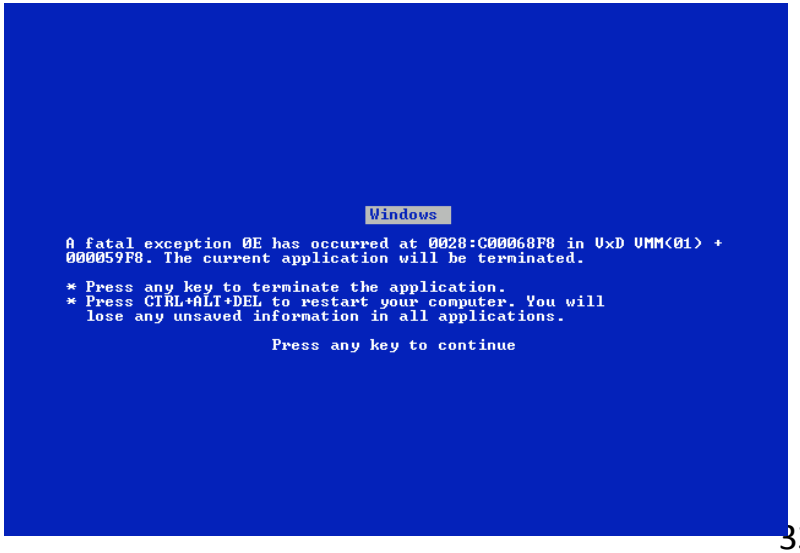
Data from the North American Carbon Program Interim Synthesis
(Courtesy of Dan Ricciuto and Yaxing Wei, ORNL)

8. Provide Documentation / Metadata

- **What** does the data set describe?
- **Why** was the data set created?
- **Who** produced the data set and **Who** prepared the metadata?
- **When** and how frequently were the data collected?
- **Where** were the data collected and with what spatial resolution? (include coordinate reference system)
- **How** was each parameter measured?
- **How** reliable are the data?; what is the uncertainty, measurement accuracy?; what problems remain in the data set?
- **What** assumptions were used to create the data set?
- **What** is the use and distribution policy of the data set? **How** can someone get a copy of the data set?
- **Provide** any references to use of data in publication(s)

9. Protect data

- Create back-up copies often
  - Ideally three copies
    - original, one on-site (external), and one off-site
  - Frequency based on need / risk
- Know that you can recover from a data loss
  - Periodically test your ability to restore information



9. Protect data (cont)

- Ensure that file transfers are done without error
  - Compare checksums before and after transfers
  - numerical value based on the number of bits in the file
    - Example tools to generate checksums

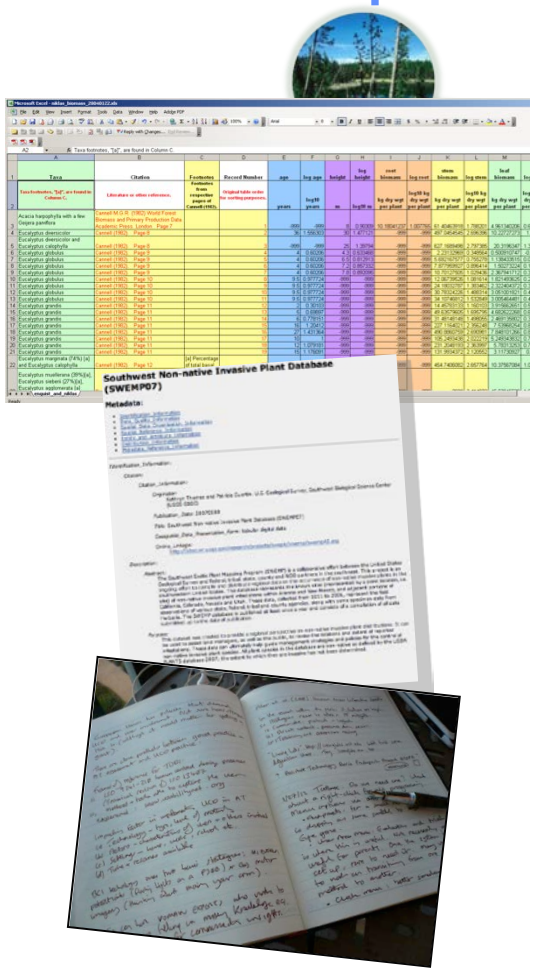
http://www.pc-tools.net/win32/md5sums/

http://corz.org/windows/software/checksum/

10. Preserve Your Data

What to preserve from the research project?

- Well-structured data files, with variables, units, and values defined
- Documentation and metadata record describing the data
- Additional information (provides context)
  - Materials from project wiki/websites
  - Files describing the project, protocols, or field sites (including photos)
  - Publication(s)



Fundamental Data Practices

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

When your project is over, where should the data be archived?

- Part of project planning
- Contact archive / data center early to find out their requirements
  - What additional data management steps would they like you to do?
- Suggested data centers / archives:
  - ORNL DAAC
  - CDIAC
  - Dryad
  - DataONE (Observation Network for Earth)
  - *Ecological Archives*

Best Practices: Conclusion

- Data management is important in today's science
- Well organized data:
  - enables researchers to work more efficiently
  - can be shared easily by collaborators
  - can potentially be re-used in ways not imagined when originally collected
- Include data management in your research workflow.
- Data Management should be a habit

Web Resources

Workshops

http://daac.ornl.gov/workshops/workshops.shtml

- This Webinar
- Other recent workshops

On-line Materials

- Data Management for Data Providers
  - ORNL DAAC Guidance on Data Management

Bibliography

- Christensen, S. W. and L. A. Hook. 2007. NARSTO Data and Metadata Reporting Guidance. Provides reference tables of chemical, physical, and metadata variable names for atmospheric measurements. Available on-line at: http://cdiac.ornl.gov/programs/NARSTO/
- Cook, R.B., R.J. Olson, P. Kanciruk, and L.A. Hook. 2001. Best Practices for Preparing Ecological Data Sets to Share and Archive. Bulletin of the Ecological Society of America, Vol. 82, No. 2, April 2001.
- Hook, L.A., S.K.S. Vannan, T.W. Beaty, R.B. Cook, and B.E. Wilson. 2010.. June 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. http://dx.doi.org/10.3334/ORNLDAAC/BestPractices-2010
- Michener, W.K. and J.W. Brunt (ed.). 2000. Ecological Data: Design, Management and Processing, Methods in Ecology, Blackwell Science. 180p.
- Michener, W. K., J. W. Brunt, J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Non-Geospatial Metadata for Ecology. Ecological Applications. 7:330-342. http://dx.doi.org/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

Additional Data Management Webinars

Encore Webinar

Part 2: Geospatial Data

3 pm EDT (UTC-4:00) Thursday, September 12